Modern Challenges in Distribution Testing

by

Gautam Kamath

B.S., Cornell University (2012)
S.M., Massachusetts Institute of Technology (2014)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2018

Author .....
Department of Electrical Engineering and Computer Science
August 31, 2018

Certified by .....
Constantinos Daskalakis
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Professor of Electrical Engineering and Computer Science
Chair, Department Committee on Graduate Students

Modern Challenges in Distribution Testing

by

Gautam Kamath

Submitted to the Department of Electrical Engineering and Computer Science
on August 31, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

Abstract

Hypothesis testing is one of the most classical problems in statistics. While it has enjoyed over a century of intense study, only recent focus has been on the small-sample regime, with interest in sample complexities and minimax rates. Our understanding of many fundamental problems is now quite mature, but there are several questions which have arisen over the last decade, which have not yet received adequate attention. The goal of this dissertation is to identify and address several contemporary challenges in distribution testing. In particular, we make progress in answering the following questions:

- Can we test distributions with tolerance to model misspecification?
- How does the complexity of distribution testing change as we consider different measures of distance?
- Can we efficiently test for membership in (potentially infinite) classes of distributions?
- How can we avoid the curse of dimensionality when testing multivariate distributions?
- Is it possible to perform hypothesis testing on sensitive data, while respecting privacy of the dataset?
- Can we design more efficient algorithms if the dataset is sampled actively?

Directions for further investigation are also discussed.

Thesis Supervisor: Constantinos Daskalakis
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

The first and most important acknowledgment belongs to my advisor, Constantinos Daskalakis. He has guided and inspired me in countless ways, ranging from technical insights for problem solving, to teaching me how to choose which problems are worth spending my time on. Despite the considerable demands on his time, he still managed to fit six-hour meetings with his students into his schedule, only cut short due to other meetings which were already postponed. He has always provided a watchful eye, ready to step in before I get into too much trouble, and prepared to go to bat for his students if necessary. I've enjoyed his support through both the easy and the difficult parts of graduate school, and I aspire to give the same mentorship to my students as he has to me. I hope that my Ph.D. is only the beginning of our collaborations together.

Taking one step back, my first real foray into theoretical computer science was led by the capable hand of Bobby Kleinberg at Cornell University. He not only taught my first course in algorithm design, but also my second course, nominated me for a theory program at the University of Washington, guided my first theory research experience, and mentored me when I was contemplating graduate schools. Anyone who has interacted with Bobby knows his intuition to be razor-sharp, and he is rarely wrong – as a result, the fact that he would spend his limited time on me gave me confidence that I lacked when I was just beginning as a researcher. If not for him, my research-life story would have ended in at least a dozen different points during my undergraduate studies.

If we want to go even further back, I would like to thank Noah Snavely for teaching my first college class in computer science (CS1114: Introduction to Computing using Matlab and Robotics), thereby giving me a good answer to what computer science actually *is*. Ramin Zabih deserves credit as well, for being the original designer of CS1114, which is one of the most effective methods of fostering undergraduate interest in computer science that I've witnessed so far. Being exposed early on to some basic algorithms, data structures, and computer vision taught me that computer science is far more about problem solving than just mindless programming. If not for this class, it's not clear that I would ever have even majored in computer science. Noah also took me under his wing early on, offering me one

of my first research experiences as a lowly second-year undergraduate.

I would like to thank all of my co-authors so far in my academic career: this includes Jayadev Acharya, Christina Brandt, Bryan Cai, Clément Canonne, Constantinos Daskalakis, Anindya De, Ilias Diakonikolas, Nishanth Dikkala, Steve Hanneke, Nicole Immorlica, Adam Kalai, Daniel M. Kane, Robert Kleinberg, Jerry Li, Ankur Moitra, Vikrant Singhal, Jacob Steinhardt, Alistair Stewart, Ziteng Sun, Christos Tzamos, Jonathan Ullman, John Wright, and Huanyu Zhang. I've learned so much from each and every one of you. Special thanks to Jayadev Acharya, Clément Canonne, Constantinos Daskalakis, Nishanth Dikkala, Jerry Li, and Christos Tzamos, with whom I feel I've spent significant time "in the trenches," or otherwise have been ever-ready to answer my nagging questions or contemplate some half-baked ideas.

Thank you to Ankur Moitra and Ronitt Rubinfeld for serving on my thesis committee. Further thanks to Jayadev Acharya, Mark Bun, Clément Canonne, Ryan Rogers, and John Wright for comments on various parts of this thesis.

During the course of my studies, I have been financially supported in various capacities (some directly and some indirectly) by a number of sources, including Akamai, Google, the National Science Foundation, the Office of Naval Research, Microsoft, MIT, the Simons Foundation, and the Sloan Foundation. I would like to thank all these sources for giving me the privilege of being able to focus on research without having to worry about the stresses of money.

In my time at MIT, the Theory of Computation group has been an anchor in my life. Rather than just a group or a community, I often prefer to call us a family, where my hope is that everyone can feel welcome. There have been a number of individuals who have helped create the lively environment that makes us the envy of other groups in CSAIL. Thanks to Michael Forbes for founding Theory Tea and Theory Retreat, Eric Price for helping support the first Theory Lunch, Cameron Musco for founding Theory Jam, Ryan and Virginia Vassilevska Williams for running Music Night, and Adam Bouland and Aaron Sidford for general contributions to the social atmosphere within the Theory group. In addition to the people mentioned above, I would like to thank Arturs Backurs, Greg Bodwin, Aloni Cohen, Nishanth Dikkala, Rati Gelashvili, Robin Hui, Rio LaVigne, Jerry Li, Sam Park, Govind

Ramnarayan, Christos Tzamos, Adrian Vladu, Matt Weinberg, John Wright, Henry Yuen, and Manolis Zampetakis, for making my time in the Theory group very memorable. At the risk of plagiarizing my Master's thesis acknowledgments, I would especially like to thank Ilya Razenshteyn for being a true Russian friend, in every sense of the phrase.

Behind the scenes, the theory group is propped up by a truly incredible group of administrative assistants. I have been supported at various points by Be Blackburn, Cree Bruins, Debbie Goodwin, Joanne Hanley, Nina Olff, and Rebecca Yadegar. These individuals have truly miraculous skills, including the ability to cause money to appear out of thin air, and an aptitude for hacking through red tape with a machete.

I would like to thank many of my officemates in G628 over the years, including Arturs Backurs, Nishanth Dikkala, Themis Gouleakis, Robin Hui, Govind Ramnarayan, and Ilya Razenshteyn. There were always meaningful conversations to be had, both during and after hours, and our office was definitely one of the most interesting ones in Stata.

Thanks to my housemates for four years at 9 Hamlin: Aloni Cohen, Daniel Grier, and Rebecca Powell. There was definitely a lot of heart in this home, and I will have fond memories of many conversations, and support during challenging times.

I also appreciate the friendship, support, and company of many individuals in the wider MIT community. Some people I would like to highlight are Sara Achour, Ariel Anders, Michel Babany, Eva Golos, Michal Grzadkowski, Marek Hempel, Twan Koolen, Albert Kwon, Danielle Pace, Andrew Sabisch, Matt Staib, Jennifer Tang, Alin Tomescu, and especially Yijin Wei. You've pressed a thumb firmly on the "life" side of my work-life balance, giving me a much-needed escape from the stresses of life at MIT otherwise, causing me to sometimes reprioritize what really matters.

I frequently tells others I love traveling as an academic. The main reason isn't due to the locations (I even enjoyed FOCS 2016 in New Brunswick!), but due to the abundance of amazing people I've met throughout the academic world. A few friends who are responsible for some fond memories during my times on the road include Zahra Ashktorab, Mark Bun, Sam Hopkins, Fotis Iliopolous, Thanasis Lianeas, Thodoris Lykouris, Jasmine Nirody, Ioannis Panageas, Aviad Rubinstein, Tselil Schramm, Thomas Steinke, and Juba Ziani.

Outside of MIT and academic friends, I have a strong circle of support, with many

close friends from Cornell I have been very fortunate to retain even years after graduating. Some include Suryansh Agarwal, Luke Chan, Taylor Chan, Jesseon Chang, Ben Greenman, Dominick Grochowina, Laura Grochowina, Jasdeep Hundal, Aparajith Kannan, Frohman Lee, Rocky Li, Robin Magalis, Anirvan Mukherjee, Elisabeth Rosen, Harry Terkelsen, Lucas Waye, Michael Wu, Stephanie Wu, Phoebe Yu, and Gregory Zecchini. These friendships have persevered so far, and I'm confident they will remain strong for many years to come.

Finally, I conclude by thanking my family, including my parents Markad and Padma, my brother and sister-in-law Anand and Archana, and my nephew Ayansh. Many academics feel that all their successes are the product of being lucky at various key points in their career. If this is the case, my first stroke of luck was being born into a family with such a strong value for education, who has supported me on the long road to completing this degree.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Hypothesis testing is one of the most classical statistical questions, with a history dating back at least three hundred years [Arb10, Lap78]. It asks the following question: can a dataset, which is assumed to be distributed according to some unknown distribution, be explained by some hypothesis model? For example, by consulting Christening records from 1629 to 1710, Arbuthnott rejected the hypothesis that the male birth rate is equal to the female birth rate, and attributed this discrepancy to divine intervention [Arb10].

Modern study in statistics can be traced back to the early 20th century, roughly originating from Pearson's introduction of the χ^2-test [Pea00]. Over the last century, these problems have enjoyed significant study (see, e.g., [Fis25, LR05]), often with the goal of computing p-values for tests of statistical significance.

However, some of the classical works in this field are focused on directions which are not entirely aligned with the needs of modern data science. Much of the analysis is targeted at the asymptotic regime, in which we allow the number of samples to go to infinity. For instance, a common goal is to understand the limiting distribution of a statistic. The drawback is that modern data analysis is frequently performed on datasets with extremely large domains, settings in which limited amounts of data may cause these asymptotic guarantees to not even be approximately true. Furthermore, an emphasis is placed on *significance* rates: that is, understanding the probability with which we reject the hypothesized model (also known as the *null hypothesis*). Somewhat less rigorously considered is the *power* of statistical tests: the probability of not rejecting the model when we are in fact looking at a different distribution

(an *alternative hypothesis*). In many works, this is analyzed with respect to limited classes of alternative hypotheses or empirically measured on several "natural" alternatives, in both cases, lacking the precise and widely applicable guarantees we desire.

Guided by these motivations, a recent direction of study has been on understanding the *sample complexity* (or equivalently, *minimax rates*) of hypothesis testing. The broad interest in these questions has caused a number of communities (including statistics, information theory, machine learning, and theoretical computer science) to converge upon this common goal. Perhaps the starting point in the statistics community can be considered the work of Ingster and coauthors [Ing94, Ing97, IS03], which studied the minimax rates of various tests. Our main focus in this thesis is the study in theoretical computer science, of which the genesis is generally considered the work Goldreich and Ron [GR00], which studied the problem of testing uniformity of a distribution for the application of testing expansion of a bounded degree graph. As this investigation is within the field of computer science, in addition to the desiderata mentioned above, there is an additional emphasis on developing algorithms and tests that are *computationally efficient*. Since this introduction, there has been a flurry of results, culminating in a tight understanding of the sample complexity of many fundamental problems of interest (see Section 1.1 for a brief history of the field so far).

While it might be tempting to declare victory in light of this success, the more applied side of the field has changed significantly over the past century. In particular, there are a number of settings and requirements in modern data science that were not dreamed of when hypothesis testing was first considered, bearing with them a number of new challenges we must face. The goal of this thesis is to identify and address several of these unresolved questions, and highlight some directions for further investigation. Specifically, the aspects of distribution testing which we focus on in this thesis are as follows:

- **Tolerance and Alternative Distances.** Classically, hypothesis testing has focused on a single null hypothesis: is our unknown distribution p equal to some model q? However, it seems unreasonable for our model to *precisely* match the unknown distribution, as small errors might have been introduced for a number of reasons. As a result, we would really like to test whether p is *close* to the model q. Additionally, while total variation is the canonical distance for distribution testing, other metrics

and divergences may be more natural for some settings. We investigate both of these concerns in Chapter 2.

- **Composite Hypotheses.** Somewhat related to the previous topic, we would also like to test if our distribution belongs to some (potentially infinite) family of distributions. For example, a researcher might want to determine if their data is distributed according to the law of *some* unimodal distribution, not a *particular* unimodal distribution. We describe results for testing such *composite* null hypotheses in Chapter 3.

- **High Dimensions.** Modern settings of data analysis involve highly multivariate domains. In these settings, the curse of dimensionality manifests: distribution testing in general high-dimensional settings necessitates a sample complexity which is exponential in the dimension. How can we perform statistical inference tasks in multivariate settings? In Chapter 4, we design specialized algorithms for *Ising models*, a common graphical model for high-dimensional datasets.

- **Data Privacy.** Statistical tasks are often performed on sensitive individual data. For instance, a medical study might operate on patient health records. Can we perform statistical procedures while ensuring *privacy* of the dataset? This is the topic we explore in Chapter 5.

- **Conditional Sampling.** Nowadays, data collection is not a static process: we may have some additional control over how our dataset is acquired. Given this extra power, can we exploit it to design more efficient algorithms? This motivates the study of the *conditional sampling* model, which we discuss in Chapter 6.

- **Other Directions.** Naturally, there are a number of directions that we will not have the opportunity to address elsewhere in this thesis. In Chapter 7, we briefly describe and discuss a few more modern challenges, and give pointers to the relevant literature.

1.1 Background, Prior Work, and Outline of Contributions

In this section, we provide a brief history of the field of distribution testing within theoretical computer science and adjacent fields. We also outline our contributions and place them in context with existing work. Discussion of some classical results will be deferred to later chapters, when they become more relevant to the topic of discussion.

Within theoretical computer science, the field of distribution testing originated as a subfield of property testing. The problem was first stated by Goldreich, Goldwasser, and Ron in [GGR96], but it was first studied in earnest by Goldreich and Ron [GR00], who were concerned with the problem of testing whether a bounded-degree graph is an expander. Their approach is based on the fact that random walks on expander graphs are rapidly mixing to the uniform distribution. This allows them to reduce their problem to *uniformity testing*: given samples from a distribution p, is it uniform over its support, or is it ε-far in total variation distance[1] from the uniform distribution? While they do not phrase their results in this language, they imply the following theorem:

Theorem 1 ([GR00])**.** *There exists a polynomial-time algorithm which, given sample access to a distribution p over $[n]$, can distinguish (with probability at least $2/3$) between the case where p is uniform over $[n]$, and the case where p is ε-far in total variation distance from being uniform. The algorithm uses $O\left(\sqrt{n}/\varepsilon^4\right)$ samples.*

At first glance, it might seem surprising that one can test whether a distribution over $[n]$ is uniform from only $O(\sqrt{n})$ samples. Indeed, with this few samples, almost all elements of the domain will never be observed. One counter-intuitive fact which provides a glimmer of hope is the well-known *birthday paradox*: if one takes only $\Theta(\sqrt{n})$ samples from the uniform distribution, collisions between the sampled elements will start to occur. This is the type of statistic exploited by Goldreich and Ron. One can observe that the uniform distribution minimizes the number of collisions between samples from p, and therefore one can test uniformity of a distribution by appropriately thresholding the number of collisions.

[1] Total variation distance, as well as several other concepts we require in this thesis, is defined in Section 1.3.

More generally, one can consider the problem of *identity testing*: given samples from a distribution p, is it equal to some particular distribution q, or is it ε-far in total variation distance from q? Batu, Fischer, Fortnow, Kumar, Rubinfeld, and White gave the first algorithm for this problem [BFF+01], which requires $\tilde{O}\left(\sqrt{n} \cdot \mathrm{poly}(\varepsilon^{-1})\right)$ samples. Note that this nearly matches the dependence on n for uniformity testing. Their method partitions the domain $[n]$ so that q is roughly uniform on each part, and then tests whether p is uniform over each part using the algorithm of Goldreich and Ron [GR00]. This method of reducing identity testing to uniformity testing is fundamental, and a number of recent works essentially do this, either by rescaling [ADK15] or splitting [DK16, Gol16] elements of the domain. Indeed, as Goldreich showed via the splitting method, uniformity testing is *complete* for testing identity [Gol16]. In other words, up to constant factors, an algorithm for testing uniformity implies an algorithm for testing identity with the same sample complexity.

The first optimal upper bounds of $O(\sqrt{n}/\varepsilon^2)$ were provided by Paninski [Pan08], and Valiant and Valiant [VV14]. The former algorithm only works when $\varepsilon = \Omega(n^{-1/4})$, and is for the special case of testing uniformity (which we now know to be complete), while the latter algorithm removed both these restrictions. Optimal algorithms for this problem have since been rediscovered several times [ADK15, DKN15b, DGPP16, DK16, DGPP18, DKW18]. To complement these upper bounds, Paninski also showed an information-theoretic lower bound of $\Omega(\sqrt{n}/\varepsilon^2)$ for uniformity testing. The dependence on n is intuitive from the fact that the $\Theta(\sqrt{n})$ in the birthday paradox is tight: if one receives fewer samples from a distribution which is uniform over a random half of the domain $[n]$, no collisions will be witnessed, and thus the distribution is indistinguishable from the uniform distribution. Obtaining the tight inverse-quadratic dependence on ε requires a more careful argument. To summarize, this line of work has resulted in the following optimal sample complexity for testing identity:

Theorem 2 ([Pan08, VV14])**.** *There exists a polynomial-time algorithm which, given sample access to a distribution p over $[n]$ and the description of a distribution q over $[n]$, can distinguish (with probability at least $2/3$) between the case where p is equal to q, and the case where p is ε-far in total variation distance from q. The algorithm uses $O\left(\sqrt{n}/\varepsilon^2\right)$ samples. Furthermore, any algorithm for this problem which succeeds with probability at least $2/3$ must use $\Omega\left(\sqrt{n}/\varepsilon^2\right)$ samples.*

While this settles the complexity of identity testing, this formulation is rather basic, and may be of limited interest in many real-world settings. For example, it seems unreasonable for the distribution p to exactly match the hypothesized model q, as error may be introduced into the process at a number of points. Therefore, it may be more motivated to study the *tolerant* case, when p and q are *close*, rather than equal. Perhaps the most natural formulation of this question is as follows: how many samples are required to distinguish the case where p is $\varepsilon/2$-close to q from the case where p is ε-far from q? Surprisingly, as shown by Valiant and Valiant [VV10a, VV10b, VV11a, VV11b], the problem is much harder than before: the sample complexity jumps to $\Theta\left(\frac{n}{\log n}\right)$. This is rather unsatisfying, as the tolerance we desire cost us a near-quadratic blow-up in the sample complexity, thus motivating study of the following question in Chapter 2:

Question 1. *Can we design sample-efficient algorithms for tolerant identity testing, potentially by considering other distance measures?*

We answer this question affirmatively, and provide a number of computationally efficient and sample-optimal tests for a number of testing problems. We show that some types of tolerance come at no cost, including tolerance in ℓ_2-distance or the χ^2-divergence.

Another way to lessen the restrictive nature of identity testing is to consider *composite hypotheses*. In addition to identity testing, the paper of Batu et al. [BFF$^+$01] also studied *independence testing*: given samples from a two-dimensional distribution p, is it a product measure? Observe that, in contrast to previous problems where the null hypothesis is a single distribution, we wish to test whether p is equal to one of infinitely many q. In a simliar vein, Batu, Kumar, and Rubinfeld [BKR04] investigated *monotonicity testing*: given samples from a distribution p over $[n]$, is its probability mass function monotone non-decreasing? Interestingly, the sample complexity is once again $\tilde{O}\left(\sqrt{n} \cdot \mathrm{poly}(\varepsilon^{-1})\right)$: the cost is comparable to that of testing identity to a single hypothesis q. This raises the main question of study in Chapter 3:

Question 2. *Are there sample-efficient tests for testing if a distribution belongs to some structured family?*

We once again answer this affirmatively for a number of families of interest by providing

sample-optimal and computationally efficient algorithms for testing many natural classes of distributions. We find that testing for an entire class of distributions often comes at no cost over testing for a single distribution q. Surprisingly, we must exploit a technical connection with Question 1.

Most of the problems mentioned so far have been studied in low-dimensional settings: all are univariate, with the exception of independence testing, which is bivariate. However, distribution testing problems frequently arise in high-dimensional settings, which are far more common in modern data analysis. Troublingly, if one embeds the lower bound construction of Paninski [Pan08] into a multivariate domain, it can be shown that the curse of dimensionality necessitates a sample complexity which is exponential in the dimension. This seems at odds with our needs to perform statistical analysis in these settings. One vestige of hope is the fact that structural assumptions on the underlying distribution often enable significant savings for testing problems. This was observed by Batu, Kumar, and Rubinfeld [BKR04], who achieved exponential savings in the sample complexity when the underlying distributions are assumed to be monotone. Note that this is testing *with* structure, not testing *for* structure, as considered in Question 2. Specifically, we investigate the following question in Chapter 4:

Question 3. *Is high-dimensional distribution testing tractable over structured classes of densities?*

If the underlying distribution is an Ising model, we show that the curse of dimensionality can be avoided. As a result, the sample complexity is polynomial in the dimension, rather than exponential.

Turning briefly to the applied side of hypothesis testing, statistical methods are now applied in an increasingly wide variety of settings. One recent area of interest is genome wide association studies (GWASs), where one tries to detect correlations between traits and genetic variants using hypothesis tests for independence. Naturally, these datasets are quite sensitive in nature, containing health information of large collections of individuals. Worryingly, it was recently shown by Homer et al. that naïve methods might allow an attacker to identify individuals who participated in such a study [HSR$^+$08], thus motivating

interest in methods which explicitly try to prevent such attacks. We turn our focus to this issue in Chapter 5:

Question 4. *Can we perform distributional hypothesis testing and property estimation while respecting privacy of the dataset?*

Using the celebrated notion of differential privacy, we give efficient private algorithms for several problems of interest. Surprisingly, in many parameter regimes of interest, privacy comes at a negligible cost in the sample complexity.

Finally, a new direction has attempted to quantify the savings enjoyed when we have stronger access to the underlying distribution. For example, suppose that, rather than simply being given a dataset, we can somehow (potentially adaptively) *gather* a dataset. This evokes the spirit of the celebrated *active learning* model in machine learning, in which one can request labels for specific datapoints. The recently-introduced conditional sampling model, in which the algorithm can elicit samples conditioned on being from query sets it specifies, attempts to capture this type of phenomenon [CFGM13, CRS14], and it has been shown that the complexity of several problems drops dramatically. For instance, identity testing requires only $\tilde{O}(1/\varepsilon^2)$ queries [FJO$^+$15], in contrast to the $O(\sqrt{n}/\varepsilon^2)$ samples in the standard model, completely eliminating the dependence on the size of the support n. However, the complexity of a number of basic questions is still not well understood – we fill several gaps in the literature in Chapter 6, contemplating the following question:

Question 5. *How far can we push the savings enabled by the conditional sampling model for distribution tesing, and how does the power of the model change when it is adaptive versus non-adaptive?*

We conclude this section by mentioning prior work on a few other interesting questions in distribution testing.

The success probability of identity testing in Theorem 2 is at least 2/3. This can be boosted to $1-\delta$ at a multiplicative cost of $O(\log(1/\delta))$ samples, as we discuss in Section 1.3. However, it turns out one can do slightly better – in some parameter regimes, one may get away with paying only a multiplicative $O(\sqrt{\log(1/\delta)})$ [HM13b, DGPP18].

As mentioned before, if one can test uniformity, one can test identity to any distribution q. As such, the uniform distribution is the *hardest* distribution, and we can actually do better on an instance-by-instance basis, as shown in [VV14, DK16, BCG17].

The harder problem of equivalence testing, when both p and q are unknown, was first studied in [BFR+00]. Optimal upper and lower bounds were given in [CDVV14], where the lower bound was based off the approach of [Val11].

Theorem 3 ([Val11, CDVV14])**.** *There exists a polynomial-time algorithm which, given sample access to distributions p and q over $[n]$, can distinguish (with probability at least $2/3$) between the case where p is equal to q, and the case where p and q are ε-far from each other in total variation distance. The algorithm uses $O\left(\max\left\{\frac{n^{2/3}}{\varepsilon^{4/3}}, \frac{n^{1/2}}{\varepsilon^2}\right\}\right)$ samples. Furthermore, any algorithm for this problem must use $\Omega\left(\max\left\{\frac{n^{2/3}}{\varepsilon^{4/3}}, \frac{n^{1/2}}{\varepsilon^2}\right\}\right)$ samples.*

In fact, one can show that it is possible to test equivalence when given unequal numbers of samples from the two distributions, as investigated in [AJOS14b, BV15, DK16].

Beyond distribution testing, there has also been significant study on several related problems of *property estimation*, some of which we describe and discuss in Section 1.3.1.2. All of these problems are in the "barely-sublinear" regime, with sample complexity $\Theta\left(\frac{n}{\log n}\right)$. Specific lines of work include Shannon and Rényi entropy estimation [Pan03, BDKR05, VV13, WY16, JVHW17, AOST17, OS17], support coverage and support size estimation [OSW16, WY18], and estimating distance between discrete distributions [VV10a, VV10b, VV11a, VV11b, JHW16, HJW16, JVHW17].

Our coverage in this section is necessarily incomplete, focusing on the results which are most relevant to our work in this thesis. For further background, surveys, and books which may be of interest, we refer the reader to [Rub12, Can15b, Gol17, BW17b].

1.2 Organization and Bibliographic Information

Most contents of this thesis have appeared previously as other publications, which we briefly outline.

In the remainder of Chapter 1, we standardize notation and overview some preliminaries that we will require for the rest of the thesis.

Chapter 2 focuses on distribution testing with alternative distance measures. This is based on the paper "Which Distribution Distances are Sublinearly Testable?" which is joint work with Constantinos Daskalakis and John Wright, and appeared in the Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms [DKW18].

Chapter 3 studies the testing of shape restrictions of distributions. This is based on the paper "Optimal Testing for Properties of Distributions," which is joint work with Jayadev Acharya and Constantinos Daskalakis, and appeared in Advances in Neural Information Processing Systems 28 [ADK15].

Chapter 4 investigates testing in structured high-dimensional domains, when the underlying distribution is known to be an Ising model. This is based on the paper "Testing Ising Models," which is joint work with Constantinos Daskalakis and Nishanth Dikkala, and appeared in the Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms [DDK18].

Chapter 5 describes results on distribution testing and property estimation with privacy constraints. This is based on two papers: the first is "Priv'IT: Private and Sample Efficient Identity Testing," which is joint work with Bryan Cai and Constantinos Daskalakis, and appeared in the Proceedings of the 34th International Conference on Machine Learning [CDK17]. The second is "INSPECTRE: Privately Estimating the Unseen," which is joint work with Jayadev Acharya, Ziteng Sun, and Huanyu Zhang, and appeared in the Proceedings of the 35th International Conference on Machine Learning [AKSZ18].

Chapter 6 discusses distribution testing when one is given conditional sampling access to the underlying distribution. This is based on two papers: the first is "A Chasm Between Identity and Equivalence Testing with Conditional Queries," which is joint work with Jayadev Acharya and Clément L. Canonne, and appeared in the proceedings of the 19th International Workshop on Randomization and Computation [ACK15b]. The second is "Anaconda: A Non-Adaptive Conditional Sampling Algorithm for Distribution Testing," which is joint work with Christos Tzamos, and is currently available as a preprint [KT18].

Chapter 7 lists some other recent directions in distribution testing, which are ripe for further investigation.

Other papers by the author over the course of his PhD studies, but not included in this

thesis include [DK14, ACK15a, DKT15, DDKT16, DKK$^+$16, DDK17, DKK$^+$17, DKK$^+$18a, DKK$^+$18b, KLSU18, HKKT18].

1.3 Preliminaries and Notation

In this thesis, we will use the symbols p and q to denote probability distributions. Distributions will usually be discrete, with support $[n] = \{1, \ldots, n\}$. To represent the probability of observing element i, we will use either $p(i)$ or p_i, the choice of which will be clear from context. Except when stated otherwise, for a set $S \subseteq [n]$ and a distribution p over $[n]$, p_S is the vector p restricted to the coordinates in S. We will call this a *restriction* of distribution p. We will generally use $\varepsilon \in (0, 1)$ to measure the accuracy of a statistical procedure: either the parameter in the "soundness" case for distribution testing or how accurately we must estimate some functional of a distribution for property estimation. $\delta \in (0, 1)$ is used to indicate the probability that the test or estimator fails (i.e., outputs an incorrect or inaccurate answer). The exception will be Chapter 5, where we use ε and δ for privacy parameters, we will use α and β in their place, respectively. The symbol m will be used for the number of samples. We will use $\mathcal{U}_n$ for the uniform distribution over $[n]$.

1.3.1 Problems Statements

The primary problem of interest in this thesis is *distribution testing*. There are many flavors which we describe in Section 1.3.1.1. Other property estimation problems we study are described in Section 1.3.1.2.

1.3.1.1 Distribution Testing

We define some classical distribution testing problems here, extensions will be defined in later chapters as they become relevant.

The most classical distribution testing problem is *identity testing*. Given an explicit description of a distribution q, the goal is to distinguish between the following two cases:

- Completeness: $p = q$;
- Soundness: $d_{\mathrm{TV}}(p, q) \geq \varepsilon$,

where $d_{\mathrm{TV}}(p, q)$ is the total variation distance between p and q. We would like for our test to be successful with probability at least $2/3$: this can be boosted to probability $1 - \delta$ by a

standard argument at the cost of a multiplicative $\log(1/\delta)$ in the sample complexity.[2] When the distribution $q = \mathcal{U}_n$, then the problem is called *uniformity testing*. In the statistics community, these problems are referred to as *one-sample testing*: this is in reference to the fact that we have samples from one unknown distribution.

In the harder case where q is not explicitly known and we only have sample access to q, then the problem is called *equivalence testing*.[3] As one might expect, this variation of the problem is called *two-sample testing* by the statistics community.

Sometimes, the completeness condition is changed from $d_{\mathrm{TV}}(p, q) = 0$ to $d_{\mathrm{TV}}(p, q) \leq \varepsilon'$. In this case, the problem is called *tolerant testing*. Often, ε' is chosen to be $\varepsilon/2$, or some other value which gives a constant factor gap between the soundness and completeness cases.

1.3.1.2 Other Problems

There are a number of other *property estimation* problems which we explore in this thesis.

Support Size. The support size of a distribution p is $S(p) = |\{x : p(x) > 0\}|$, the number of symbols with non-zero probability values. However, notice that estimating $S(p)$ from samples can be hard due to the presence of symbols with negligible, yet non-zero probabilities. To circumvent this issue, [RRSS09] proposed to study the problem when the smallest probability is bounded. Let $\Delta_{\geq \frac{1}{n}} \triangleq \{p \in \Delta : p(x) \in \{0\} \cup [1/n, 1]\}$ be the set of all distributions where all non-zero probabilities have value at least $1/n$. Given samples from $p \in \Delta_{\geq \frac{1}{n}}$, our goal is to estimate $S(p)$ up to $\pm\varepsilon n$.

Support Coverage. For a distribution p, and an integer k, let $S_k(p) = \sum_x (1-(1-p(x))^k)$, be the expected number of symbols that appear when we obtain k independent samples from the distribution p. The objective is, given samples from p, to estimate $S_k(p)$ up to an additive $\pm\varepsilon k$.

Support coverage arises in many ecological and biological studies [CCG+12] to quantify

[2] The argument is as follows: run the test independently $O(\log(1/\delta))$ times, and take the majority result. Since each result is correct with probability at least $2/3$, then, by a Chernoff bound, the correct result will be the majority with probability $1 - \delta$.

[3] The problem is also known as *closeness testing*, though we will generally use the former term in this thesis.

the number of *new* elements (gene mutations, species, words, etc.) that can be expected to be seen in the future. Good and Toulmin [GT56] proposed an estimator that can extrapolate by a factor of up to 2: it can use $k/2$ samples to estimate $S_k(p)$.

Entropy. The Shannon entropy of a distribution p is $H(p) = \sum_x p(x) \log \frac{1}{p(x)}$. $H(p)$ is a central object in information theory [CT06a], and also arises in many fields such as machine learning [Now12], neuroscience [BWM97, NBdRvS04], and others. Estimating $H(p)$ is unfortunately impossible with any finite number of samples due to the possibility of infinite support. To circumvent this, a natural approach is to consider distributions in Δ_n, where Δ_n is all discrete distributions over at most n symbols. The goal is to estimate the entropy of a distribution in Δ_n up to $\pm\varepsilon$.

Distance between Distributions. The ℓ_1-distance between a distribution p and q is $\|p-q\|_1 = \sum_i |p(i)-q(i)| \in [0,1]$. This is most frequently studied when $q = \mathcal{U}_n$. The goal is to estimate $\|p-q\|_1$ up to an additive ε. Estimating the distance between distributions is closely related to the problem of tolerant distribution testing, when we want to determine whether two distributions are close or far in ℓ_1 distance. For more discussion on this connection and distribution testing, see [PRR06, DKW18].

1.3.2 Measures of Distance between Distributions

In this thesis, a number of different distances and divergences will be core to our work.

Definition 1. *The* total variation distance *or* statistical distance *between p and q is defined as*

$$d_{\mathrm{TV}}(p,q) = \max_{S \subseteq [n]} p(S) - q(S) = \frac{1}{2} \sum_{i \in [n]} |p_i - q_i| = \frac{1}{2} \|p - q\|_1 \in [0,1].$$

Note that, up to a factor of two, this is equivalent to the ℓ_1 distance *between p and q.*

Definition 2. *The* KL divergence *between p and q is defined as*

$$d_{\mathrm{KL}}(p,q) = \sum_{i \in [n]} p_i \ln \frac{p_i}{q_i} \in [0, \infty).$$

This definition uses the convention that $0 \ln 0 = 0$.

Definition 3. *The* symmetric KL divergence *between* p *and* q *is defined as*

$$d_{\mathrm{SKL}}(p,q) = d_{\mathrm{KL}}(p,q) + d_{\mathrm{KL}}(q,p) = \sum_{i \in [n]} p_i \ln \frac{p_i}{q_i} + q_i \ln \frac{q_i}{p_i} \in [0, \infty).$$

Definition 4. *The* Hellinger distance *between* p *and* q *is defined as*

$$d_{\mathrm{H}}(p,q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i \in [n]} \left(\sqrt{p_i} - \sqrt{q_i}\right)^2} \in [0,1].$$

Definition 5. *The* χ^2-divergence *(or* chi-squared divergence*) between* p *and* q *is defined as*

$$d_{\chi^2}(p,q) = \sum_{i \in [n]} \frac{(p_i - q_i)^2}{q_i} \in [0, \infty).$$

Definition 6. *The* ℓ_2 distance *between* p *and* q *is defined as*

$$d_{\ell_2}(p,q) = \sqrt{\sum_{i \in [n]} (p_i - q_i)^2} = \|p - q\|_2 \in [0,1].$$

We also define these distances for restrictions of distributions p_S and q_S by replacing the summations over $i \in [n]$ with summations over $i \in S$.

We have the following relationships between these distances. These are well-known for distributions, i.e., see [GS02], but we prove them more generally for restrictions of distributions in Section 1.3.2.1.

Proposition 1. *Letting* p_S *and* q_S *be restrictions of distributions* p *and* q *to* $S \subseteq [n]$*,*

$$d_{\mathrm{H}}^2(p_S, q_S) \leq d_{\mathrm{TV}}(p_S, q_S) \leq \sqrt{2} d_{\mathrm{H}}(p_S, q_S) \leq \sqrt{\sum_{i \in S} (q_i - p_i) + d_{\mathrm{KL}}(p_S, q_S)} \leq \sqrt{d_{\chi^2}(p_S, q_S)}.$$

Furthermore, $d_{\mathrm{KL}}(p_S, q_S) \leq d_{\mathrm{SKL}}(p_S, q_S)$.

We recall that d_{ℓ_2} fits into the picture by its relationship with total variation distance:

Proposition 2. *Letting p and q be distributions over $[n]$,*

$$d_{\ell_2}(p,q) \le 2d_{\mathrm{TV}}(p,q) \le \sqrt{n} d_{\ell_2}(p,q).$$

The second inequality follows from Cauchy-Schwarz.

We will also need to following bound for Hellinger distance:

Proposition 3. $2d^2_{\mathrm{H}}(p,q) \le \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{p_i + q_i} \le 4d^2_{\mathrm{H}}(p,q)$.

Proof. Expanding the Hellinger-squared distance,

$$d^2_{\mathrm{H}}(p,q) = \frac{1}{2}\sum_{i=1}^{n}(\sqrt{p_i} - \sqrt{q_i})^2 = \frac{1}{2}\sum_{i=1}^{n}\frac{(p_i - q_i)^2}{(\sqrt{p_i} + \sqrt{q_i})^2}.$$

The fact now follows because $(p_i + q_i) \le (\sqrt{p_i} + \sqrt{q_i})^2 \le 2(p_i + q_i)$. □

The quantity $\sum_{i=1}^{n}(p_i - q_i)^2/(p_i + q_i)$ is sometimes called the *triangle distance*. However, we see here that it is essentially the Hellinger distance (up to constant factors).

1.3.2.1 Proof of Proposition 1

Recall that we will prove this for restrictions of probability distributions to subsets of the support – in other words, we do not assume $\sum_{i\in S} p_i = \sum_{i\in S} q_i = 1$, we only assume that $\sum_{i\in S} p_i \le 1$ and $\sum_{i\in S} q_i \le 1$.

$d^2_{\mathrm{H}}(p_S, q_S) \le d_{\mathrm{TV}}(p_S, q_S)$:

$$\begin{aligned} d^2_{\mathrm{H}}(p_S, q_S) &= \frac{1}{2}\sum_{i\in S}(\sqrt{p_i} - \sqrt{q_i})^2 \\ &\le \frac{1}{2}\sum_{i\in S}|\sqrt{p_i} - \sqrt{q_i}|(\sqrt{p_i} + \sqrt{q_i}) \\ &= \frac{1}{2}\sum_{i\in S}|p_i - q_i| \\ &= d_{\mathrm{TV}}(p_S, q_S). \end{aligned}$$

$d_{\mathrm{TV}}(p_S, q_S) \leq \sqrt{2} d_{\mathrm{H}}(p_S, q_S)$:

$$\begin{aligned}
d_{\mathrm{TV}}^2(p_S, q_S) &= \frac{1}{4}\left(\sum_{i\in S} |p_i - q_i|\right)^2 \\
&= \frac{1}{4}\left(\sum_{i\in S} |\sqrt{p_i} - \sqrt{q_i}|\,(\sqrt{p_i} + \sqrt{q_i})\right)^2 \\
&\leq \frac{1}{4}\left(\sum_{i\in S} |\sqrt{p_i} - \sqrt{q_i}|^2\right)\left(\sum_{i\in S} (\sqrt{p_i} + \sqrt{q_i})^2\right) \\
&\leq d_{\mathrm{H}}^2(p_S, q_S) \cdot \frac{1}{2}\left(\sum_{i\in S} (\sqrt{p_i} + \sqrt{q_i})^2\right) \\
&= d_{\mathrm{H}}^2(p_S, q_S) \cdot \left(\sum_{i\in S} p_i + \sum_{i\in S} q_i - d_{\mathrm{H}}^2(p_S, q_S)\right) \\
&\leq d_{\mathrm{H}}^2(p_S, q_S) \cdot \left(2 - d_{\mathrm{H}}^2(p_S, q_S)\right) \\
&\leq 2 d_{\mathrm{H}}^2(p_S, q_S).
\end{aligned}$$

Taking the square root of both sides gives the result. The second inequality is Cauchy-Schwarz.

$2d_{\mathrm{H}}^2(p_S, q_S) \leq \sum_{i\in S}(q_i - p_i) + d_{\mathrm{KL}}(p_S, q_S)$:

$$\begin{aligned}
2d_{\mathrm{H}}^2(p_S, q_S) &= \sum_{i\in S}(q_i + p_i) - 2\sum_{i\in S}\sqrt{p_i q_i} \\
&= \sum_{i\in S}(q_i + p_i) - 2\left(\left(\sum_{j\in S} p_j\right)\sum_{i\in S}\frac{p_i}{\sum_{j\in S} p_j}\sqrt{\frac{q_i}{p_i}}\right) \\
&\leq \sum_{i\in S}(q_i + p_i) - 2\left(\left(\sum_{j\in S} p_j\right)\exp\left(\frac{1}{2}\sum_{i\in S}\frac{p_i}{\sum_{j\in S} p_j}\log\frac{q_i}{p_i}\right)\right) \\
&\leq \sum_{i\in S}(q_i + p_i) - 2\left(\left(\sum_{j\in S} p_j\right)\left(1 + \frac{1}{2}\sum_{i\in S}\frac{p_i}{\sum_{j\in S} p_j}\log\frac{q_i}{p_i}\right)\right) \\
&= \sum_{i\in S}(q_i - p_i) - \left(\sum_{i\in S} p_i \log\frac{q_i}{p_i}\right) \\
&= \sum_{i\in S}(q_i - p_i) + d_{\mathrm{KL}}(p_S, q_S).
\end{aligned}$$

The first inequality is Jensen's, and the second is $1 + x \leq \exp(x)$.

$d_{\mathrm{KL}}(p_S, q_S) \leq \sum_{i \in S}(p_i - q_i) + d_{\chi^2}(p_S, q_S)$:

$$\begin{aligned}
d_{\mathrm{KL}}(p_S, q_S) &= \left(\sum_{j \in S} p_j\right)\left(\sum_{i \in S} \frac{p_i}{\sum_{j \in S} p_j} \log \frac{p_i}{q_i}\right) \\
&\leq \left(\sum_{j \in S} p_j\right)\left(\log \frac{1}{\sum_{j \in S} p_j} \sum_{i \in S} \frac{p_i^2}{q_i}\right) \\
&= \left(\sum_{j \in S} p_j\right)\left(\log\left(\frac{1}{\sum_{j \in S} p_j}\left(d_{\chi^2}(p_S, q_S) + 2\sum_{i \in S} p_i - \sum_{i \in S} q_i\right)\right)\right) \\
&= \left(\sum_{j \in S} p_j\right)\left(\log\left(2 + \frac{1}{\sum_{j \in S} p_j}\left(d_{\chi^2}(p_S, q_S) - \sum_{i \in S} q_i\right)\right)\right) \\
&\leq \left(\sum_{j \in S} p_j\right)\left(1 + \frac{1}{\sum_{j \in S} p_j}\left(d_{\chi^2}(p_S, q_S) - \sum_{i \in S} q_i\right)\right) \\
&= \sum_{i \in S}(p_i - q_i) + d_{\chi^2}(p_S, q_S).
\end{aligned}$$

The first inequality is Jensen's, and the second is $1 + x \leq \exp(x)$.

$d_{\mathrm{KL}}(p_S, q_S) \leq d_{\mathrm{SKL}}(p_S, q_S)$ This is immediate from non-negativity of KL divergence.

1.3.3 Convergence Bounds

One general purpose tool is the celebrated Dvoretzky-Kiefer-Wolfowitz (DKW) inequality. This gives a generic approach for learning an *arbitrary* distribution in Kolmogorov distance with only $O(1/\varepsilon^2)$ samples. This is in contrast to learning in total variation distance, which generally requires tailored methods for every distribution class of interest.

Lemma 1 ([DKW56],[Mas90])**.** *Let $\hat{p}_m$ be the empirical distribution generated by m i.i.d. samples from a distribution p. We have that*

$$\Pr[d_{\mathrm{K}}(p, \hat{p}_m) \geq \varepsilon] \leq 2e^{-2m\varepsilon^2}.$$

In particular, if $m = \Omega(\log(1/\delta)/\varepsilon^2)$, then $\Pr[d_{\mathrm{K}}(p, \hat{p}_m) \geq \varepsilon] \leq \delta$.

We will make extensive use of Chernoff-style bounds in this work. Recall that the Binomial(n, p) distribution describes the distribution of the number of successes when we run n independent Bernoulli trials, each with success probability p.

Lemma 2 (Chernoff Bound for Binomials)**.** *Let* $X \sim \text{Binomial}(n, p)$ *and* $\mu = \mathbf{E}[X] = np$*. Then*

$$\forall \delta \in [0, 1), \qquad \Pr[\,|X - \mu| \geq \delta\mu\,] \leq 2 \exp\left(-\frac{\delta^2 \mu}{3}\right).$$

We will also need a similar Chernoff-style bound for the *hypergeometric* distribution. The Hypergeometric(n, K, N) distribution describes the distribution of the number of successes when we draw n times without replacement from a population of size N, in which K objects have the pertinent feature (and thus count as successes). Note that if the drawing were done with replacement, and $K/N = p$, then this would be equivalent to Binomial(n, p). Sampling without replacement introduces *negative correlation* between the probability of each draw being successful. This type of negative correlation generally "helps" with concentration, allowing one to prove similar concentration bounds (see, e.g., [Chv79, DR96], Theorem 1.17 of [AD11]).

Lemma 3 (Chernoff Bound for Hypergeometrics)**.** *Let* $X \sim \text{Hypergeometric}(n, K, N)$ *and* $\mu = \mathbf{E}[X] = nK/N$*. Then,*

$$\forall \delta \in [0, 1), \qquad \Pr[\,|X - \mu| \geq \delta\mu\,] \leq 2 \exp\left(-\frac{\delta^2 \mu}{3}\right).$$

1.3.4 Poisson Sampling

At certain points, our algorithms will employ Poisson sampling. Rather than taking a fixed number of m samples from a distribution p, we instead draw Poisson(m) samples. More precisely, we first sample $m' \sim \text{Poisson}(m)$, and then draw m' samples from p. While this procedure might seem odd and indirect, it has a perhaps surprising benefit. Namely, letting N_i be the number of occurrences of element i, all N_i will be independent and distributed as Poisson$(m \cdot p_i)$. This is in contrast to the standard sampling procedure: the N_i's will be marginally distributed as Binomial(m, p_i), but there will exist significant correlations. Injecting this independence over the N_i's makes the analysis significantly easier in certain

cases, as we can focus on individual symbols without worrying about correlations. One concern could be that m' is much larger than m, which would significantly increase the sample complexity. However, since Poisson(m) is tightly concentrated around its mean m, we have that $m' = \Theta(m)$ with high probability. Therefore, any algorithm which draws Poisson(m) samples can be converted to an algorithm with a fixed budget of (say) $10m$ samples. The first step would be to draw $m' \sim$ Poisson(m): if $m' \leq 10m$, use that many samples and discard the rest. On the other hand, if $m' > 10m$ (which occurs with negligible probability), then the output can be arbitrary.

Chapter 2

Testing with Tolerance and Alternative Distances

2.1 Introduction

Up to this point, most of the discussion on the problem of testing whether p is equal to some hypothesis q, or is far in total variation distance from q. In this chapter, we focus on relaxing both of these restrictions. In particular, the two core problems will be the following:

1. Can we handle when p is *close* to q, rather than equal?

2. Can we test for when p and q are far in *other distances* besides total variation?

To give an example of why these type of issues may arise, we give a concrete example. Suppose we want to test whether the sizes of some population of insects are normally distributed around their mean by sampling insects and measuring their sizes. Of course, our models are usually imperfect. In our insect example, perhaps our estimation of the mean and variance of the insect sizes is a bit off. Furthermore, the sizes will clearly always be positive numbers. Yet a Normal distribution could still be a good fit. To get a meaningful testing problem some slack may be introduced, turning the problem into that of distinguishing whether $d_1(p, q) \leq \varepsilon_1$ versus $d_2(p, q) \geq \varepsilon_2$, for some distance measures $d_1(\cdot, \cdot)$ and $d_2(\cdot, \cdot)$ between distributions over $[n]$ and some choice of ε_1 and ε_2 which may potentially depend on $[n]$ or even q. Regardless, for the problem to be well-defined, the sets of distributions

$\mathcal{C} = \{p \mid d_1(p,q) \leq \varepsilon_1\}$ and $\mathcal{F} = \{p \mid d_2(p,q) \geq \varepsilon_2\}$ should be disjoint. In fact, as our goal is to distinguish between $p \in \mathcal{C}$ and $p \in \mathcal{F}$ from samples, we cannot possibly draw the right conclusion with probability 1 or detect the most minute deviations of p from $\mathcal{C}$ or $\mathcal{F}$. So our guarantee should be probabilistic, and there should be some "gap" between the sets $\mathcal{C}$ and $\mathcal{F}$. In sum, the problem is the following:

> (d_1, d_2)-*Identity Testing*: Given an explicit description of a distribution q over $[n]$, sample access to a distribution p over $[n]$, and bounds $\varepsilon_1 \geq 0$, and $\varepsilon_2, \delta > 0$, distinguish with probability at least $1 - \delta$ between $d_1(p,q) \leq \varepsilon_1$ and $d_2(p,q) \geq \varepsilon_2$, whenever p satisfies one of these two inequalities.

A related problem is when we have sample access to both p and q. For example, we might be interested in whether two populations of insects have distributions that are close or far. The resulting problem is the following:

> (d_1, d_2)-*Equivalence (or Closeness) Testing*: Given sample access to distributions p and q over $[n]$, and bounds $\varepsilon_1 \geq 0$, and $\varepsilon_2, \delta > 0$, distinguish with probability at least $1 - \delta$ between $d_1(p,q) \leq \varepsilon_1$ and $d_2(p,q) \geq \varepsilon_2$, whenever p, q satisfy one of these two inequalities.

As mentioned before, the primary focus of prior work has been on the case where $\varepsilon_1 = 0$ and d_2 is the total variation distance. There are several other sub-optimal results known for various combinations of d_1, d_2, ε_1 and ε_2, and for many combinations there are no known testers. A more extensive discussion of the literature is provided in Section 2.1.2.

The goal of this chapter is to *provide a complete mapping of the optimal sample complexity required to obtain computationally efficient testers for identity testing and equivalence testing under the most commonly used notions of distances d_1 and d_2.* Our results are summarized in Tables 2.1, 2.2, and 2.3 and discussed in detail in Section 2.1.1. In particular, we obtain computationally efficient and sample optimal testers for distances d_1 and d_2 ranging in the set $\{\ell_2$-distance, total variation distance, Hellinger distance, Kullback-Leibler divergence, χ^2-divergence$\}$,[1] and for combinations of these distances and choice of errors ε_1 and ε_2 which

[1] These distances are nicely nested, as discussed in Section 1.3.2, from the weaker ℓ_2 to the stronger χ^2-divergence.

give rise to meaningful testing problems as discussed above. The sample complexities stated in the tables are for probability of error 1/3. Throwing in extra factors of $O(\log 1/\delta)$ boosts the probability of error to $1 - \delta$, as usual.

Our motivation for this work is primarily the fundamental nature of identity and equivalence testing, as well as of the distances under which we study these problems. It is also the fact that, even though distribution testing is by now a mature subfield of information theory, property testing, and sublinear-time algorithms, several of the testing questions that we consider have had unknown statuses prior to our work. This gap is accentuated by the fact that, as we establish, closely related distances may have radically different behavior. To give a quick example, it is easy to see that χ^2-divergence is the second-order Taylor expansion of KL-divergence. Yet, as we show, the sample complexity for identity testing changes radically when d_2 is taken to be total variation or Hellinger distance, and d_1 transitions from χ^2 to KL or weaker distances; see Table 2.1. Similar fragility phenomena are identified by our work for equivalence testing, when we switch from total variation to Hellinger distance, as seen in Tables 2.2 and 2.3.

Adding to the fundamental nature of the problems we consider here, we should also emphasize that a clear understanding of the different tradeoffs mapped out by our work is critical at this point for the further development of the distribution testing field, as recent experience has established. We provide a couple of recent examples where testing with tolerance and alternative distances proves to be critical. One application in [ADK15] is that of *composite* hypothesis testing, where we wish to test if p belongs to some *class* of distributions. For instance, one could ask if the density of p is monotone increasing. It turns out that testing with d_1 being the χ^2-divergence is crucial for this application. This work is the focus of Chapter 3, and we defer further discussion to then.

Another example supporting our expectation can be found in recent work of Daskalakis and Pan [DP17]. They study equivalence testing of Bayesian networks under total variation distance. Bayesian networks are flexible models expressing combinatorial structure in high-dimensional distributions in terms of a directed acyclic graph (DAG) specifying their conditional dependence structure. The challenge in testing Bayes nets is that their support scales exponentially in the number of nodes, and hence naive applications of known

equivalence tests lead to sample complexities that are exponential in the number of nodes, even when the in-degree δ of the underlying DAGs is bounded. To address this challenge, Daskalakis and Pan establish "localization-of-distance" results of the following form, for various choices of distance d: "If two Bayes nets p and q are ε-far in total variation distance, then there exists a small set of nodes S (whose size is $\Delta + 1$, where Δ is again the maximum in-degree of the underlying DAG where p and q are defined) such that the marginal distributions of p and q over the nodes of set S are ε'-far under distance d." When they take d to be total variation distance, they can show $\varepsilon' = \Omega(\varepsilon/m)$, where m is the number of nodes in the underlying DAG (i.e. the dimension). Given this localization of distance, to test whether two Bayes nets p and q satisfy $p = q$ versus $d_{\mathrm{TV}}(p, q) \geq \varepsilon$, it suffices to test, for all relevant marginals p_S and q_S whether $p_S = q_S$ versus $d_{\mathrm{TV}}(p_S, q_S) = \Omega(\varepsilon/m)$. From Table 2.2 it follows that this requires sample size superlinear in m, which is suboptimal. Interestingly, when they take d to be the square Hellinger distance, they can establish a localization-of-distance result with $\varepsilon' = \varepsilon^2/2m$. By Table 2.2, to test each S they need sample complexity that is linear in m, leading to an overall dependence of the sample complexity on m that is $\tilde{O}(m)$,[2] which is optimal up to log factors. More recent applications of our algorithm for identity testing for Hellinger distance include [DDG18, ABDK18]. Again, switching to a different distance results in near-optimal overall sample complexity, and our table is guidance as to where the bottlenecks and opportunities lie.

Finally, we comment that tolerant testing (i.e., when $\varepsilon_1 > 0$) is perhaps one of the most interesting questions in the design of practically useful testers. Indeed, as mentioned before, in many statistical settings there may be model misspecification. For example, why should one expect to be receiving samples from *precisely* the uniform distribution? As such, one may desire that a tester is *robust* to small errors, and accepts all distributions which are *close* to uniform. Unfortunately, Valiant and Valiant [VV11a] ruled out the possibility of a strongly sublinear tester which has total variation tolerance, showing that such a problem requires $\Theta\left(\frac{n}{\log n}\right)$ samples. This raises the following question: Which distances can a tester be tolerant to, while maintaining a strongly sublinear sample complexity? We outline what is possible.

[2]The extra log factors are to guarantee that the tests performed on all sets S of size $\delta + 1$ succeed.

	$d_{\mathrm{TV}}(p,q) \geq \varepsilon$	$d_{\mathrm{H}}(p,q) \geq \varepsilon/\sqrt{2}$	$d_{\mathrm{KL}}(p,q) \geq \varepsilon^2$	$d_{\chi^2}(p,q) \geq \varepsilon^2$
$p = q$	$\Omega\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$[Pan08]		Untestable [Theorem 10]	
$d_{\chi^2}(p,q) \leq \varepsilon^2/4$		$O\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$[Theorem 4]		
$d_{\mathrm{KL}}(p,q) \leq \varepsilon^2/4$	$\Omega\left(\frac{n}{\log n}\right)$[Theorem 11]			
$d_{\mathrm{H}}(p,q) \leq \varepsilon/2\sqrt{2}$				
$d_{\mathrm{TV}}(p,q) \leq \varepsilon/2$ or $\varepsilon^2/4$[4]		$O\left(\frac{n}{\log n}\right)$[Corollary 3]		

Table 2.1: Identity Testing. Rows correspond to completeness of the tester, and columns correspond to soundness.

2.1.1 Results

Our results are pictorially presented in Tables 2.1, 2.2, and 2.3. We note that these tables are intended to provide only references to the *sample complexity* of each testing problem, rather than exhaustively cover all prior work. As such, several references are deferred to Section 2.1.2. In Tables 2.1 and 2.2, each cell contains the complexity of testing whether two distributions are close in the distance for that row, versus far in the distance for that column.[3] These distances and their relationships are covered in detail in Section 2.2, but we note that the distances are scaled and transformed such that problems become harder as we traverse the table down or to the right. In other words, lower bounds hold for cells which are down or to the right in the table, and upper bounds hold for cells which are up or to the left; problems with the same complexity are shaded with the same color. The dark grey boxes indicate problems which are not well-defined, i.e. two distributions could simultaneously be close in KL and far in χ^2-divergence.

We highlight some of our results:

1. We give an $O(\sqrt{n}/\varepsilon^2)$ sample algorithm for identity testing whether $d_{\chi^2}(p,q) \leq \varepsilon^2/4$ or $d_{\mathrm{H}}(p,q) \geq \varepsilon/\sqrt{2}$ (Theorem 4). This is the first algorithm which achieves the optimal dependence on both n and ε for identity testing with respect to Hellinger distance (even non-tolerantly). We note that a $O(\sqrt{n}/\varepsilon^4)$ algorithm was known, due to optimal identity testers for total variation distance and the quadratic relationship between total

[3]Note that we chose constants in our theorem statements for simplicity of presentation, and they may not match the constants presented in the table. This can be remedied by appropriate changing of constants in the algorithms and constant factor increases in the sample complexity.

[4]We note that we must use $\varepsilon/2$ or $\varepsilon^2/4$ depending on whether we are testing with respect to TV or Hellinger. For more details and other discussion of the $n/\log n$ region of this chart, see Section 2.1.1.2.

	$d_{\mathrm{TV}}(p,q) \geq \varepsilon$	$d_{\mathrm{H}}(p,q) \geq \varepsilon/\sqrt{2}$	$d_{\mathrm{KL}}(p,q) \geq \varepsilon^2$	$d_{\chi^2}(p,q) \geq \varepsilon^2$
$p = q$	$O\left(\max\left\{\frac{n^{1/2}}{\varepsilon^2}, \frac{n^{2/3}}{\varepsilon^{4/3}}\right\}\right)$[CDVV14] $\Omega\left(\max\left\{\frac{n^{1/2}}{\varepsilon^2}, \frac{n^{2/3}}{\varepsilon^{4/3}}\right\}\right)$[CDVV14]	$O\left(\min\left\{\frac{n^{3/4}}{\varepsilon^2}, \frac{n^{2/3}}{\varepsilon^{8/3}}\right\}\right)$[Theorem 8] $\Omega\left(\min\left\{\frac{n^{3/4}}{\varepsilon^2}, \frac{n^{2/3}}{\varepsilon^{8/3}}\right\}\right)$[DK16]	Untestable [Theorem 10]	
$d_{\chi^2}(p,q) \leq \varepsilon^2/4$	$\Omega\left(\frac{n}{\log n}\right)$ [Theorem 12]			
$d_{\mathrm{KL}}(p,q) \leq \varepsilon^2/4$				
$d_{\mathrm{H}}(p,q) \leq \varepsilon/2\sqrt{2}$				
$d_{\mathrm{TV}}(p,q) \leq \varepsilon/2$ or $\varepsilon^2/4^4$		$O\left(\frac{n}{\log n}\right)$[Corollary 3]		

Table 2.2: Equivalence Testing. Rows correspond to completeness of the tester, and columns correspond to soundness.

	Identity Testing	Equivalence Testing
$d(p,q) \leq f_d(n,\varepsilon)$ vs. $d_{\ell_2}(p,q) \geq \varepsilon$	$\Theta\left(\frac{1}{\varepsilon^2}\right)$ [Corollary 2]	$\Theta\left(\frac{1}{\varepsilon^2}\right)$ [Corollary 2]
$d_{\ell_2}(p,q) \leq \frac{\varepsilon}{\sqrt{n}}$ vs. $d_{\mathrm{TV}}(p,q) \geq \varepsilon$	$\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [Theorem 5]	$\Theta\left(\max\left\{\frac{n^{1/2}}{\varepsilon^2}, \frac{n^{2/3}}{\varepsilon^{4/3}}\right\}\right)$ [Theorem 7]
$d_{\ell_2}(p,q) \leq \frac{\varepsilon^2}{\sqrt{n}}$ vs. $d_{\mathrm{H}}(p,q) \geq \varepsilon$	$\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [Theorem 6]	$\Theta\left(\min\left\{\frac{n^{3/4}}{\varepsilon^2}, \frac{n^{2/3}}{\varepsilon^{8/3}}\right\}\right)$[Theorem 8]

Table 2.3: ℓ_2 Testing. $f_d(n,\varepsilon)$ is a quantity such that $d(p,q) \leq f_d(n,\varepsilon)$ and $d_{\ell_2}(p,q) \geq \varepsilon$ are disjoint.

variation and Hellinger distance. Note that this immediately implies an algorithm with the same sample complexity for identity testing whether $d_{\chi^2}(p,q) \leq \varepsilon^2/4$ or $d_{\mathrm{TV}}(p,q) \geq \varepsilon$, which was also shown in [ADK15].

2. In the case of identity testing, a stronger form of tolerance (i.e., KL divergence instead of χ^2) causes the sample complexity to jump to $\Omega(n/\log n)$ (Theorem 11). We find this a bit surprising, as χ^2-divergence is the second-order Taylor expansion of KL divergence, so one might expect that the testing problems have comparable complexities.

3. In the case of equivalence testing, *even* χ^2*-tolerance* comes at the cost of an $\Omega(n/\log n)$ sample complexity (Theorem 12). This is a qualitative difference from identity testing, where χ^2-tolerance came at no cost.

4. However, in both identity and equivalence testing, ℓ_2 tolerance comes at no additional cost (Theorems 5, 6, 7, and 8). Thus, in many cases, ℓ_2 tolerance is the best one can do if one wishes to maintain a strongly sublinear sample complexity.

From a technical standpoint, our algorithms are χ^2-statistical tests, and most closely resemble those of [ADK15] and [CDVV14] (similar χ^2-tests were employed in [VV17a, DKN15b,

CDGR16]). However, crucial changes are required to satisfy the more stringent requirements of testing with respect to Hellinger distance. In our identity tester for Hellinger, we deal with this different distance measure by pruning light domain elements of q less aggressively than in the total variation tester of [ADK15], in combination with a preliminary test to reject early if the difference between p and q is contained exclusively within the set of light elements – this is a new issue that cannot arise when testing with respect to total variation distance. In our equivalence tester for Hellinger, we follow an approach, similar to [CDVV14] and [DK16], of analyzing the light and heavy domain elements separately, with the challenge that the algorithm does not know which elements are which. Finally, to achieve ℓ_2 tolerance in these cases, we use a "mixing" strategy in which instead of testing based solely on samples from p and q, we mix in some number (depending on our application) of samples from the uniform distribution. At a high level, the purpose of mixing is to make our distributions *well-conditioned*, i.e. to ensure that all probability values are sufficiently large. Such a strategy was recently employed by Goldreich in [Gol16] for uniformity testing.

2.1.1.1 Comments on ℓ_2-tolerance

ℓ_2 tolerance has been indirectly considered in [GR00, BFF$^+$01, BFR$^+$13] through their weak tolerance for total variation distance and the relationship with ℓ_2 distance, though these results have suboptimal sample complexity. Our equivalence testing results improve upon [CDVV14] by adding ℓ_2-tolerance. We note that [DK16] also provides ℓ_2-tolerant testers (as well as [DKN15b] for the case of uniformity), comparable to those obtained in Theorems 5, 6, and 8, though this tolerance is not explicitly analyzed in their paper. This can be seen by noting that the underlying tester from [CDVV14] is tolerant, and the "flattening" operation they apply reduces the ℓ_2-distance between the distributions. The testers in [DK16] are those of Propositions 2.7, 2.10, and 2.15, combined with the observation of Remark 2.8. We rederive these results for completeness, and to show a direct way of proving ℓ_2-tolerance. Note that Theorem 8 also improves upon Proposition 2.15 of [DK16] by removing log factors in the sample complexity.

2.1.1.2 Comments on the $\Theta(n/\log n)$ Results

Our upper bounds in the bottom-left portion of the table are based off the total variation distance estimation algorithm of Jiao, Han, and Weissman [JHW16], where an $\Theta(n/\log n)$ complexity is only derived for $\varepsilon \geq 1/\operatorname{poly}(n)$. Similarly, in [VV10a], the lower bounds are only valid for constant ε. We believe that the precise characterization is a very interesting open problem. In the present work, we focus on the case of constant ε for these testing problems.

We wish to draw attention to the bottom row of the table, and note that the two testing problems are $d_{\mathrm{TV}}(p,q) \leq \varepsilon/2$ versus $d_{\mathrm{TV}}(p,q) \geq \varepsilon$, and $d_{\mathrm{TV}}(p,q) \leq \varepsilon^2/4$ versus $d_{\mathrm{H}}(p,q) \geq \varepsilon/\sqrt{2}$. This difference in parameterization is required to make the two cases in the testing problem disjoint. With this parameterization, we conjecture that the latter problem has a greater dependence on ε as it goes to 0 (namely, ε^{-4} versus ε^{-2}), so we colour the box a slightly darker shade of orange.

2.1.2 Related Work

[Wag15, DBNNR11, GMV06] also consider testing problems with other distances, namely ℓ_p distances, earth mover's distance (also known as Wasserstein distance), and various f-divergences.

Tolerant identity testing (where $\varepsilon_1 = O(\varepsilon)$ and d_1 is total variation distance) was studied in [VV10a, VV10b, VV11a, VV11b], through the (equivalent) lens of estimating total variation distance between distributions. In these works, $\Theta\left(n/\log n\right)$ bounds were proven for the sample complexity. Several other related problems (e.g., support size and entropy estimation) share the same sample complexity, and have enjoyed significant study [AOST17, WY16, ADOS17]. The closest related results to our work are those on estimating distances between distributions [JHW16, JVHW17, HJW16].

χ^2-tolerance (when d_1 is χ^2-divergence and $\varepsilon_1 = O(\varepsilon^2)$) was introduced and applied by [ADK15] for testing families of distributions, e.g., testing if a distribution is monotone or far from being monotone. This result will be discussed more in Chapter 3.

Testing with respect to Hellinger distance was applied in [DP17, ABDK18] for testing

Bayes networks, and [DDG18] for testing Markov chains. Due to tensorization properties, Hellinger distance is more natural for testing problems in some multivariate settings, and we believe it will arise more frequently as this new direction matures. Prior to our work, Hellinger testing was studied for equivalence testing in [DK16].

Most of our tests in this work are based around χ^2-statistics. We note that the χ^2-statistic for testing hypothesis is prevalent in statistics providing optimal error exponents in the large-sample regime. A similar modification to the χ^2-statistic as we use here (i.e., subtracting the count N_i from $(N_i - mq_i)^2$) was previously used in [Zel87]. To the best of our knowledge, in the small-sample regime, *modified-versions* of the χ^2-statistic have only been used somewhat recently. Some instances include equivalence testing in [ADJ$^+$12, CDVV14], uniformity testing of monotone distributions in [AJOS13], and identity testing in [DKN15b, VV17a]. The latter two papers also apply subtraction modifications, similar to our work and [Zel87]. The statistic of [ADJ$^+$12] is an unbiased statistic for estimating the χ^2-distance between two *unknown* distributions.

2.1.3 Organization

The organization of this chapter is as follows. In Section 2.2, we state preliminaries and notation used in this chapter. In Sections 2.3 and 2.4, we prove upper bounds for identity testing and equivalence testing (respectively) based on χ^2-style statistics. In Section 2.5, we prove upper bounds for distribution testing based on distance estimation. Finally, in Section 2.6, we prove testing lower bounds.

2.2 Preliminaries

Proposition 4. *Given a number $\delta \in [0,1]$ and a discrete distribution $r = (r_1, \ldots, r_n)$, define*

$$r^{+\delta} := (1-\delta) \cdot r + \delta \cdot (\tfrac{1}{n}, \ldots, \tfrac{1}{n}).$$

Then given two discrete distributions $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$,

$$d_{\mathrm{TV}}(p^{+\delta}, q^{+\delta}) = (1-\delta) d_{\mathrm{TV}}(p, q), \quad d_{\ell_2}(p^{+\delta}, q^{+\delta}) = (1-\delta) d_{\ell_2}(p, q).$$

In addition, $d_{\mathrm{H}}(p^{+\delta}, q^{+\delta}) \geq d_{\mathrm{H}}(p, q) - 2\sqrt{\delta}$.

Proof. The statements for total variation and ℓ_2 distance are immediate. As for the Hellinger distance, we have by the triangle inequality that

$$d_{\mathrm{H}}(p, q) \leq d_{\mathrm{H}}(p, p^{+\delta}) + d_{\mathrm{H}}(p^{+\delta}, q^{+\delta}) + d_{\mathrm{H}}(q^{+\delta}, q).$$

We can bound the first term by

$$d_{\mathrm{H}}^2(p, p^{+\delta}) \leq d_{\mathrm{TV}}(p, p^{+\delta}) = \tfrac{1}{2} \cdot \|\delta \cdot p - \delta \cdot (\tfrac{1}{n}, \ldots, \tfrac{1}{n})\|_1 \leq \delta,$$

where the last step is by the triangle inequality, and a similar argument bounds the third term by $\sqrt{\delta}$ as well. Thus, $d_{\mathrm{H}}(p^{+\delta}, q^{+\delta}) \geq d_{\mathrm{H}}(p, q) - 2\sqrt{\delta}$. □

A similar technique was employed in [Gol16].

2.3 Upper Bounds for Identity Testing

In this section, we prove the following theorems for identity testing.

Theorem 4. *There exists an algorithm for identity testing between p and q distinguishing the cases:*

- $d_{\chi^2}(p, q) \leq \varepsilon^2$;
- $d_{\mathrm{H}}(p, q) \geq \varepsilon$.

The algorithm uses $O\left(\frac{n^{1/2}}{\varepsilon^2}\right)$ samples.

Theorem 5. *There exists an algorithm for identity testing between p and q distinguishing the cases:*

- $d_{\ell_2}(p,q) \le \frac{\varepsilon}{\sqrt{n}}$;
- $d_{\mathrm{TV}}(p,q) \ge \varepsilon$.

The algorithm uses $O\left(\frac{n^{1/2}}{\varepsilon^2}\right)$ *samples.*

Theorem 6. *There exists an algorithm for identity testing between* p *and* q *distinguishing the cases:*

- $d_{\ell_2}(p,q) \le \frac{\varepsilon^2}{\sqrt{n}}$;
- $d_{\mathrm{H}}(p,q) \ge \varepsilon$.

The algorithm uses $O\left(\frac{n^{1/2}}{\varepsilon^2}\right)$ *samples.*

We prove Theorem 4 in Section 2.3.1, and Theorems 5 and 6 in Section 2.3.2.

2.3.1 Identity Testing with Hellinger Distance and χ^2-Tolerance

We prove Theorem 4 by analyzing Algorithm 1. We will set $c_1 = \frac{1}{100}, c_2 = \frac{6}{25}$, and let C be a sufficiently large constant.

Algorithm 1 χ^2-close versus Hellinger-far testing algorithm

Input: ε; an explicit distribution q; sample access to a distribution p
Implicitly define $\mathcal{A} \leftarrow \{i : q_i \ge c_1\varepsilon^2/n\}$, $\bar{\mathcal{A}} \leftarrow [n] \setminus \mathcal{A}$
Let $\hat{p}$ be the empirical distribution[5] from drawing $m_1 = \Theta(1/\varepsilon^2)$ samples from p
if $\hat{p}(\bar{\mathcal{A}}) \ge \frac{3}{4}c_2\varepsilon^2$ **then**
    return REJECT
end if
Draw a multiset S of Poisson(m_2) samples from p, where $m_2 = C\sqrt{n}/\varepsilon^2$
Let N_i be the number of occurrences of the ith domain element in S
Let S' be the set of domain elements observed in S
$Z \leftarrow \sum_{i \in S' \cap \mathcal{A}} \frac{(N_i - m_2 q_i)^2 - N_i}{m_2 q_i} + m_2(1 - q(S' \cap \mathcal{A}))$
if $Z \le \frac{3}{2}m_2\varepsilon^2$ **then**
    return ACCEPT
else
    return REJECT
end if

We note that the sample and time complexity are both $O(\sqrt{n}/\varepsilon^2)$. We draw $m_1 + m_2 = \Theta(\sqrt{n}/\varepsilon^2)$ samples total. All steps of the algorithm only involve inspecting domain elements

where a sample falls, and it runs linearly in the number of such elements. Indeed, Step 10 of the algorithm is written in an unusual way in order to ensure the running time of the algorithm is linear.

We first analyze the test in Step 4 of the algorithm. Folklore results state that with probability at least 99/100, this preliminary test will reject any p with $p(\bar{\mathcal{A}}) \geq c_2\varepsilon^2$, it will not reject any p with $p(\bar{\mathcal{A}}) \leq \frac{c_2}{2}\varepsilon^2$, and behavior for any other p is arbitrary. Condition on the event the test does not reject for the remainder of the proof. Note that since both thresholds here are $\Theta(\varepsilon^2)$, it only requires $m_1 = \Theta(1/\varepsilon^2)$ samples, rather than the "non-extreme" regime, where we would require $\Theta(1/\varepsilon^4)$ samples.

Remark 1. *We informally refer to this "extreme" versus "non-extreme" regime in distribution testing. To give an example of what we mean in these two cases, consider distinguishing $Ber(1/2)$ from $Ber(1/2+\varepsilon)$. The complexity of this problem is $\Theta(1/\varepsilon^2)$, and we consider this to be in the non-extreme regime. On the other hand, distinguishing $Ber(\varepsilon)$ from $Ber(2\varepsilon)$ has a sample complexity of $\Theta(1/\varepsilon)$, and we consider this to be in the extreme regime.*

We justify that any p which may be rejected in Step 5 (i.e., any p such that $p(\bar{\mathcal{A}}) > \frac{c_2}{2}\varepsilon^2$) has the property that $d_{\chi^2}(p,q) > \varepsilon^2$ (in other words, we do not wrongfully reject any p).

Consider a p such that $p(\bar{\mathcal{A}}) \geq \frac{c_2}{2}\varepsilon^2$. Note that $d_{\chi^2}(p,q) \geq d_{\chi^2}(p_{\bar{\mathcal{A}}}, q_{\bar{\mathcal{A}}})$, which we lower bound as follows:

$$\begin{aligned}
d_{\chi^2}(p_{\bar{\mathcal{A}}}, q_{\bar{\mathcal{A}}}) &= \sum_{i\in\bar{\mathcal{A}}} \frac{(p_i - q_i)^2}{q_i} \\
&\geq \frac{n}{c_1\varepsilon^2} \sum_{i\in\bar{\mathcal{A}}} (p_i - q_i)^2 \\
&\geq \frac{n}{c_1\varepsilon^2} \cdot \frac{1}{n} \left(\sum_{i\in\bar{\mathcal{A}}} (p_i - q_i) \right)^2 \\
&\geq \frac{n}{c_1\varepsilon^2} \frac{\varepsilon^4 \left(\frac{c_2}{2} - c_1\right)^2}{n} \\
&= \frac{\left(\frac{c_2}{2} - c_1\right)^2}{c_1} \varepsilon^2
\end{aligned}$$

The first inequality is by the definition of $\bar{\mathcal{A}}$, the second is by Cauchy-Schwarz, and the third is since $p(\bar{\mathcal{A}}) \geq \frac{c_2}{2}\varepsilon^2$ and $q(\bar{\mathcal{A}}) \leq c_1\varepsilon^2$. By our setting of c_1 and c_2, this implies that

$d_{\chi^2}(p,q) > \varepsilon^2$, and we are not rejecting any p which should be accepted.

For the remainder of the proof, we will implicitly assume that $p(\bar{\mathcal{A}}) \leq c_2\varepsilon^2$.

Let

$$Z' = \sum_{i \in \mathcal{A}} \frac{(N_i - m_2 q_i)^2 - N_i}{m_2 q_i}.$$

Note that the statistic Z can be rewritten as follows:

$$\begin{aligned}
Z &= \sum_{i \in S' \cap \mathcal{A}} \frac{(N_i - m_2 q_i)^2 - N_i}{m_2 q_i} + m_2(1 - q(S' \cap \mathcal{A})) \\
&= \sum_{i \in S' \cap \mathcal{A}} \frac{(N_i - m_2 q_i)^2 - N_i}{m_2 q_i} + \sum_{i \in \mathcal{A} \setminus S'} m_2 q_i + m_2 q(\bar{\mathcal{A}}) \\
&= \sum_{i \in S' \cap \mathcal{A}} \frac{(N_i - m_2 q_i)^2 - N_i}{m_2 q_i} + \sum_{i \in \mathcal{A} \setminus S'} \frac{(N_i - m_2 q_i)^2 - N_i}{m_2 q_i} + m_2 q(\bar{\mathcal{A}}) \\
&= Z' + m_2 q(\bar{\mathcal{A}})
\end{aligned}$$

We proceed by analyzing Z'. First, note that it has the following expectation and variance:

Proposition 5.

$$\mathbf{E}[Z'] = m_2 \cdot \sum_{i \in \mathcal{A}} \frac{(p_i - q_i)^2}{q_i} = m_2 \cdot d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) \tag{2.1}$$

$$\mathbf{Var}[Z'] = \sum_{i \in \mathcal{A}} \left[2\frac{p_i^2}{q_i^2} + 4m_2 \cdot \frac{p_i \cdot (p_i - q_i)^2}{q_i^2} \right] \tag{2.2}$$

Proof. We start by analyzing the mean:

$$\begin{aligned}
\mathbf{E}\left[Z'\right] &= \sum_{i\in\mathcal{A}} \mathbf{E}\left[\frac{(N_i - m_2 q_i)^2 - N_i}{m_2 q_i}\right] \\
&= \sum_{i\in\mathcal{A}} \frac{\mathbf{E}\left[N_i^2\right] - 2m_2 q_i \mathbf{E}\left[N_i\right] + m_2^2 q_i^2 - \mathbf{E}\left[N_i\right]}{m_2 q_i} \\
&= \sum_{i\in\mathcal{A}} \frac{m_2^2 p_i^2 + m_2 p_i - 2m_2^2 q_i p_i + m_2^2 q_i^2 - m_2 p_i}{m_2 q_i} \\
&= m_2 \sum_{i\in\mathcal{A}} \frac{(p_i - q_i)^2}{q_i} \\
&= m_2 \cdot d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}})
\end{aligned}$$

Next, we analyze the variance. Let $\lambda_i = \mathbf{E}\left[N_i\right] = m_2 p_i$ and $\lambda_i' = m_2 q_i$.

$$\begin{aligned}
\mathbf{Var}[Z'] &= \sum_{i\in\mathcal{A}} \frac{1}{\lambda_i'^2} \mathbf{Var}(N_i - \lambda_i)^2 + 2(N_i - \lambda_i)(\lambda_i - \lambda_i') - (N_i - \lambda_i) \\
&= \sum_{i\in\mathcal{A}} \frac{1}{\lambda_i'^2} \mathbf{Var}(N_i - \lambda_i)^2 + (N_i - \lambda_i)(2\lambda_i - 2\lambda_i' - 1) \\
&= \sum_{i\in\mathcal{A}} \frac{1}{\lambda_i'^2} \mathbf{E}\left[(N_i - \lambda_i)^4 + 2(N_i - \lambda_i)^3(2\lambda_i - 2\lambda_i' - 1) + (N_i - \lambda_i)^2(2\lambda_i - 2\lambda_i' - 1)^2 - \lambda_i^2\right] \\
&= \sum_{i\in\mathcal{A}} \frac{1}{\lambda_i'^2} [3\lambda_i^2 + \lambda_i + 2\lambda_i(2\lambda_i - 2\lambda_i' - 1) + \lambda_i(2\lambda_i - 2\lambda_i' - 1)^2 - \lambda_i^2] \\
&= \sum_{i\in\mathcal{A}} \frac{1}{\lambda_i'^2} [2\lambda_i^2 + \lambda_i + 4\lambda_i(\lambda_i - \lambda_i') - 2\lambda_i + \lambda_i(4(\lambda_i - \lambda_i')^2 - 4(\lambda_i - \lambda_i') + 1)] \\
&= \sum_{i\in\mathcal{A}} \frac{1}{\lambda_i'^2} [2\lambda_i^2 + 4\lambda_i(\lambda_i - \lambda_i')^2] \\
&= \sum_{i\in\mathcal{A}} \left[2\frac{p_i^2}{q_i^2} + 4m_2 \cdot \frac{p_i \cdot (p_i - q_i)^2}{q_i^2}\right] \qquad (2.3)
\end{aligned}$$

The third equality is by noting the random variable has expectation λ_i and the fourth equality substitutes the values of centralized moments of the Poisson distribution. □

We require the following two lemmas, which state that the mean of the statistic is separated in the two cases, and that the variance is bounded. The proofs largely follow the proofs of two similar lemmas in [ADK15].

Lemma 4. *If $d_{\chi^2}(p,q) \le \varepsilon^2$, then $\mathbf{E}[Z'] \le m_2\varepsilon^2$. If $d_{\mathrm{H}}(p,q) \ge \varepsilon$, then $\mathbf{E}[Z'] \ge (2 - c_1 - c_2)m_2\varepsilon^2$.*

Proof. The former case is immediate from (2.1).

For the latter case, note that

$$d_{\mathrm{H}}^2(p,q) = d_{\mathrm{H}}^2(p_{\mathcal{A}}, q_{\mathcal{A}}) + d_{\mathrm{H}}^2(p_{\bar{\mathcal{A}}}, q_{\bar{\mathcal{A}}}).$$

We upper bound the latter term as follows:

$$\begin{aligned} d_{\mathrm{H}}^2(p_{\bar{\mathcal{A}}}, q_{\bar{\mathcal{A}}}) &\le d_{\mathrm{TV}}(p_{\bar{\mathcal{A}}}, q_{\bar{\mathcal{A}}}) \\ &= \frac{1}{2}\sum_{i\in\bar{\mathcal{A}}} |p_i - q_i| \\ &\le \frac{1}{2}\left(p(\bar{\mathcal{A}}) + q(\bar{\mathcal{A}})\right) \\ &\le \left(\frac{c_1 + c_2}{2}\right)\varepsilon^2 \end{aligned}$$

The first inequality is from Proposition 1, and the third inequality is from our prior condition that $p(\bar{\mathcal{A}}) \le c_2\varepsilon^2$.

Since $d_{\mathrm{H}}^2(p,q) \ge \varepsilon^2$, this implies $d_{\mathrm{H}}^2(p_{\mathcal{A}}, q_{\mathcal{A}}) \ge \left(1 - \frac{c_1+c_2}{2}\right)\varepsilon^2$. Proposition 1 further implies that $d_{\chi^2}(p_{\mathcal{A}}, q_{\mathcal{A}}) \ge (2 - c_1 - c_2)\,\varepsilon^2$. The lemma follows from (2.1). □

Lemma 5. *If $d_{\chi^2}(p,q) \le \varepsilon^2$, then $\mathbf{Var}[Z'] = O(m_2^2\varepsilon^4)$. If $d_{\mathrm{H}}(p,q) \ge \varepsilon$, then $\mathbf{Var}[Z'] \le O(\mathbf{E}[Z']^2)$. The constant in both expressions can be made arbitrarily small with the choice of the constant C.*

Proof. We bound the terms of (2.2) separately, starting with the first.

$$
\begin{aligned}
2\sum_{i\in\mathcal{A}}\frac{p_i^2}{q_i^2} &= 2\sum_{i\in\mathcal{A}}\left(\frac{(p_i-q_i)^2}{q_i^2}+\frac{2p_iq_i-q_i^2}{q_i^2}\right)\\
&= 2\sum_{i\in\mathcal{A}}\left(\frac{(p_i-q_i)^2}{q_i^2}+\frac{2q_i(p_i-q_i)+q_i^2}{q_i^2}\right)\\
&\le 2n+2\sum_{i\in\mathcal{A}}\left(\frac{(p_i-q_i)^2}{q_i^2}+2\frac{(p_i-q_i)}{q_i}\right)\\
&\le 4n+4\sum_{i\in\mathcal{A}}\frac{(p_i-q_i)^2}{q_i^2}\\
&\le 4n+\frac{4n}{c_1\varepsilon^2}\sum_{i\in\mathcal{A}}\frac{(p_i-q_i)^2}{q_i}\\
&= 4n+\frac{4n}{c_1\varepsilon^2}\frac{E[Z']}{m_2}\\
&\le 4n+\frac{4}{c_1C}\sqrt{n}E[Z'] \qquad (2.4)
\end{aligned}
$$

The second inequality is the AM-GM inequality, the third inequality uses that $q_i \ge \frac{c_1\varepsilon^2}{n}$ for all $i \in \mathcal{A}$, the last equality uses (2.1), and the final inequality substitutes a value $m_2 \ge C\frac{\sqrt{n}}{\varepsilon^2}$.

The second term can be similarly bounded:

$$
\begin{aligned}
4m_2\sum_{i\in\mathcal{A}}\frac{p_i(p_i-q_i)^2}{q_i^2} &\le 4m_2\left(\sum_{i\in\mathcal{A}}\frac{p_i^2}{q_i^2}\right)^{1/2}\left(\sum_{i\in\mathcal{A}}\frac{(p_i-q_i)^4}{q_i^2}\right)^{1/2}\\
&\le 4m_2\left(4n+\frac{4}{c_1C}\sqrt{n}E[Z']\right)^{1/2}\left(\sum_{i\in\mathcal{A}}\frac{(p_i-q_i)^4}{q_i^2}\right)^{1/2}\\
&\le 4m_2\left(2\sqrt{n}+\frac{2}{\sqrt{c_1C}}n^{1/4}E[Z']^{1/2}\right)\left(\sum_{i\in\mathcal{A}}\frac{(p_i-q_i)^2}{q_i}\right)\\
&= \left(8\sqrt{n}+\frac{8}{\sqrt{c_1C}}n^{1/4}E[Z']^{1/2}\right)E[Z'].
\end{aligned}
$$

The first inequality is Cauchy-Schwarz, the second inequality uses (2.4), the third inequality uses the monotonicity of the ℓ_p norms, and the equality uses (2.1).

Combining the two terms, we get

$$
\mathbf{Var}[Z'] \le 4n+\left(8+\frac{4}{c_1C}\right)\sqrt{n}\mathbf{E}[Z']+\frac{8}{\sqrt{c_1C}}n^{1/4}\mathbf{E}[Z']^{3/2}.
$$

We now consider the two cases in the statement of our lemma.

- When $d_{\chi^2}(p,q) \le \varepsilon^2$, we know from Lemma 4 that $\mathbf{E}[Z'] \le m_2\varepsilon^2$. Combined with a choice of $m_2 \ge C\frac{\sqrt{n}}{\varepsilon^2}$ and the above expression for the variance, this gives:

$$\begin{aligned}\mathbf{Var}[Z'] &\le \frac{4}{C^2}m_2^2\varepsilon^4 + \left(\frac{8}{C} + \frac{4}{c_1C^2}\right)m_2^2\varepsilon^4 + \frac{8}{C\sqrt{c_1}}m_2^2\varepsilon^4 \\ &= \left(\frac{8}{C} + \frac{8}{C\sqrt{c_1}} + \frac{4}{C^2} + \frac{4}{c_1C^2}\right)m_2^2\varepsilon^4 = O(m_2^2\varepsilon^4).\end{aligned}$$

- When $d_{\mathrm{H}}(p,q) \ge \varepsilon$, Lemma 4 and $m_2 \ge C\frac{\sqrt{n}}{\varepsilon^2}$ give:

$$\mathbf{E}[Z'] \ge (2 - c_1 - c_2)m_2\varepsilon^2 \ge C(2 - c_1 - c_2)\sqrt{n}.$$

  Similar to before, combining this with our expression for the variance we get:

$$\begin{aligned}\mathbf{Var}[Z'] &\le \left(\frac{8}{C(2-c_1-c_2)} + \frac{8}{C\sqrt{c_1(2-c_1-c_2)}} + \frac{4}{C^2(2-c_1-c_2)^2} + \frac{4}{C^2c_1(2-c_1-c_2)}\right)\mathbf{E}[Z']^2 \\ &= O(\mathbf{E}[Z']^2).\end{aligned}$$

□

To conclude the proof, we consider the two cases.

- Suppose $d_{\chi^2}(p,q) \le \varepsilon^2$. By Lemma 4 and the definition of $\mathcal{A}$, we have that $\mathbf{E}[Z] \le (1+c_1)m_2\varepsilon^2$. By Lemma 5, $\mathbf{Var}[Z] = O(m_2^2\varepsilon^4)$. Therefore, for constant C sufficiently large, Chebyshev's inequality implies $\Pr(Z > \frac{3}{2}m_2\varepsilon^2) \le 1/10$.

- Suppose $d_{\mathrm{H}}(p,q) \ge \varepsilon$. By Lemma 4, we have that $\mathbf{E}[Z'] \ge (2 - c_1 - c_2)m_2\varepsilon^2$. By Lemma 5, $\mathbf{Var}[Z'] = O(\mathbf{E}[Z']^2)$. Therefore, for constant C sufficiently large, Chebyshev's inequality implies $\Pr(Z' < \frac{3}{2}m_2\varepsilon^2) \le 1/10$. Since $Z \ge Z'$, $\Pr(Z < \frac{3}{2}m_2\varepsilon^2) \le 1/10$ as well.

2.3.2 Identity Testing with ℓ_2 Tolerance

In this section, we sketch the algorithms required to achieve ℓ_2 tolerance for identity testing. Since the algorithms and analysis are very similar to those of Algorithm 1 of [ADK15] and

Algorithm 1, the full details are omitted.

First, we prove Theorem 5. The algorithm is Algorithm 1 of [ADK15], but instead of testing on p and q, we instead test on $p^{+\frac{1}{2}}$ and $q^{+\frac{1}{2}}$, as defined in Proposition 4. By this proposition, this operation preserves total variation and ℓ_2 distance, up to a factor of 2, and also makes it so that the minimum probability element of $q^{+\frac{1}{2}}$ is at least $1/2n$. In the case where $d_{\ell_2}(p,q) \leq \frac{\varepsilon}{\sqrt{n}}$, we have the following upper bound on $\mathbf{E}[Z]$:

$$\mathbf{E}[Z] = m \sum_{i \in \mathcal{A}} \frac{(p_i - q_i)^2}{q_i} \leq O\left(m \cdot n \cdot d_{\ell_2}^2(p,q)\right) \leq O(m\varepsilon^2).$$

This is the same bound as in Lemma 2 of [ADK15]. The rest of the analysis follows identically to that of Algorithm 1 of [ADK15], giving us Theorem 5.

Next, we prove Theorem 6. We observe that Algorithm 1 as stated can be considered as ℓ_2-tolerant instead of χ^2-tolerant, if desired. First, we do not wrongfully reject any p (i.e., those with $d_{\ell_2}(p,q) \leq \frac{\varepsilon^2}{\sqrt{n}}$) in Step 5. This is because we reject in this step if there is $\geq \Omega(\varepsilon^2)$ total variation distance between p and q (witnessed by the set $\bar{\mathcal{A}}$), which implies that p and q are far in ℓ_2-distance by Proposition 2. It remains to prove an upper bound on $\mathbf{E}[Z']$ in the case where $d_{\ell_2}(p,q) \leq \frac{\varepsilon^2}{\sqrt{n}}$.

$$\mathbf{E}[Z'] = m_2 d_{\chi^2}(p,q) = m_2 \sum_{i \in \mathcal{A}} \frac{(p_i - q_i)^2}{q_i} \leq O\left(m_2 \cdot \left(\frac{n}{\varepsilon^2}\right) \cdot d_{\ell_2}^2(p,q)\right) \leq O(m_2\varepsilon^2).$$

We note that this is the same bound as in Lemma 4. With this bound on the mean, the rest of the analysis is identical to that of Theorem 4, giving us Theorem 6.

2.4 Upper Bounds for Equivalence Testing

In this section, we prove the following theorems for equivalence testing.

Theorem 7. *There exists an algorithm for equivalence testing between p and q distinguishing the cases:*

- $d_{\ell_2}(p,q) \leq \frac{\varepsilon}{2\sqrt{n}}$

- $d_{\mathrm{TV}}(p,q) \geq \varepsilon$

The algorithm uses $O\left(\max\left\{\frac{n^{2/3}}{\varepsilon^{4/3}}, \frac{n^{1/2}}{\varepsilon^2}\right\}\right)$ *samples.*

Theorem 8. *There exists an algorithm for equivalence testing between* p *and* q *distinguishing the cases:*

- $d_{\ell_2}(p,q) \leq \frac{\varepsilon^2}{32\sqrt{n}}$
- $d_{\mathrm{H}}(p,q) \geq \varepsilon$

The algorithm uses $O\left(\min\left\{\frac{n^{2/3}}{\varepsilon^{8/3}}, \frac{n^{3/4}}{\varepsilon^2}\right\}\right)$ *samples.*

Consider drawing Poisson(m) samples from two unknown distributions $p = (p_1, \ldots, p_n)$ and $q = (q_1, \ldots, q_n)$. Given the resulting histograms $\boldsymbol{X}$ and $\boldsymbol{Y}$, [CDVV14] define the following statistic:

$$\boldsymbol{Z} = \sum_{i=1}^{n} \frac{(\boldsymbol{X}_i - \boldsymbol{Y}_i)^2 - \boldsymbol{X}_i - \boldsymbol{Y}_i}{\boldsymbol{X}_i + \boldsymbol{Y}_i}. \tag{2.5}$$

This can be viewed as a modification to the empirical triangle distance applied to $\boldsymbol{X}$ and $\boldsymbol{Y}$. Both of our equivalence testing upper bounds will be obtained by appropriate thresholding of the statistic $\boldsymbol{Z}$.

The organization of this section is as follows. We proceed to prove some basic properties of $\boldsymbol{Z}$. In Section 2.4.1, we prove Theorem 7. In Section 2.4.2, we prove Theorem 8.

Some facts about Z. Chan et al. [CDVV14] give the following expressions for the mean and variance of $\boldsymbol{Z}$.

Proposition 6 ([CDVV14])**.** *Consider the function*

$$f(x) = \left(1 - \frac{1 - e^{-x}}{x}\right).$$

Then for any subset $A \subseteq [n]$*,*

$$\mathbf{E}[\boldsymbol{Z}_A] = \sum_{i \in A} \frac{(p_i - q_i)^2}{p_i + q_i} m \cdot f(m(p_i + q_i)). \tag{2.6}$$

As a result, $\boldsymbol{Z}$ is mean-zero when $p = q$. Furthermore,

$$\mathbf{Var}[\boldsymbol{Z}] \leq 2\min\{m, n\} + \sum_{i=1}^{n} 5m\frac{(p_i - q_i)^2}{p_i + q_i}.$$

Applying Proposition 3, we immediately have the following corollary.

Corollary 1. $\mathbf{Var}[\boldsymbol{Z}] \leq 2\min\{m, n\} + 20md_{\mathrm{H}}(p, q)^2$.

Without the corrective factor of $f(m(p_i + q_i))$, Equation (2.6) would just be m times the triangle distance between p and q. Our goal then is to understand the function $f(x)$ and how it affects this quantity. Aside from the removable discontinuity at $x = 0$, f is a monotonically increasing function, and for $x > 0$, it is strictly bounded between 0 and 1. Furthermore, for $x > 0$ there are roughly two "regimes" that $f(x)$ exhibits: when $x < 1$, where $f(x)$ is well-approximated by $x/2$, and when $x \geq 1$, where $f(x)$ is "morally the constant one," slowly increasing from e^{-1} to 1. In fact, we have the following explicit bound on $f(x)$.

Fact 1. *For all $x > 0$, $f(x) \leq \min\{1, x\}$.*

In terms of $f(m(p_i + q_i))$, these regimes correspond to whether $p_i + q_i$ is less than or greater than $\frac{1}{m}$. Hence, the expression for the mean of $\boldsymbol{Z}$ (i.e. Equation (2.6) for $A = [n]$) splits in two: those terms for "large" $p_i + q_i$ look roughly like the triangle distance (times m), and those terms for "small" $p_i + q_i$ look roughly like the ℓ_2^2 distance (times m^2). This is why we have given ourselves the flexibility to consider subsets A of the domain.

We will now prove several upper and lower bounds on $\mathbf{E}[\boldsymbol{Z}_A]$, based in part on whether we will apply them in the large or small $p_i + q_i$ regime. Let us begin with a pair of upper bounds.

Proposition 7. *Suppose for every $i \in A$, $p_i + q_i \geq \delta$. Then*

$$\mathbf{E}[\boldsymbol{Z}_A] \leq \frac{m}{\delta} d_{\ell_2}^2(p_A, q_A).$$

Proof. Because $f(x) \le 1$ for all $x > 0$,

$$\mathbf{E}[\boldsymbol{Z}_A] = \sum_{i \in A} \frac{(p_i - q_i)^2}{p_i + q_i} m \cdot f(m(p_i + q_i)) \le \sum_{i \in A} \frac{(p_i - q_i)^2}{p_i + q_i} m \le \frac{m}{\delta} \sum_{i \in A} (p_i - q_i)^2 = \frac{m}{\delta} d_{\ell_2}^2(p_A, q_A).$$

□

Proposition 8. $\mathbf{E}[\boldsymbol{Z}] \le m^2 d_{\ell_2}^2(p, q)$.

Proof. Let L be the set of i such that $m(p_i + q_i) \ge 1$. Then $\mathbf{E}[\boldsymbol{Z}] = \mathbf{E}[\boldsymbol{Z}_L] + \mathbf{E}[\boldsymbol{Z}_{\overline{L}}]$, and by Proposition 7, $\mathbf{E}[\boldsymbol{Z}_L] \le m^2 d_{\ell_2}^2(p_L, q_L)$. On the other hand, by Fact 1, $f(x) \le x$, and therefore

$$\mathbf{E}[\boldsymbol{Z}_{\overline{L}}] = \sum_{i \in \overline{L}} \frac{(p_i - q_i)^2}{p_i + q_i} m \cdot f(m(p_i + q_i)) \le \sum_{i \in \overline{L}} (p_i - q_i)^2 m^2 = m^2 d_{\ell_2}^2(p_{\overline{L}}, q_{\overline{L}}).$$

The proof is completed by noting that $d_{\ell_2}^2(p_L, q_L) + d_{\ell_2}^2(p_{\overline{L}}, q_{\overline{L}}) = d_{\ell_2}^2(p, q)$. □

Now we give a pair of lower bounds.

Proposition 9. *Suppose for every $i \in A$, $m(p_i + q_i) \ge 1$. Then*

$$\mathbf{E}[\boldsymbol{Z}_A] \ge \frac{2m}{3} d_{\mathrm{H}}^2(p_A, q_A).$$

Proof. Because $f(x)$ is monotonically increasing and $f(1) = 1/e$,

$$\mathbf{E}[\boldsymbol{Z}_A] = m \sum_{i \in A} \frac{(p_i - q_i)^2}{p_i + q_i} f(m(p_i + q_i)) \ge m \sum_{i \in A} \frac{(p_i - q_i)^2}{p_i + q_i} f(1) \ge \frac{2m}{e} d_{\mathrm{H}}^2(p_A, q_A),$$

where the first step is by Proposition 6 and the last is by Proposition 3. The result follows from $e \le 3$. □

The next proposition is essentially the second half of the proof of Lemma 4 from [CDVV14].

Proposition 10. *For any subset A,*

$$\mathbf{E}[\boldsymbol{Z}_A] \ge \left(\frac{4m^2}{2|A| + m \cdot (p(A) + q(A))} \right) \cdot d_{\mathrm{TV}}^2(p_A, q_A),$$

where we write $p(A) = \sum_{i \in A} p(i)$ and likewise for $q(A)$.

Proof. Consider the function $g(x) = xf(x)^{-1}$. Then $g(x) \le 2 + x$ for nonnegative x. Furthermore,

$$\frac{(p_i - q_i)^2}{g(m(p_i + q_i))} = \frac{(p_i - q_i)^2}{m(p_i + q_i)} \left(1 - \frac{1 - e^{-m(p_i+q_i)}}{m(p_i + q_i)}\right),$$

which, from Proposition 6, is $\frac{1}{m^2} \cdot \mathbf{E}[\boldsymbol{Z}_{\{i\}}]$. As a result,

$$d_{\mathrm{TV}}^2(p_A, q_A) = \frac{1}{4}\left(\sum_{i\in A} |p_i - q_i|\right)^2 = \frac{1}{4}\left(\sum_{i\in A} |p_i - q_i| \cdot \frac{\sqrt{g(m(p_i + q_i))}}{\sqrt{g(m(p_i + q_i))}}\right)^2$$
$$\le \frac{1}{4}\left(\sum_{i\in A} \frac{(p_i - q_i)^2}{g(m(p_i + q_i))}\right) \cdot \left(\sum_{i\in A} g(m(p_i + q_i))\right) \le \frac{1}{4m^2} \cdot \mathbf{E}[\boldsymbol{Z}_A] \cdot (2|A| + m \cdot (p(A) + q(A))),$$

where the first inequality is Cauchy-Schwarz. Rearranging finishes the proof. □

2.4.1 Equivalence Testing with ℓ_2 Tolerance

In this section, we prove Theorem 7. We will take the number of samples to be

$$m = \max\left\{C \cdot \frac{n^{2/3}}{\varepsilon^{4/3}}, C^{3/2} \cdot \frac{n^{1/2}}{\varepsilon^2}\right\}, \tag{2.7}$$

where C is some constant which can be taken to be 10^{10}.

Rather than drawing samples from p or q, our algorithm draws samples from $p^{+1/2}$ and $q^{+1/2}$. By Proposition 4, we have the following guarantees in the two cases:

$$\text{(Case 1): } d_{\ell_2}(p^{+1/2}, q^{+1/2}) \le \frac{\varepsilon}{4\sqrt{n}}, \qquad \text{(Case 2): } d_{\mathrm{TV}}(p^{+1/2}, q^{+1/2}) \ge \frac{\varepsilon}{2}.$$

Furthermore, for any $i \in [n]$, we know the i-th coordinates of $p^{+1/2}$ and $q^{+1/2}$ are both at least $\frac{1}{2n}$. Henceforth, we will write p' and q' for $p^{+1/2}$ and $q^{+1/2}$, respectively.

In Case 1, if we apply Proposition 7 with $A = [n]$ and $\delta = \frac{1}{n}$ and Proposition 8,

$$\mathbf{E}[\boldsymbol{Z}] \le \min\{m^2, mn\} \cdot d_{\ell_2}^2(p', q') \le \min\{m^2, mn\} \cdot \frac{\varepsilon^2}{16n} \le \frac{m^2}{4(2m + 2n)} \cdot \varepsilon^2.$$

On the other hand, in Case 2, applying Proposition 10 with $A = [n]$,

$$\mathbf{E}[\boldsymbol{Z}] \geq \frac{4m^2}{2m+2n} \cdot d_{\mathrm{TV}}(p', q')^2 \geq \frac{m^2}{2m+2n} \cdot \varepsilon^2.$$

Our algorithm therefore thresholds $\boldsymbol{Z}$ on the value $\frac{5m^2}{8(2m+2n)}\varepsilon^2$, outputting "close" if it's below this value and "far" otherwise.

The two bounds in (2.7) meet when $C^3\varepsilon^{-4} = n$, which is exactly when $m = n$. When $m \leq n$, the first bound applies, and when $m > n$ the second bound applies. As a result, we will split our analysis into the two cases.

Lemma 6. *The tester succeeds in the $m \leq n$ case of Theorem 7.*

Proof. By Corollary 1

$$\mathbf{Var}[\boldsymbol{Z}] \leq 2\min\{m, n\} + 20md_{\mathrm{H}}(p', q')^2 \leq 22m,$$

where we used the fact that $d_{\mathrm{H}}(p', q') \leq 1$. In Case 1, by Chebyshev's inequality,

$$\Pr\left[\boldsymbol{Z} \geq \frac{5m^2}{8(2m+2n)}\varepsilon^2\right] \leq \frac{\mathbf{Var}[\boldsymbol{Z}]}{\left(\frac{3m^2}{8(2m+2n)}\varepsilon^2\right)^2} = O\left(\frac{m}{\frac{m^4}{n^2}\varepsilon^4}\right) = O\left(\frac{n^2}{m^3\varepsilon^4}\right).$$

In Case 2,

$$\Pr\left[\boldsymbol{Z} \leq \frac{5m^2}{8(2m+2n)}\varepsilon^2\right] \leq \frac{64\mathbf{Var}[\boldsymbol{Z}]}{9\mathbf{E}[\boldsymbol{Z}]^2} = O\left(\frac{m}{\frac{m^4}{n^2}\varepsilon^4}\right) = O\left(\frac{n^2}{m^3\varepsilon^4}\right).$$

Both of these bounds can be made arbitrarily small constants by setting C sufficiently large. □

Lemma 7. *The tester succeeds in the $m \geq n$ case of Theorem 7.*

Proof. We first consider Case 1. By Proposition 6,

$$\mathbf{Var}[\boldsymbol{Z}] \leq 2\min\{m, n\} + \sum_{i=1}^{n} 5m\frac{(p'_i - q'_i)^2}{p'_i + q'_i} \leq 2n + 5mnd_{\ell_2}^2(p', q') \leq 2n + \tfrac{5}{16}m\varepsilon^2.$$

Then, we have that

$$\Pr\left[\boldsymbol{Z} \geq \frac{5m^2}{8(2m+2n)}\varepsilon^2\right] \leq \frac{\mathbf{Var}[\boldsymbol{Z}]}{\left(\frac{3m^2}{8(2m+2n)}\varepsilon^2\right)^2} = O\left(\frac{n}{m^2\varepsilon^4} + \frac{m\varepsilon^2}{m^2\varepsilon^4}\right) = O\left(\frac{n}{m^2\varepsilon^4} + \frac{1}{m\varepsilon^2}\right).$$

Next, we focus on Case 2. Write L for the set of $i \in [n]$ such that $m(p'_i + q'_i) \geq 1$. Then $d^2_{\mathrm{H}}(p'_{\overline{L}}, q'_{\overline{L}}) \leq \frac{1}{2}\sum_{i\in\overline{L}}(p'_i + q'_i) \leq n/2m$. As a result, by Corollary 1

$$\mathbf{Var}[\boldsymbol{Z}] \leq 2\min\{m, n\} + 20md^2_{\mathrm{H}}(p', q') \leq 12n + 20md^2_{\mathrm{H}}(p'_L, q'_L).$$

By Proposition 9, $\mathbf{E}[\boldsymbol{Z}] \geq \frac{2m}{3}d^2_{\mathrm{H}}(p'_L, q'_L)$. Hence,

$$\begin{aligned}\Pr\left[\boldsymbol{Z} \leq \frac{5m^2}{8(2m+2n)}\varepsilon^2\right] &\leq \frac{64\mathbf{Var}[\boldsymbol{Z}]}{9\mathbf{E}[\boldsymbol{Z}]^2} = O\left(\frac{n}{\mathbf{E}[\boldsymbol{Z}]^2} + \frac{md^2_{\mathrm{H}}(p'_L, q'_L)}{\mathbf{E}[\boldsymbol{Z}]^2}\right) \\ &= O\left(\frac{n}{\mathbf{E}[\boldsymbol{Z}]^2} + \frac{1}{\mathbf{E}[\boldsymbol{Z}]}\right) = O\left(\frac{n}{m^2\varepsilon^4} + \frac{1}{m\varepsilon^2}\right).\end{aligned}$$

Both of these bounds can be made arbitrarily small constants by setting C sufficiently large. □

2.4.2 Equivalence Testing with Hellinger Distance

In this section, we prove Theorem 8. We will take the number of samples to be

$$m = \min\left\{C \cdot \frac{n^{2/3}}{\varepsilon^{8/3}}, C^{3/4} \cdot \frac{n^{3/4}}{\varepsilon^2}\right\},$$

where C is some constant which can be taken to be 10^{10}.

Rather than drawing samples from p or q, our algorithm draws samples from $p^{+\delta}$ and $q^{+\delta}$ for $\delta = \varepsilon^2/32$. By Proposition 4, we have the following guarantees in the two cases:

$$\text{(Case 1): } d_{\ell_2}(p, q) \leq \frac{\varepsilon^2}{32\sqrt{n}}, \qquad \text{(Case 2): } d_{\mathrm{H}}(p, q) \geq \frac{1}{2}\varepsilon.$$

Furthermore, for any $i \in [n]$, we know the i-th coordinates of $p^{+\delta}$ and $q^{+\delta}$ are both at least $\frac{\varepsilon^2}{32n}$. Henceforth, we will write p' and q' for $p^{+\delta}$ and $q^{+\delta}$, respectively.

The two bounds meet when $C^{3/4}\varepsilon^{-2} = n^{1/4}$, which is exactly when $m = n$. When $m \le n$, the first bound applies, and when $m > n$ the second bound applies. As a result, we will split our analysis into the two cases.

Lemma 8. *The tester succeeds in the $m \le n$ case of Theorem 8.*

Proof. In Case 1, if we apply Proposition 8,

$$\mathbf{E}[\boldsymbol{Z}] \le m^2 \cdot d_{\ell_2}^2(p', q') \le \frac{m^2\varepsilon^4}{32^2 n}.$$

On the other hand, in Case 2, applying Proposition 10 with $A = [n]$,

$$\mathbf{E}[\boldsymbol{Z}] \ge \left(\frac{4m^2}{2n+2m}\right) \cdot d_{\mathrm{TV}}(p', q')^2 \ge \left(\frac{4m^2}{2n+2m}\right) \cdot d_{\mathrm{H}}(p', q')^4 \ge \frac{m^2\varepsilon^4}{16n}.$$

Our algorithm therefore thresholds $\boldsymbol{Z}$ on the value $\frac{m^2\varepsilon^4}{128n}$, outputting "close" if it's below this value and "far" otherwise.

By Corollary 1

$$\mathbf{Var}[\boldsymbol{Z}] \le 2\min\{m, n\} + 20md_{\mathrm{H}}(p', q')^2 \le 22m,$$

where we used the fact that $d_{\mathrm{H}}(p', q') \le 1$. In Case 1,

$$\Pr\left[\boldsymbol{Z} \ge \frac{m^2\varepsilon^4}{128n}\right] \le \frac{\mathbf{Var}[\boldsymbol{Z}]}{\left(\frac{m^2\varepsilon^4}{256n}\right)^2} = O\left(\frac{m}{\frac{m^4}{n^2}\varepsilon^8}\right) = O\left(\frac{n^2}{m^3\varepsilon^8}\right).$$

In Case 2,

$$\Pr\left[\boldsymbol{Z} \le \frac{m^2\varepsilon^4}{128n}\right] \le \frac{64\mathbf{Var}[\boldsymbol{Z}]}{49\mathbf{E}[\boldsymbol{Z}]^2} = O\left(\frac{m}{\frac{m^4}{n^2}\varepsilon^8}\right) = O\left(\frac{n^2}{m^3\varepsilon^8}\right).$$

Both of these bounds can be made arbitrarily small constants by setting C sufficiently large. □

Lemma 9. *The tester succeeds in the $m > n$ case of Theorem 8.*

Proof. In Case 1, if we apply Proposition 7 with $A = [n]$ and $\delta = \frac{\varepsilon^2}{16n}$ and Proposition 8,

$$\mathbf{E}[\boldsymbol{Z}] \le \min\left\{m^2, 16\frac{mn}{\varepsilon^2}\right\} \cdot d_{\ell_2}^2(p', q') \le \min\left\{m^2, 16\frac{mn}{\varepsilon^2}\right\} \cdot \frac{\varepsilon^4}{32^2 n} = \min\left\{\frac{m^2\varepsilon^4}{32^2 n}, \frac{m\varepsilon^2}{64}\right\}.$$

Case 2 is more complicated. We will need to define the set of "large" coordinates $L = \{i : m(p'_i + q'_i) \ge 1\}$ and the set of "small" coordinates $S = [n] \setminus L$. Applying Proposition 10 to S, we have

$$\mathbf{E}[\boldsymbol{Z}_S] \ge \left(\frac{4m^2}{2|S| + m \cdot (p'(S) + q'(S))}\right) \cdot d_{\mathrm{TV}}^2(p'_S, q'_S) \ge \frac{4m^2}{3n} d_{\mathrm{TV}}^2(p'_S, q'_S),$$

where $m \cdot (p'(S) + q'(S)) \le n$ by the definition of S. If we also apply Proposition 9 to L, we get

$$\mathbf{E}[\boldsymbol{Z}] = \mathbf{E}[\boldsymbol{Z}_S] + \mathbf{E}[\boldsymbol{Z}_L] \ge \frac{4m^2}{3n} d_{\mathrm{TV}}^2(p'_S, q'_S) + \frac{2m}{3} d_{\mathrm{H}}^2(p'_L, q'_L) \ge \min\left\{\frac{m^2\varepsilon^4}{48n}, \frac{m\varepsilon^2}{12}\right\},$$

where the last step follows because $d_{\mathrm{H}}^2(p'_S, q'_S) + d_{\mathrm{H}}^2(p'_L, q'_L) = d_{\mathrm{H}}^2(p', q')$ and $d_{\mathrm{TV}}^2(p'_S, q'_S) \ge d_{\mathrm{H}}^4(p'_S, q'_S)$. As a result, we threshold $\boldsymbol{Z}$ on the value

$$\frac{1}{2} \cdot \min\left\{\frac{m^2\varepsilon^4}{48n}, \frac{m\varepsilon^2}{12}\right\},$$

outputting "close" if it's below this value and "far" otherwise.

In Case 1, by Proposition 6,

$$\mathbf{Var}[\boldsymbol{Z}] \le 2\min\{m, n\} + \sum_{i=1}^{m} 5m\frac{(p'_i - q'_i)^2}{p'_i + q'_i} \le 2n + \frac{80mn}{\varepsilon^2}\|p' - q'\|_2^2 \le 2n + \frac{5}{64}m\varepsilon^2.$$

Hence, by Chebyshev's inequality,

$$\begin{aligned}\Pr\left[\boldsymbol{Z} \geq \frac{1}{2}\cdot\min\left\{\frac{m^2\varepsilon^4}{48n}, \frac{m\varepsilon^2}{12}\right\}\right] &\leq \frac{\mathbf{Var}[\boldsymbol{Z}]}{(\frac{1}{8}\cdot\min\left\{\frac{m^2\varepsilon^4}{48n}, \frac{m\varepsilon^2}{12}\right\})^2} \\ &\leq O\left(\frac{n}{(\frac{m^2\varepsilon^4}{n})^2} + \frac{n}{(m\varepsilon^2)^2} + \frac{m\varepsilon^2}{(\frac{m^2\varepsilon^4}{n})^2} + \frac{m\varepsilon^2}{(m\varepsilon^2)^2}\right) \\ &= O\left(\frac{n^3}{m^4\varepsilon^8} + \frac{n}{m^2\varepsilon^4} + \frac{n^2}{m^3\varepsilon^6} + \frac{1}{m\varepsilon^2}\right).\end{aligned}$$

This can be made an arbitrarily small constant by setting C sufficiently large.

In Case 2, by Corollary 1,

$$\Pr\left[\boldsymbol{Z} \leq \frac{\mathbf{E}[\boldsymbol{Z}]}{2}\right] \leq \frac{4\mathbf{Var}[\boldsymbol{Z}]}{\mathbf{E}[\boldsymbol{Z}]^2} \leq \frac{8n + 80md_{\mathrm{H}}(p', q')^2}{\mathbf{E}[\boldsymbol{Z}]^2}. \tag{2.8}$$

Because $d_{\mathrm{H}}(p', q')^2 = d_{\mathrm{H}}^2(p'_S, q'_S) + d_{\mathrm{H}}^2(p'_L, q'_L)$, either $d_{\mathrm{H}}^2(p'_S, q'_S)$ or $d_{\mathrm{H}}^2(p'_L, q'_L)$ is at least $\frac{1}{2}d_{\mathrm{H}}^2(p', q')$. Suppose that $d_{\mathrm{H}}^2(p'_S, q'_S) \geq \frac{1}{2}d_{\mathrm{H}}^2(p', q')$. We note that

$$md_{\mathrm{H}}^2(p'_S, q'_S) = \frac{m}{2}\sum_{i\in S}(\sqrt{p'_i} - \sqrt{q'_i})^2 \leq \frac{m}{2}\sum_{i\in S}|p'_i + q'_i| \leq \frac{n}{2},$$

by the definition of S. Thus,

$$(2.8) \leq \frac{8n + 160md_{\mathrm{H}}^2(p'_S, q'_S)}{(\frac{4m^2}{3n}d_{\mathrm{TV}}^2(p'_S, q'_S))^2} \leq \frac{88n}{(\frac{4m^2}{3n}d_{\mathrm{TV}}^2(p'_S, q'_S))^2} = O\left(\frac{n^3}{m^4 d_{\mathrm{TV}}^4(p'_S, q'_S)}\right) \leq O\left(\frac{n^3}{m^4\varepsilon^8}\right),$$

where the last step used the fact that $d_{\mathrm{TV}}(p'_S, q'_S) \geq d_{\mathrm{H}}^2(p'_S, q'_S) \geq \frac{1}{2}d_{\mathrm{H}}^2(p', q') \geq \frac{1}{2}\varepsilon^2$.

In the case when $d_{\mathrm{H}}^2(p'_L, q'_L) \geq \frac{1}{2}d_{\mathrm{H}}^2(p', q')$,

$$(2.8) \leq \frac{8n + 160md_{\mathrm{H}}^2(p'_L, q'_L)}{(\frac{2m}{3}d_{\mathrm{H}}^2(p'_L, q'_L))^2} = O\left(\frac{n}{m^2 d_{\mathrm{H}}^4(p'_L, q'_L)} + \frac{1}{md_{\mathrm{H}}^2(p'_L, q'_L)}\right) \leq O\left(\frac{n}{m^2\varepsilon^4} + \frac{1}{m\varepsilon^2}\right).$$

This can be made an arbitrarily small constant by setting C sufficiently large. □

2.5 Upper Bounds Based on Estimation

We start by showing a simple meta-algorithm – in short, it says that if a testing problem is well-defined (i.e., has appropriate separation between the cases) and we can estimate one of the distances, it can be converted to a testing algorithm.

Theorem 9. *Suppose there exists an $m(n,\varepsilon)$-sample algorithm which, given sample access to distributions p and q over $[n]$ and parameter ε, estimates some distance $d(p,q)$ up to an additive ε with probability at least $2/3$. Consider distances $d_X(\cdot,\cdot), d_Y(\cdot,\cdot)$ and $\varepsilon_1, \varepsilon_2 > 0$ such that $d_Y(p,q) \geq \varepsilon_2 \to d_X(p,q) > 3\varepsilon_1/2$ and $d_X(p,q) \leq \varepsilon_1 \to d_Y(p,q) < 2\varepsilon_2/3$, and $d(\cdot,\cdot)$ is either $d_X(\cdot,\cdot)$ or $d_Y(\cdot,\cdot)$.*

Then there exists an algorithm for equivalence testing between p and q distinguishing the cases:

- $d_X(p,q) \leq \varepsilon_1$;
- $d_Y(p,q) \geq \varepsilon_2$.

The algorithm uses either $m(n, O(\varepsilon_1))$ or $m(n, O(\varepsilon_2))$ samples, depending on whether $d = d_X$ or d_Y.

Proof. Suppose that $d = d_X$, the other case follows similarly. Using the $m(n, \varepsilon_1/4)$ samples, obtain an estimate $\hat{\tau}$ of $d_X(p,q)$, accurate up to an additive $\varepsilon_1/4$. If $\hat{\tau} \leq 5\varepsilon_1/4$, output that $d_X(p,q) \leq \varepsilon_1$, else output that $d_Y(p,q) \geq \varepsilon_2$. Conditioning on the correctness of the estimation algorithm, correctness for the case when $d_X(p,q) \leq \varepsilon_1$ is immediate, and correctness for the case when $d_Y(p,q) \geq \varepsilon_2$ follows from the separation between the cases. □

It is folklore that a distribution over $[n]$ can be ε-learned in ℓ_2-distance with $O(1/\varepsilon^2)$ samples (see, e.g., [CDVV14, Wag15] for a reference). By triangle inequality, this implies that we can estimate the ℓ_2 distance between p and q up to an additive $O(\varepsilon)$ with $O(1/\varepsilon^2)$ samples, leading to the following corollary.

Corollary 2. *There exists an algorithm for equivalence testing between p and q distinguishing the cases:*

- $d(p,q) \leq f(n,\varepsilon)$;
- $d_{\ell_2}(p,q) \geq \varepsilon$,

where $d(\cdot,\cdot)$ is a distance and $f(n,\varepsilon)$ is such that $d_{\ell_2}(p,q) \geq \varepsilon \to d(p,q) \geq 3f(n,\varepsilon)/2$ and $d(p,q) \leq f(n,\varepsilon) \to d_{\ell_2}(p,q) \leq 2\varepsilon/3$. The algorithm uses $O(1/\varepsilon^2)$ samples.

Finally, we note that total variation distance between p and q can be additively estimated up to a constant using $O(n/\log n)$ samples [LC06, VV11b, JHW16], leading to the following corollary:

Corollary 3. *For constant $\varepsilon > 0$, there exists an algorithm for equivalence testing between p and q distinguishing the cases:*

- $d_{\mathrm{TV}}(p,q) \leq \varepsilon^2/4$;
- $d_{\mathrm{H}}(p,q) \geq \varepsilon/\sqrt{2}$.

The algorithm uses $O(n/\log n)$ samples.

2.6 Lower Bounds for Testing with Tolerance and Alternative Distances

We start with a simple lower bound, showing that identity testing with respect to KL divergence is impossible. A similar observation was made in [BFR$^+$00].

Theorem 10. *No finite sample test can perform identity testing between p and q distinguishing the cases:*

- $p = q$;
- $d_{\mathrm{KL}}(p,q) \geq \varepsilon^2$.

Proof. Simply take $q = (1,0)$ and let p be either $(1,0)$ or $(1-\delta,\delta)$, for $\delta > 0$ tending to zero. Then $p = q$ in the first case and $d_{\mathrm{KL}}(p,q) = \infty$ in the second, but distinguishing between these two possibilities for p takes $\Omega(\delta^{-1}) \to \infty$ samples. □

Next, we prove our lower bound for KL tolerant identity testing.

Theorem 11. *There exist constants $0 < s < c$, such that any algorithm for identity testing between p and q distinguishing the cases:*

- $d_{\mathrm{KL}}(p, q) \leq s$;
- $d_{\mathrm{TV}}(p, q) \geq c$;

requires $\Omega(n/\log n)$ samples.

Proof. Let $q = (\frac{1}{n}, \ldots, \frac{1}{n})$ be the uniform distribution. Let $R(\cdot, \cdot)$ denote the *relative earth-mover distance* (see [VV10a] for the definition). By Theorem 1 of [VV10a], for any $\delta < \frac{1}{4}$ there exist sets of distributions $\mathcal{C}$ and $\mathcal{F}$ (for *close* and *far*) such that:

- For every $p \in \mathcal{C}$, $R(p, q) = O(\delta |\log \delta|)$.
- For every $p \in \mathcal{F}$ there exists a distribution r which is uniform over $n/2$ elements such that $R(p, r) = O(\delta |\log \delta|)$.
- Distinguishing between $p \in \mathcal{C}$ and $p \in \mathcal{F}$ requires $\Omega(\frac{\delta n}{\log(n)})$ samples.

Now, if $p \in \mathcal{C}$ then

$$d_{\mathrm{KL}}(p, q) = \sum_{i=1}^{n} p_i \log\left(\frac{p_i}{1/n}\right) = \log(n) - H(p) \leq O(\delta |\log(\delta)|),$$

where $H(p)$ is the Shannon entropy of p, and here we used the fact that $|H(p) - H(q)| \leq R(p, q)$, which follows from Fact 5 of [VV10a]. On the other hand, if $q \in \mathcal{F}$, let r be the corresponding distribution which is uniform over $n/2$ elements. Then

$$\frac{1}{2} = d_{\mathrm{TV}}(p, q) \leq d_{\mathrm{TV}}(q, p) + d_{\mathrm{TV}}(p, r) \leq d_{\mathrm{TV}}(q, p) + O(\delta |\log \delta|),$$

where we used the triangle inequality and the fact that $d_{\mathrm{TV}}(p, r) \leq R(p, r)$ (see [VV10a] page 4). As a result, if we set δ to be some small constant, $s = O(\delta |\log(\delta)|)$, and $c = \frac{1}{2} - O(\delta |\log \delta|)$, then this argument shows that distinguish $d_{\mathrm{KL}}(p, q) \leq s$ versus $d_{\mathrm{TV}}(p, q) \geq c$ requires $\Omega(n/\log n)$ samples. □

Finally, we conclude with our lower bound for χ^2-tolerant equivalence testing.

Theorem 12. *There exists a constant $\varepsilon > 0$ such that any algorithm for equivalence testing between p and q distinguishing the cases:*

- $d_{\chi^2}(p, q) \leq \varepsilon^2/4$*;*
- $d_{\mathrm{TV}}(p, q) \geq \varepsilon$*;*

requires $\Omega(n/\log n)$ samples.

Proof. We reduce the problem of distinguishing $d_{\mathrm{H}}(p, q) \leq \frac{1}{\sqrt{48}}\varepsilon$ from $d_{\mathrm{TV}}(p, q) \geq 3\varepsilon$ to this. Define the distributions

$$p' = \frac{2}{3}p + \frac{1}{3}q, \qquad q' = \frac{1}{3}p + \frac{2}{3}q.$$

Then m samples from p' and q' can be simulated by m samples from p and q. Furthermore,

$$d_{\mathrm{H}}(p', q') \leq \frac{1}{\sqrt{48}}\varepsilon, \qquad d_{\mathrm{TV}}(p', q') = \frac{1}{3}d_{\mathrm{TV}}(p, q) \geq \varepsilon,$$

where we used the fact that Hellinger distance satisfies the data processing inequality. But then, in the “close" case,

$$d_{\chi^2}(p', q') = \sum_{i=1}^{n} \frac{(p'_i - q'_i)^2}{q'_i} \leq 3\sum_{i=1}^{n} \frac{(p'_i - q'_i)^2}{p'_i + q'_i} \leq 12 d_{\mathrm{H}}^2(p', q') \leq \frac{1}{4}\varepsilon^2,$$

where we used the fact that $p'_i \leq 2q'_i$ and Proposition 3. Hence, this problem, which requires $\Omega(n/\log n)$ samples (by the relationship between total variation and Hellinger distance, and the lower bound for testing total variation-close versus -far of [VV10a]), reduces to the problem in the proposition, and so that requires $\Omega(n/\log n)$ samples as well. □

Chapter 3

Testing Shape-Restricted Families of Distributions

3.1 Introduction

In this chapter, our focus is on *composite* hypothesis testing. Much of prior work focuses on testing for a single null hypothesis q. However, is a rather unrealistic scenario for hypothesis testing – it seems implausible that we would have a precise guess for the unknown distribution. In the previous chapter, we considered one natural way of relaxing this restriction, where we considered testing whether p is in an ε_1-neighborhood of q (in distance measure d_1). In this chapter, we consider an alternative relaxation: we wish to test whether the unknown distribution belongs to some *structured class* of distributions: is $p \in \mathcal{C}$, or is it far from all such representations (i.e., $d_{\mathrm{TV}}(p, \mathcal{C}) \geq \varepsilon$). For example, one might wish to ask if the distribution p follows *some* Binomial distribution, rather than a *particular* Binomial distribution. More precisely, our problem is the following:

> Given a class of distributions $\mathcal{C}$, some $\varepsilon > 0$, and sample access to an unknown distribution p over a discrete support, how many samples are required to distinguish between $p \in \mathcal{C}$ versus $d_{\mathrm{TV}}(p, \mathcal{C}) > \varepsilon$?

In some cases, composite hypothesis testing may be rather simple. For instance, if we wish to test whether p is equal to one of $O(1)$ hypotheses, this can be done with $O(1) \cdot O(\sqrt{n}/\varepsilon^2)$

samples. The real challenge arises when the class $\mathcal{C}$ is infinite. In this case, perhaps the natural extension is to generate an $O(\varepsilon)$-net over $\mathcal{C}$, and perform tolerant testing against every distribution in this net. Even disregarding the size of this net, testing against a single hypothesis requires a testing algorithm which is tolerant in total variation distance, which requires $\Omega(n/\log n)$ samples [VV17b], so this approach seems infeasible. Another approach is to use a generalized likelihood ratio test, but their behavior is not well-understood in our regime, and optimizing likelihood over our classes becomes computationally intense.

In the finite sample regime we consider in this thesis, this type of question has received relatively limited attention. The primary classes of interest have been monotonicity [BKR04, BFRV11] and independence [BFF$^+$01, AAK$^+$07, LRR13, RX14]. Even then, the known upper bounds seem to be quite distant from the information-theoretic lower bounds, and the true sample complexity of these problems is unclear.

3.1.1 Results

In this work, we prove a framework for testing whether an unknown distribution p belongs to some class $\mathcal{C}$, or $d_{\mathrm{TV}}(p, \mathcal{C}) \geq \varepsilon$. We apply this framework in order to obtain sample-optimal and computationally efficient testers for a number of natural shape restrictions to a distribution. Our contributions are the following:

1. For the class $\mathcal{C} = \mathcal{M}_n^d$ of monotone distributions over $[n]^d$ we require an optimal $\Theta\left(\frac{n^{d/2}}{\varepsilon^2}\right)$ number of samples, where prior work requires $\Omega\left(\frac{\sqrt{n}\log n}{\varepsilon^4}\right)$ samples for $d = 1$ and $\tilde{\Omega}\left(n^{d-\frac{1}{2}}\mathrm{poly}\left(\frac{1}{\varepsilon}\right)\right)$ for $d > 1$ [BKR04, BFRV11]. Our results improve the exponent of n with respect to d, shave all logarithmic factors in n, and improve the exponent of ε by at least a factor of 2.

   (a) A useful building block and interesting byproduct of our analysis is extending Birgé's oblivious decomposition for single-dimensional monotone distributions [Bir87] to monotone distributions in $d \geq 1$, and to the stronger notion of χ^2-distance. See Section 3.5.1.

   (b) Moreover, we show that $O(\log^d n)$ samples suffice to learn a monotone distribution over $[n]^d$ in χ^2-distance. See Lemma 11 for the precise statement.

2. For the class $\mathcal{C} = \Pi_d$ of product distributions over $[n_1] \times \cdots \times [n_d]$, our algorithm requires $O\left(\left(\left(\prod_\ell n_\ell\right)^{1/2} + \sum_\ell n_\ell\right)/\varepsilon^2\right)$ samples. We note that a product distribution is one where all marginals are independent, so this is equivalent to testing if a collection of random variables are all independent. In the case where n_ℓ's are large, then the first term dominates, and the sample complexity is $O((\prod_\ell n_\ell)^{1/2}/\varepsilon^2)$. In particular, when d is a constant and all n_ℓ's are equal to n, we achieve the optimal sample complexity of $\Theta(n^{d/2}/\varepsilon^2)$. To the best of our knowledge, this is the first result for $d \geq 3$, and when $d = 2$, this improves the previously known complexity from $O\left(\frac{n}{\varepsilon^6}\text{polylog}(n/\varepsilon)\right)$ [BFF+01, LRR13], significantly improving the dependence on ε and shaving all logarithmic factors.

3. For the classes $\mathcal{C} = \mathcal{LCD}_n$, $\mathcal{C} = \mathcal{MHR}_n$ and $\mathcal{C} = \mathcal{U}_n$ of log-concave, monotone-hazard-rate and unimodal distributions over $[n]$, we require an optimal $\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ number of samples. Our testers for $\mathcal{LCD}_n$ and $\mathcal{C} = \mathcal{MHR}_n$ are to our knowledge the first for these classes for the low sample regime we are studying—see [HVK05] and its references for statistics literature on the asymptotic regime. Our tester for $\mathcal{U}_n$ improves the dependence of the sample complexity on ε by at least a factor of 2 in the exponent, and shaves all logarithmic factors in n, compared to testers based on testing monotonicity.

   (a) A useful building block and important byproduct of our analysis are the first computationally efficient algorithms for properly learning log-concave and monotone-hazard-rate distributions, to within ε in total variation distance, from $\text{poly}(1/\varepsilon)$ samples, independent of the domain size n. See Corollaries 8 and 10. Again, these are the first computationally efficient algorithms to our knowledge in the low sample regime. [CDSS14, ADLS17] provide algorithms for density estimation, which are non-proper, i.e. will approximate an unknown distribution from these classes with a distribution that does not belong to these classes. On the other hand, the statistics literature focuses on maximum-likelihood estimation in the asymptotic regime—see e.g. [CS10] and its references.

4. For all the above classes we obtain matching lower bounds, showing that the sample complexity of our testers is optimal with respect to n, ε and when applicable d. See

Section 3.10. Our lower bounds are based on extending Paninski's lower bound for testing uniformity [Pan08].

Our tester follows a very intuitive "learn then test" approach, though there are some perhaps unexpected technical twists in order to achieve the optimal sample complexity. More precisely, we could imagine the following two-step procedure:

1. **Learn:** Estimate a distribution $q \in \mathcal{C}$ which best approximates p;

2. **Test:** Test whether p and q are close or far.

Naturally, if p and q are close, then we can conclude that $p \in \mathcal{C}$, and if they are far, we know that $d_{\mathrm{TV}}(p, \mathcal{C}) \geq \varepsilon$. The major unspecified parameter in this approach is which *distance* to use: in what metric should we try to approximate p? Arguably the most natural approach would be to choose the total variation distance – in this case, the second step asks us to distinguish whether p and q are close or far in total variation distance. Unfortunately, this approach hits a roadblock: as shown by [VV17b] (and discussed in Chapter 2), this testing problem requires $\Omega(n/\log n)$ samples. As a result, the second step alone would seem to preclude an $O(\sqrt{n}/\varepsilon^2)$ sample algorithm.

To avoid this information-theoretic lower bound, we must instead consider a *relaxation* of the previous testing problem. In particular, rather than learning in total variation distance, we will instead consider the χ^2-*distance*. As a result, the second step will be realized as the following testing problem:

- p and q are $O(\varepsilon^2)$-close in χ^2-distance; this case corresponds to $p \in \mathcal{C}$.

- p and q are $\Omega(\varepsilon)$-far in total variation distance; this case corresponds to $d_{\mathrm{TV}}(p, \mathcal{C}) > \varepsilon$.

It can be verified (cf. Proposition 1) that this testing problem is easier than the previous one with total variation tolerance. Indeed, this is precisely the setting in which Theorem 4 (from Chapter 2) is designed to work[1], and the sample complexity of the second step drops to $O(\sqrt{n}/\varepsilon^2)$. Note that we achieve dramatic savings by considering the χ^2-distance rather

[1]Theorem 4 is actually designed to solve a *harder* problem (testing for farness in *Hellinger* distance), but the result for total variation follows from Proposition 1.

than the total variation distance – this serves as another motivation for studying distribution testing with alternative distances, as we have done in Chapter 2.

With this χ^2-tolerant identity testing primitive in place, we are ready to turn our focus to the testing of specific classes. In particular, it turns out the first step (obtaining an estimate q of the unknown distribution p) is quite cheap for many natural classes, including all the ones we consider, and thus, the overall cost of our test is dominated by the second step. While many of these learning problems have been studied significantly in the total variation setting (see, e.g., [CDSS14, ADLS17]), there has been less exploration when one considers χ^2-learning. Nonetheless, we show that this is still possible, with a mild increase in the sample complexity. However, the cost is still very cheap: for instance, estimating log-concave or monotone hazard rate distributions in χ^2-distance requires only $\mathrm{poly}(1/\varepsilon)$ samples, while estimating monotone or unimodal distributions requires $\mathrm{poly}(\log n, 1/\varepsilon)$ samples.

Our base tester is combined with the afore-mentioned extension of Birgé's decomposition theorem to test monotone distributions in Section 3.5 (see Theorem 14 and Corollary 4), and is also used to test independence of distributions in Section 3.7 (see Theorem 17).

Naturally, there are several bells and whistles that we need to add to the above skeleton to accommodate all classes of distributions that we are considering. For log-concave and monotone-hazard distributions, we are unable to obtain a cheap (in terms of samples) learner that χ^2-approximates the unknown distribution p throughout its support. Still, we can identify a subset of the support where the χ^2-approximation is tight and which captures almost all the probability mass of p. We extend our tester to accommodate excluding subsets of the support from the χ^2-approximation. See Theorems 18 and 19 in Sections 3.8 and 3.9.

For unimodal distributions, we are even unable to identify a large enough subset of the support where the χ^2-approximation is guaranteed to be tight. But we can show that there exists a light enough piece of the support (in terms of probability mass under p) that we can exclude to make the χ^2-approximation tight. Given that we only use Chebyshev's inequality to prove the concentration of the test statistic, it would seem that our lack of knowledge of the piece to exclude would involve a union bound and a corresponding increase in the required number of samples. We avoid this through a careful application of Kolmogorov's max inequality in our setting. See Theorem 16 of Section 3.6.

3.1.2 Related Work

Shape restrictions have played a vast role in probabilistic modeling and testing, and we are unable to cover this area in its entirety. It suffices to say that the classes of distributions that we study are fundamental, motivating extensive literature on their learning and testing [BBBB72]. In the recent times, there has been work on shape restricted statistics, pioneered by Jon Wellner, and others. [JW09, BW10] study estimation of monotone and k-monotone densities, and [BJRP13, SW14] study estimation of log-concave distributions.

As we have mentioned, statistics has focused on the asymptotic regime as the number of samples tends to infinity. Instead we are considering the low sample regime and are more stringent about the behavior of our testers, requiring two-sided guarantees. We want to accept if the unknown distribution is in our class of interest, and also reject if it is far from the class. For this problem, as discussed above, there are few results when $\mathcal{C}$ is a whole class of distributions. Closer related to our work is the line of papers [BKR04, ACS10, BFRV11] for monotonicity testing, albeit these papers have sub-optimal sample complexity as discussed above. Testing independence of random variables has a long history in statisics [RS81, Agr12]. The theoretical computer science community has also considered the problem of testing independence of random variables [BFF$^+$01, AAK$^+$07, LRR13, RX14]. While our results sharpen the case where the variables are over domains of equal size, they demonstrate an interesting asymmetric upper bound when this is not the case. More recently, Acharya and Daskalakis provide optimal testers for the family of Poisson Binomial Distributions [AD15].

Contemporaneous work of Canonne et al [CDGR16] provides a generic algorithm and lower bounds for the single-dimensional families of distributions considered here. We note that their algorithm has a sample complexity which is suboptimal in both n and ε, while our algorithms are optimal. Their algorithm also extends to mixtures of these classes, though some of these extensions are not computationally efficient. They also provide a framework for proving lower bounds, giving the optimal bounds for many classes when ε is sufficiently large with respect to $1/n$. In comparison, we provide these lower bounds unconditionally by modifying Paninski's construction [Pan08] to suit the classes we consider.

There has been a great deal of work on testing for structure subsequent to the initial pub-

lication of this work as [ADK15]. Our technique was applied and extended in [Can16] for testing k-histogram distributions, improving upon previous results for this problem by [ILR12]. [DK16] has improved results for testing independence on domains with different size in each dimension. [CDS17] applies Fourier techniques for testing additional classes of distributions. [BC17, DKS17] focus on the new problem of "generalized" uniformity testing: here, one wishes to test against the class of all distributions which are uniform over some (unknown) subset of $[n]$. [OZ18] work on the hypercube, testing whether distributions are uniform when restricted to any k coordinates. Finally, in a relatively new direction, [CDKS18] study testing for *conditional* independence.

Our work provides the first efficient algorithm for proper learning of log-concave distribution, even in total variation distance. After our work, [DKS16] provided a more efficient algorithm for this task.

The above works mentioned focus on testing *for* structure. There have also been a number of works on more efficient testing *with* structure, which is discussed more in Chapter 4.

3.2 Preliminaries

In this work, we will consider the following classes of distributions:

- Monotone distributions over $[n]^d$ (denoted by $\mathcal{M}_n^d$), for which $i \lesssim j$ implies $f_i \geq f_j$[2];
- Unimodal distributions over $[n]$ (denoted by $\mathcal{UM}_n$), for which there exists an i^* such that f_i is non-decreasing for $i \leq i^*$ and non-increasing for $i \geq i^*$;
- Log-concave distributions over $[n]$ (denoted by $\mathcal{LCD}_n$), the sub-class of unimodal distributions for which $f_{i-1}f_{i+1} \leq f_i^2$;
- Monotone hazard rate (MHR) distributions over $[n]$ (denoted by $\mathcal{MHR}_n$), for which $i < j$ implies $\frac{f_i}{1-F_i} \leq \frac{f_j}{1-F_j}$.

Definition 7. *An* η-effective support *of a distribution* p *is any set* S *such that* $p(S) \geq 1-\eta$.

[2]This definition describes monotone non-increasing distributions. By symmetry, identical results hold for monotone non-decreasing distributions.

The *flattening* of a function f over a subset S is the function $\bar{f}$ such that $\bar{f}_i = p(S)/|S|$.

Definition 8. *Let p be a distribution, and let $I_1, \ldots, I_k$ be a partition of the domain. The flattening of p with respect to $I_1, \ldots, I_k$ is the distribution $\bar{p}$ which is the flattening of p over the intervals $I_1, \ldots, I_k$.*

3.3 Overview

Our algorithm for testing a distribution p can be decomposed into three steps.

Near-proper learning in χ^2-distance. Our first step requires a learning algorithm with very specific guarantees. In proper learning, we are given sample access to a distribution $p \in \mathcal{C}$, where $\mathcal{C}$ is some class of distributions, and we wish to output $q \in \mathcal{C}$ such that p and q are close in total variation distance. In our setting, given sample access to $p \in \mathcal{C}$, we wish to output q such that q is *close* to $\mathcal{C}$ in total variation distance, and p and q are close in χ^2-distance on an effective support[3] of p. From an information theoretic standpoint, this problem is harder than proper learning, since χ^2-distance is more restrictive than total variation distance. Nonetheless, this problem can be shown to have comparable sample complexity to proper learning for the structured classes we consider.

Computation of distance to class. The next step is to see if the hypothesis q is close to the class $\mathcal{C}$ or not. Since we have an explicit description of q, this step requires no further samples from p, i.e. it is purely computational. If we find that q is far from the class $\mathcal{C}$, then it must be that $p \notin \mathcal{C}$, as otherwise the guarantees from the previous step would imply that q is close to $\mathcal{C}$. Thus, if it is not, we can terminate the algorithm at this point.

χ^2-testing. At this point, the previous two steps guarantee that our distribution q is such that:

- If $p \in \mathcal{C}$, then p and q are close in χ^2-distance on a (known) effective support of p;

[3]We also require the algorithm to output a description of an effective support for which this property holds. This requirement can be slightly relaxed, as we show in our results for testing unimodality.

- If $d_{\mathrm{TV}}(p, \mathcal{C}) \geq \varepsilon$, then p and q are far in total variation distance.

We can distinguish between these two cases using $O(\sqrt{n}/\varepsilon^2)$ samples with the statistical test of Theorem 4.

Using the above three-step approach, our tester, as described in the next section, can directly test monotonicity, log-concavity, and monotone hazard rate. With an extra modification, using Kolmogorov's max inequality, it can also test unimodality.

3.4 A Testing Framework

Our main result in this section is Theorem 13. This will follow from our χ^2-tolerant Hellinger identity tester (Theorem 4).

Theorem 13. *Suppose we are given $\varepsilon \in (0, 1]$, a class of probability distributions $\mathcal{C}$, sample access to a distribution p over $[n]$, and an explicit description of a distribution q with the following properties:*

Property 1. *$d_{\mathrm{TV}}(q, \mathcal{C}) \leq \frac{\varepsilon}{2}$.*

Property 2. *If $p \in \mathcal{C}$, then $d_{\chi^2}(p, q) \leq \frac{\varepsilon^2}{500}$.*

Then there exists an algorithm with the following guarantees:

- *If $p \in \mathcal{C}$, the algorithm outputs* Accept *with probability at least $2/3$;*
- *If $d_{\mathrm{TV}}(p, \mathcal{C}) \geq \varepsilon$, the algorithm outputs* Reject *with probability at least $2/3$.*

The time and sample complexity of this algorithm are $O\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$.

Proof. We note that combined with Proposition 1, Theorem 4 implies a test which can distinguish between the following two cases:

- $d_{\chi^2}(p, q) \leq \varepsilon^2/8$;
- $d_{\mathrm{TV}}(p, q) \geq \varepsilon/2$.

At this point, the proof follows by using this test on p and q. If $p \in \mathcal{C}$, then $d_{\chi^2}(p, q) \leq \frac{\varepsilon^2}{500}$, and we fall into the first case. On the other hand, if $d_{\mathrm{TV}}(p, \mathcal{C}) \geq \varepsilon$, then $d_{\mathrm{TV}}(q, \mathcal{C}) \leq \varepsilon/2$, and by triangle inequality, $d_{\mathrm{TV}}(p, q) \geq \varepsilon/2$. These two conditions fall into the two cases of the test, and hence we can distinguish them. □

3.4.1 Class-Specific Modifications

As stated in Theorem 13, Property 2 requires that q is $O(\varepsilon^2)$-close in χ^2-distance to p over its entire domain. For the class of monotone distributions, we are able to efficiently obtain such a q, which immediately implies sample-optimal learning algorithms for this class. However, for some classes, we cannot learn a q with such strong guarantees, and we must consider modifications to our base testing algorithm.

For example, for log-concave and monotone hazard rate distributions, we can obtain a distribution q and a set S with the following guarantees:

- If $p \in \mathcal{C}$, then $d_{\chi^2}(p_S, q_S) \leq O(\varepsilon^2)$ and $p(S) \geq 1 - O(\varepsilon)$;
- If $d_{\mathrm{TV}}(p, \mathcal{C}) \geq \varepsilon$, then $d_{\mathrm{TV}}(p, q) \geq \varepsilon/2$.

In this scenario, the tester will simply pretend the support of p and q is S, ignoring any samples and support elements in $[n] \setminus S$. Analysis of this tester is extremely similar to what was presented in Chapter 2. In particular, we can still show that the statistic Z will be separated in the two cases. When $p \in \mathcal{C}$, excluding $[n] \setminus S$ will only reduce Z. On the other hand, when $d_{\mathrm{TV}}(p, \mathcal{C}) \geq \varepsilon$, since $p(S) \geq 1 - O(\varepsilon)$, p and q must still be far on the remaining support, and we can show that Z is still sufficiently large. Therefore, a small modification allows us to handle this case with the same sample complexity of $O(\sqrt{n}/\varepsilon^2)$.

A further modification can handle even weaker learning guarantees. We could handle the previous case because the tester "knows what we don't know" – it can explicitly ignore the support over which we do not have a χ^2-closeness guarantee. A more difficult case is when there may be a low measure interval hidden in our effective support, over which p and q have a large χ^2-distance. While we may have insufficient samples to reliably identify this interval, it may still have a large effect on our statistic. A naive solution would be to consider a tester

which tries all possible "guesses" for this "bad" interval, but a union bound would incur an extra logarithmic factor in the sample complexity. We manage to avoid this cost through a careful analysis involving Kolmogorov's max inequality, maintaining the $O(\sqrt{n}/\varepsilon^2)$ sample complexity even in this more difficult case.

Being more precise, we can handle cases where we can obtain a distribution q and a set of intervals $S = \{I_1, \ldots, I_b\}$ with the following guarantees:

- If $p \in \mathcal{C}$, then $p(S) \geq 1 - O(\varepsilon)$, $p(I_j) = \Theta(p(S)/b)$ for all $j \in [b]$, and there exists a set $T \subseteq [b]$ such that $|T| \geq b - t$ (for $t = O(1)$) and $d_{\chi^2}(p_R, q_R) \leq O(\varepsilon^2)$, where $R = \cup_T I_j$;
- If $d_{\mathrm{TV}}(p, \mathcal{C}) \geq \varepsilon$, then $d_{\mathrm{TV}}(p, q) \geq \varepsilon/2$.

This allows us to additionally test against the class of unimodal distributions.

The tester requires that an effective support is divided into several intervals of roughly equal measure. It computes our statistic over each of these intervals, and we let our statistic Z be the sum of all but the largest t of these values. In the case when $p \in \mathcal{C}$, Z will only become smaller by performing this operation. We use Kolmogorov's maximal inequality to show that Z remains large when $d_{\mathrm{TV}}(p, \mathcal{C}) \geq \varepsilon$. More details on this tester are provided in Section 3.6.

3.5 Testing Monotonicity

As an application of our testing framework, we will demonstrate how to test for monotonicity. Let $d \geq 1$, and $\mathbf{i} = (i_1, \ldots, i_d), \mathbf{j} = (j_1, \ldots, j_d) \in [n]^d$. We say $\mathbf{i} \succcurlyeq \mathbf{j}$ if $i_l > j_l$ for $l = 1, \ldots, d$.

Definition 9. *A distribution p over $[n]^d$ is monotone (decreasing) if for all $\mathbf{i} \succcurlyeq \mathbf{j}$, $p_{\mathbf{i}} \leq p_{\mathbf{j}}$.*

Our main result of this section is as follows:

Theorem 14. *For any $d \geq 1$, there exists an algorithm for testing monotonicity over $[n]^d$ with sample complexity*

$$O\left(\frac{n^{d/2}}{\varepsilon^2} + \left(\frac{d \log n}{\varepsilon^2}\right)^d \cdot \frac{1}{\varepsilon^2}\right)$$

and time complexity $O\left(\frac{n^{d/2}}{\varepsilon^2} + \mathrm{poly}(\log n, 1/\varepsilon)^d\right)$.

In particular, this implies the following optimal algorithms for monotonicity testing for all $d \geq 1$:

Corollary 4. *Fix any $d \geq 1$, and suppose $\varepsilon > \frac{\sqrt{d \log n}}{n^{1/4}}$. Then there exists an algorithm for testing monotonicity over $[n]^d$ with sample complexity $O\left(n^{d/2}/\varepsilon^2\right)$.*

Our analysis starts with a structural lemma about monotone distributions. In [Bir87], Birgé showed that any monotone distribution p over $[n]$ can be *obliviously* decomposed into $O(\log(n)/\varepsilon)$ intervals, such that the flattening $\bar{p}$ (recall Definition 8) of p over these intervals is ε-close to p in total variation distance. [AJOS14a] extend this result, giving a bound between the χ^2-distance of p and $\bar{p}$. We strengthen these results by extending them to monotone distributions over $[n]^d$. In particular, we partition the domain $[n]^d$ of p into $O((d \log(n)/\varepsilon^2)^d)$ rectangles, and compare it with $\bar{p}$, the flattening over these rectangles.

Lemma 10. *Let $d \geq 1$. There is an oblivious decomposition of $[n]^d$ into $O((d \log(n)/\varepsilon^2)^d)$ rectangles such that for any monotone distribution p over $[n]^d$, its flattening $\bar{p}$ over these rectangles satisfy $d_{\chi^2}(p, \bar{p}) \leq \varepsilon^2$.*

This effectively reduces the support size to logarithmic in n. At this point, we can apply the Laplace estimator (along the lines of [KOPS15]) and learn a q such that if p was monotone, then q will be $O(\varepsilon^2)$-close in χ^2-distance.

Lemma 11. *Let $d \geq 1$, and p be a monotone distribution over $[n]^d$. There is an algorithm which outputs a distribution q such that $\mathbf{E}\left[d_{\chi^2}(p, q)\right] \leq \frac{\varepsilon^2}{500}$. The time and sample complexity are both $O((d \log(n)/\varepsilon^2)^d/\varepsilon^2)$.*

The final step before we apply our χ^2-tester is to compute the distance between q and $\mathcal{M}_n^d$. This subroutine is similar to the one introduced by [BKR04]. The key idea is to write a linear program, which searches for any distribution f which is close to q in total variation distance. We note that the desired properties of f (i.e., monotonicity, normalization, and ε-closeness to q) are easy to enforce as linear constraints. If we find that such an f exists, we will apply our χ^2-test to q. If not, we output REJECT, as this is sufficient evidence to conclude that $p \notin \mathcal{M}_n^d$. Note that the linear program operates over the oblivious decomposition used

in our structural result, so the complexity is polynomial in $(d\log(n)/\varepsilon)^d$, rather than the naive n^d.

At this point, we have precisely the guarantees needed to apply Theorem 13, directly implying Theorem 14. The proof of Lemmas 10 and 11 are in in Sections 3.5.1 and 3.5.2, respectively.

3.5.1 Structure of Monotone Distributions

Birgé [Bir87] showed that any monotone distribution is estimated to a total variation ε with a $O(\log(n)/\varepsilon)$-piecewise constant distribution. Moreover, the intervals over which the output is constant is independent of the distribution p. This result, was strengthened to the Kullback-Leibler divergence by [AJOS14a] to study the compression of monotone distributions. They upper bound the KL divergence by χ^2-distance and then bound the χ^2-distance. We extend this result to $[n]^d$. We divide $[n]^d$ into b^d rectangles as follows. Let $\{I_1, \ldots, I_b\}$ be a partition of $[n]$ into consecutive intervals defined as:

$$|I_j| = \begin{cases} 1 & \text{for } 1 \le j \le \frac{b}{2}, \\ \lfloor 2(1+\gamma)^{j-b/2} \rfloor & \text{for } \frac{b}{2} < j \le b. \end{cases}$$

For $\mathbf{j} = (j_1, \ldots, j_d) \in [b]^d$, let $I_{\mathbf{j}} \triangleq I_{j_1} \times I_{j_2} \times \ldots \times I_{j_d}$.

The χ^2-distance between p and $\bar{p}$ can be bounded as

$$\begin{aligned} d_{\chi^2}(p, \bar{p}) &= \left[\sum_{\mathbf{j} \in [b]^d} \sum_{\mathbf{i} \in I_{\mathbf{j}}} \frac{p_{\mathbf{i}}^2}{\bar{p}_{\mathbf{i}}^2} \right] - 1 \\ &\le \left[\sum_{\mathbf{j} \in [b]^d} p_{\mathbf{j}}^+ |I_{\mathbf{j}}| \right] - 1 \end{aligned}$$

For $\mathbf{j} = (j_1, \ldots, j_d) \in [b]^d$, let $\mathbf{j}^* = (j_1^*, \ldots, j_d^*)$ be

$$j_i^* = \begin{cases} j_i & \text{if } j_i \leq b/2 + 1 \\ j_i - 1 & \text{otherwise.} \end{cases}$$

We bound the expression above as follows.

Let $T \subseteq [d]$ be any subset of d. Suppose the size of T is ℓ. Let $\bar{T}$ be the set of all $\mathbf{j}$ that satisfy $\mathbf{j}_i = b/2 + 1$ for $i \in T$. In other words, over the dimensions determined by T, the value of the index is equal to $d/2 + 1$. The map $\mathbf{j} \to \mathbf{j}^*$ restricted to T is one-to-one, and since at most $d - \ell$ of the coordinates drop,

$$|I_{\mathbf{j}}| \leq |I_{\mathbf{j}^*}| \cdot (1+\gamma)^{d-\ell}.$$

Since there are ℓ coordinates that do not change, and each of them have $2(1+\gamma)$ coordinates, we obtain

$$\begin{aligned} \sum_{\mathbf{j} \in \bar{T}} p_{\mathbf{j}} &\leq \sum_{\mathbf{j} \in \bar{T}} p^-_{\mathbf{j}^*} \cdot |I_{\mathbf{j}}| \cdot (2(1+\gamma))^{\ell} \cdot (1+\gamma)^{d-\ell} \\ &= \sum_{\mathbf{j} \in \bar{T}} p^-_{\mathbf{j}^*} \cdot |I_{\mathbf{j}^*}| \cdot 2^{\ell}(1+\gamma)^{d}. \end{aligned}$$

Since the mapping is one-to-one, the probability of observing as element in $\bar{T}$ is the probability of observing $b/2 + 1$ in ℓ coordinates, which is at most $(2/(b+2))^{\ell}$ under any monotone distribution. Therefore,

$$\sum_{\mathbf{j} \in \bar{T}} p_{\mathbf{j}} \leq \left(\frac{2}{b+2}\right)^{\ell} \cdot 2^{\ell}(1+\gamma)^{d}.$$

For any ℓ there are $\binom{d}{\ell}$ choices for T. Therefore,

$$\begin{aligned} d_{\chi^2}(p, \bar{p}) &\leq \sum_{\ell=0}^{d} \binom{d}{\ell} \left(\frac{4}{b+2}\right)^{\ell} (1+\gamma)^d - 1 \\ &= (1+\gamma)^d \left(1 + \frac{4}{b+2}\right)^d - 1 \\ &= \left(1 + \gamma + \frac{4}{b+2} + \frac{4\gamma}{b+2}\right)^d - 1 \end{aligned}$$

Recall that $\gamma = 2\log(n)/b > 1/b$, implies that the expression above is at most $(1+2\gamma)^d - 1$. This implies Lemma 10.

3.5.2 Learning Monotone Distributions

Our algorithm requires a distribution q satisfying the properties discussed earlier. We learn a monotone distribution from samples as follows.

Before proving this result, we prove a general result for χ^2-learning of arbitrary discrete distributions, adapting the result from [KOPS15]. For a distribution p, and a partition of the domain into b intervals $I_1, \ldots, I_b$, let $\bar{p}_i = p(I_i)/|I_i|$ be the flattening of p over these intervals. We saw that for monotone distributions there exists a partition of the domain such that $\bar{p}$ is *close* to the underlying distribution in χ^2-distance.

Suppose we are given m samples from a distribution p and a partition $I_1, \ldots, I_b$. Let m_j be the number of samples that fall in I_j. For $i \in I_j$, let

$$q_i \triangleq \frac{1}{|I_j|} \frac{m_j + 1}{m + b}.$$

Let $S_j = \sum_{i \in I_j} p_i^2$. The expected χ^2-distance between p and q can be bounded as follows.

$$\begin{aligned}
\mathbf{E}\left[d_{\chi^2}(p,q)\right] &= \left[\sum_{j=1}^{b}\sum_{i\in I_j}\sum_{\ell=0}^{m}\binom{m}{\ell}(p(I_j))^{\ell}(1-p(I_j))^{m-\ell}\frac{p_i^2}{(\ell+1)/(|I_j|(m+b))}\right]-1 \\
&= \left[\frac{m+b}{m+1}\sum_{j=1}^{b}\frac{S_j}{\bar{p}(I_j)/|I_j|}\left(\sum_{\ell=0}^{m}\binom{m+1}{\ell+1}(p(I_j))^{\ell+1}(1-p(I_j))^{m+1-\ell+1}\right)\right]-1 \\
&= \left[\frac{m+b}{m+1}\sum_{j=1}^{b}\frac{S_j}{\bar{p}(I_j)/|I_j|}\left(1-(1-p(I_j)^{m+1}\right)\right]-1 \\
&\leq \left[\frac{m+b}{m+1}\sum_{j=1}^{b}\frac{S_j}{\bar{p}(I_j)/|I_j|}\right]-1 \\
&= \left[\frac{m+b}{m+1}\left(d_{\chi^2}(p,\bar{p})+1\right)\right]-1 \\
&= \frac{m+b}{m+1}\cdot d_{\chi^2}(p,\bar{p})+\frac{b}{m+1}. \qquad (3.1)
\end{aligned}$$

Suppose $\gamma = O(\log(n)/b)$, and $b = O(d \cdot \log(n)/\varepsilon^2)$. Then, by Lemma 10,

$$d_{\chi^2}(p,\bar{p}) \leq \varepsilon^2. \qquad (3.2)$$

Combining this with (3.1) gives Lemma 11.

3.6 Testing Unimodality

One striking feature of Birgé's result is that the decomposition of the domain is oblivious to the samples, and therefore to the unknown distribution. However, such an oblivious decomposition will not work for the unimodal distribution, since the mode is unknown. Suppose we know where the mode of the unknown distribution might be, then the problem can be decomposed into monotone functions over two intervals. Therefore, in theory, one can modify the monotonicity testing algorithm by iterating over all the possible n modes. Indeed, by applying a union bound, it then follows that

Theorem 15 (Follows from Theorem 14)**.** *For $\varepsilon > 1/n^{1/4}$, there exists an algorithm for testing unimodality over $[n]$ with sample complexity $O\left(\frac{\sqrt{n}}{\varepsilon^2}\log n\right)$.*

However, this is unsatisfactory, since our lower bound (and as we will demonstrate, the true complexity of this problem) is $\sqrt{n}/\varepsilon^2$. We overcome the logarithmic barrier introduced by the union bound, by employing a non-oblivious decomposition of the domain, and using Kolmogorov's max-inequality.

Our main result for testing unimodality is the following theorem.

Theorem 16. *Suppose $\varepsilon > n^{-1/4}$. Then there exists an algorithm for testing unimodality over $[n]$ with sample complexity $O(\sqrt{n}/\varepsilon^2)$.*

Proof. Recall that to circumvent Birgé's decomposition, we want to decompose the interval into disjoint intervals such that the probability of each interval is about $O(1/b)$, where b is a parameter, specified later. In particular we consider a decomposition of $[n]$ with the following properties:

1. For each element i with probability at least $1/b$, there is an $I_\ell = \{i\}$.

2. There are at most two intervals with $p(I) \leq 1/2b$.

3. Every other interval I satisfies $p(I) \in \left[\frac{1}{2b}, \frac{2}{b}\right]$.

Let $I_1, \ldots, I_L$ denote the partition of $[n]$ corresponding to these intervals. Note that $L = O(b)$.

Claim 1. *There is an algorithm that takes $O(b \log b)$ samples and outputs $I_1, \ldots, I_L$ satisfying the properties above.*

The first step in our algorithm is to estimate the *total probability* within each of these intervals. In particular,

Lemma 12. *There is an algorithm that takes $m' = O(b \log b/\varepsilon^2)$ samples from a distribution p, and with probability at least 9/10 outputs a distribution $\bar{q}$ that is constant on each I_L. Moreover, for any j such that $p(I_j) > 1/2b$, $\bar{q}(I_j) \in (1 \pm \varepsilon)p(I_j)$.*

Proof. Consider any interval I_j with $p(I_j) \geq 1/2b$. The number of samples N_{I_j} that fall in that interval is distributed as $Binomial(m', p(I_j))$. Then by Chernoff bounds for $m' >$

$12b \log b/\varepsilon^2$,

$$\Pr\left[\, |N_{I_j} - m'p(I_j)| > \varepsilon m'p(I_j) \,\right] \leq 2\exp\left(\varepsilon^2 m'p(I_j)/2\right) \tag{3.3}$$

$$\leq \frac{1}{b^2}, \tag{3.4}$$

where the last inequality uses the fact that $p(I_j) \geq 1/2b$. □

The next step is estimate the distance of q from $\mathcal{U}_n$. This is possible by a simple dynamic program, similar to the one used for monotonicity. If the estimated distance is more than $\varepsilon/2$, we output REJECT.

Our next step is to remove certain intervals. This will be to ensure that when the underlying distribution is unimodal, we are able to estimate the distribution *multiplicatively* over the remaining intervals. In particular, we do the following preprocessing step:

- $A = \emptyset$.
- For interval I_j,
  - If

$$q(I_j) \notin ((1-\varepsilon)\cdot q(I_{j+1}), (1+\varepsilon)\cdot q(I_{j+1})) \quad \text{OR} \tag{3.5}$$

$$q(I_j) \notin ((1-\varepsilon)\cdot q(I_{j-1}), (1+\varepsilon)\cdot q(I_{j-1}))\,, \tag{3.6}$$

    add I_j to A.
- Add the (at most 2) intervals with mass at most $1/2b$ to A.
- Add all intervals j with $q(I_j)/|I_j| < \varepsilon/50n$ to A

If the distribution is unimodal, we can prove the following about the set of intervals A^c.

Lemma 13. *If p is unimodal then,*

- $p(I_{A^c}) \geq 1 - \varepsilon/25 - 1/b - O\left(\log n/(\varepsilon b)\right).$

- *Except* at most one *interval in* A^c *every other interval* I_j *satisfies,*

$$\frac{p_j^+}{p_j^-} \leq (1+\varepsilon).$$

If this holds, then the χ^2-distance between p and q constrained to A^c, is at most ε^2. This lemma follows from the following result.

Lemma 14. *Let* $C > 2$. *For a unimodal distribution over* $[n]$, *there are at most* $\frac{4\log(50n/\varepsilon)}{C\varepsilon}$ *intervals* I_j *that satisfy* $\frac{p_j^+}{p_j^-} < (1+\varepsilon/C)$.

Proof. To the contrary, if there are more than $\frac{4\log(50n/\varepsilon)}{C\varepsilon}$ intervals, then at least half of them are on one side of the mode, however this implies that the ratio of the largest probability and smallest probability is at least $(1+\varepsilon/C)^j$, and if $j > \frac{2\log(50n/\varepsilon)}{C\varepsilon}$, is at least $50n/\varepsilon$, contradicting that we have removed all such elements. □

We have one additional pre-processing step here. We compute $q(A^c)$ and if it is smaller than $1-\varepsilon/25$, we output Reject.

Suppose there are L' intervals in A^c. Then, except at most one interval in L' we know that the χ^2-distance between p and q is at most ε^2 when p is unimodal, and the TV distance between p and q is at least $\varepsilon/2$ over A^c. We propose the following simple modification to take into account, the one interval that might introduce a high χ^2-distance in spite of having a small total variation. If we knew the interval, we can simply remove it and proceed. Since we do not know where the interval lies, we do the following.

1. Let Z_j be the χ^2-statistic over the ith interval in A^c, computed with $O(\sqrt{n}/\varepsilon^2)$ samples.
2. Let Z_l be the largest among all Z_j's.
3. If $\sum_{j,j\neq l} Z_j > m\varepsilon^2/10$, output Reject.
4. Output Accept.

The objective of removing the largest χ^2-statistic is our substitute for not knowing the largest interval. We now prove the correctness of this algorithm.

Case 1 $p \in \mathcal{UM}_n$**:** We only concentrate on the final step. The χ^2-statistic over all but one interval are at most $c \cdot m\varepsilon^2$, and the variance is bounded as before. Since we remove the largest statistic, the expected value of the new statistic is *strictly dominated* by that of these intervals. Therefore, the algorithm outputs ACCEPT with at least the same probability as if we removed the spurious interval.

Case 2 $p \notin \mathcal{UM}_n$**:** This is the hard case to prove for unimodal distributions. We know that the χ^2-statistic is large in this case, and we therefore have to prove that it remains large even after removing the largest test statistic Z_l.

We invoke Kolmogorov's Maximal Inequality to this end.

Lemma 15 (Kolmogorov's Maximal Inequality)**.** *For independent zero mean random variables $X_1, \ldots, X_L$ with finite variance, let $S_\ell = X_1 + \ldots X_\ell$. Then for any $\lambda > 0$,*

$$\Pr\left[\max_{1 \le \ell \le L} |S_\ell| \ge \lambda\right] \le \frac{1}{\lambda^2} \cdot \mathbf{Var}\left(S_L\right). \tag{3.7}$$

As a corollary, it follows that $\Pr[\max_\ell |X_\ell| > 2\lambda] \le \frac{1}{\lambda^2} \cdot \mathbf{Var}\left(S_L\right)$.

In the case we are interested in, we let $X_i = Z_\ell - \mathbf{E}\left[Z_\ell\right]$. Then, similar to the computations before, and the fact that each interval has a small mass, it follows that that the variance of the summation is at most $\mathbf{E}\left[Z_\ell\right]^2/100$. Taking $\lambda = \mathbf{E}\left[S_L - m\varepsilon^2/3\right]^2/100$, it follows that the statistic does not fall below to $\sqrt{n}$. □

3.7 Testing Independence

Let $\mathcal{X} \triangleq [n_1] \times \ldots \times [n_d]$, and let Π_d be the class of all product distributions over $\mathcal{X}$. We first bound the χ^2-distance between product distributions in terms of the individual coordinates.

Lemma 16. *Let $p = p^1 \times p^2 \ldots \times p^d$, and $q = q^1 \times q^2 \ldots \times q^d$ be two distributions in Π_d. Then*

$$d_{\chi^2}\left(p, q\right) = \prod_{\ell=1}^{d} (1 + d_{\chi^2}\left(p^\ell, q^\ell\right)) - 1.$$

Proof. By the definition of χ^2-distance

$$d_{\chi^2}(p,q) = \sum_{\mathbf{i}\in\mathcal{X}} \frac{\left(p_i^\ell\right)^2}{q_i^\ell} - 1 \tag{3.8}$$

$$= \prod_{\ell=1}^{d} \left[\sum_{i\in[n_\ell]} \frac{\left(p_i^\ell\right)^2}{q_i^\ell} \right] - 1 \tag{3.9}$$

$$= \prod_{\ell=1}^{d} \left(1 + d_{\chi^2}\left(p^\ell, q^\ell\right)\right) - 1. \tag{3.10}$$

□

Along the lines of learning monotone distributions in χ^2-distance we obtain the following result.

Lemma 17. *There is an algorithm that takes*

$$O\left(\sum_{\ell=1}^{d} \frac{n_\ell}{\varepsilon^2}\right)$$

samples from a distribution p in Π_d and outputs a distribution $q \in \Pi_d$ such that with probability at least $5/6$,

$$d_{\chi^2}(p,q) \le O(\varepsilon^2).$$

This fits precisely in our framework of tolerant χ^2-ℓ_1 testing. In particular, applying Theorem 13, we obtain the following result.

Theorem 17. *For any $d \ge 1$, there exists an algorithm for testing independence of random variables over $[n_1] \times \ldots [n_d]$ with sample and time complexity*

$$O\left(\frac{(\prod_{\ell=1}^{d} n_\ell)^{1/2} + \sum_{\ell=1}^{d} n_\ell}{\varepsilon^2}\right).$$

The following corollaries are immediate.

Corollary 5. *Suppose $\prod_{\ell=1}^{d} n_\ell^{1/2} \ge \sum_{\ell=1}^{d} n_\ell$. Then there exists an algorithm for testing independence over $[n_1] \times \cdots \times [n_d]$ with sample complexity $\Theta((\prod_{\ell=1}^{d} n_\ell)^{1/2}/\varepsilon^2)$.*

In particular,

Corollary 6. *There exists an algorithm for testing if two distributions over $[n]$ are independent with sample complexity $\Theta(n/\varepsilon^2)$.*

We conclude by proving Lemma 17.

Proof of Lemma 17: In this section we prove Lemma 17. The proof is analogous to the proof for learning monotone distributions, and hinges on the following result of [KOPS15]. Given m samples from a distribution q over n elements, the add-1 estimator (Laplace estimator) q satisfies:

$$\mathbf{E}\left[d_{\chi^2}(p,q)\right] \leq \frac{n}{m+1}.$$

Now, suppose p is a product distribution over $\mathcal{X} = [n_1] \times \cdots \times [n_d]$. We simply perform the add-1 estimation over each coordinate independently, giving a distribution $q^1 \times \cdots \times q^d$. Since p is a product distribution the estimates in each coordinate is independent. Therefore, a simple application of the previous result and independence of the coordinates implies

$$\begin{aligned}
\mathbf{E}\left[d_{\chi^2}(p,q)\right] &= \prod_{l=1}^{d} \left(1 + \mathbf{E}\left[d_{\chi^2}(p^l,q^l)\right]\right) - 1 \\
&\leq \prod_{l=1}^{d} \left(1 + \frac{n_l}{m+1}\right) - 1 \\
&\leq \exp\left(\frac{\sum_l n_l}{m+1}\right) - 1, \qquad (3.11)
\end{aligned}$$

where (3.11) follows from $e^x \geq 1 + x$. Using $e^x \leq 1 + 2x$ for $0 \leq x \leq 1$, we have

$$\mathbf{E}\left[d_{\chi^2}(p,q)\right] \leq 2\frac{\sum_l n_l}{m+1}, \qquad (3.12)$$

when $m \geq \sum_l n_l$. Therefore, following an application of Markov's inequality, when $m = \Omega((\sum_l n_l)/\varepsilon^2)$, Lemma 17 is proved. □

3.8 Testing Log-Concavity

In this section we describe our results for testing log-concavity of distributions. Our main result is as follows:

Theorem 18. *There exists an algorithm for testing log-concavity over $[n]$ with sample complexity*

$$O\left(\frac{\sqrt{n}}{\varepsilon^2} + \frac{1}{\varepsilon^5}\right)$$

and time complexity $\mathrm{poly}(n, 1/\varepsilon)$.

In particular, this implies the following optimal tester for this class:

Corollary 7. *Suppose $\varepsilon > 1/n^{1/5}$. Then there exists an algorithm for testing log-concavity over $[n]$ with sample complexity $O\left(\sqrt{n}/\varepsilon^2\right)$.*

Our algorithm will fit into the structure of our general framework. We first perform a very particular type of learning algorithm, whose guarantees are summarized in the following lemma:

Lemma 18. *Given $\varepsilon > 0$ and sample access to a distribution p, there exists an algorithm with the following guarantees:*

- *If $p \in \mathcal{LCD}_n$, the algorithm outputs a distribution $q \in \mathcal{LCD}_n$ and an $O(\varepsilon)$-effective support S of p such that $d_{\chi^2}(p_S, q_S) \leq \frac{\varepsilon^2}{500}$ with probability at least $5/6$;*
- *If $d_{\mathrm{TV}}(p, \mathcal{LCD}_n) \geq \varepsilon$, the algorithm either outputs a distribution $q \in \mathcal{LCD}_n$ or* Reject.

The sample complexity is $O(1/\varepsilon^5)$ and the time complexity is $\mathrm{poly}(n, 1/\varepsilon)$.

We note that as a corollary, one immediately obtains a $O(1/\varepsilon^5)$ proper learning algorithm for log-concave distributions. The result is immediate from the first item of Lemma 18 and Proposition 1. We can actually do a bit better – in the proof of Lemma 18, we partition $[n]$ into intervals of probability mass $\Theta(\varepsilon^{3/2})$. If one instead partitions into intervals of probability mass $\Theta(\varepsilon/\log(1/\varepsilon))$ and works directly with total variation distance instead of χ^2-distance, one can show that $\tilde{O}(1/\varepsilon^4)$ samples suffice.

Corollary 8. *Given $\varepsilon > 0$ and sample access to a distribution $p \in \mathcal{LCD}_n$, there exists an algorithm which outputs a distribution $q \in \mathcal{LCD}_n$ such that $d_{\mathrm{TV}}(p, q) \leq \varepsilon$. The sample complexity is $\tilde{O}(1/\varepsilon^4)$ and the time complexity is* $\mathrm{poly}(n, 1/\varepsilon)$.

Then, given the guarantees of Lemma 18, Theorem 18 follows from Theorem 13[4].

Proof of Lemma 18: We first draw samples from p and obtain a $O(1/\varepsilon^{3/2})$-piecewise constant distribution f by appropriately flattening the empirical distribution. The proof is now in two parts. In the first part, we show that if $p \in \mathcal{LCD}_n$ then f will be close to p in χ^2-distance over its effective support. The second part involves proper learning of p. We will use a linear program on f to find a distribution $q \in \mathcal{LCD}_n$. This distribution is such that if $p \in \mathcal{LCD}_n$, then $d_{\chi^2}(p, q)$ is small, and otherwise the algorithm will either output some $q \in \mathcal{LCD}_n$ (with no other relevant guarantees) or Reject.

We first construct f. Let $\hat{p}$ be the empirical distribution obtained by sampling $O(1/\varepsilon^5)$ samples from p. By Lemma 1, with probability at least $5/6$, $d_{\mathrm{K}}(p, \hat{p}) \leq \varepsilon^{5/2}/10$. In particular, note that $|p_i - \hat{p}_i| \leq \varepsilon^{5/2}/10$. Condition on this event in the remainder of the proof.

Let a be the minimum i such that $p_i \geq \varepsilon^{3/2}/5$, and let b be the maximum i satisfying the same condition. Let $M = \{a, \ldots, b\}$ or $\emptyset$ if a and b are undefined. By the guarantee provided by the DKW inequality, $p_i \geq \varepsilon^{3/2}/10$ for all $i \in M$. Furthermore, $\hat{p}_i \in p_i \pm \varepsilon^{3/2}/10 \in (1 \pm \varepsilon) \cdot p_i$. For each $i \in M$, let $f_i = \hat{p}_i$. We note that $|M| = O(1/\varepsilon)$, so this contributes $O(1/\varepsilon)$ constant pieces to f.

We now divide the rest of the domain into t intervals, all but constantly many of measure $\Theta(\varepsilon^{3/2})$ (under p). This is done via the following iterative procedure. As a base case, set $r_0 = 0$. Define I_j as $[l_j, r_j]$, where $l_j = r_{j-1} + 1$ and r_j is the largest $j \in [n]$ such that $\hat{p}(I_j) \leq 9\varepsilon^{3/2}/10$. The exception is if I_j would intersect M – in this case, we "skip" M: set $r_j = a - 1$ and $l_{j+1} = b + 1$. If such a j exists, denote it by j^*. We note that $p(I_j) \leq \hat{p}(I_j) + \varepsilon^{5/2}/10 \leq \varepsilon^{3/2}$. Furthermore, for all j except j^* and t, $r_j + 1 \notin M$, so $p(I_j) \geq 9\varepsilon^{3/2}/10 - \varepsilon^{3/2}/5 - \varepsilon^{5/2}/10 \geq 3\varepsilon^{3/2}/5$. Observe that this lower bound implies that $t \leq \frac{2}{\varepsilon^{3/2}}$ for ε sufficiently small.

[4]To be more precise, we require the modification of Theorem 13 which is described in Section 3.4.1, in order to handle the case where the χ^2-distance guarantees only hold for a known effective support.

Part 1. For this part of the algorithm, we only care about the guarantees when $p \in \mathcal{LCD}_n$, so we assume this is the case.

For the domain $[n] \setminus M$, we let f be the flattening of $\hat{p}$ over the intervals $I_1, \dots I_t$. To analyze f, we need a structural property of log-concave distributions due to Chan, Diakonikolas, Servedio, and Sun [CDSS14]. This essentially states that a log-concave distribution cannot have a sudden increase in probability.

Lemma 19 (Lemma 4.1 in [CDSS14])**.** *Let p be a distribution over $[n]$ that is non-decreasing and log-concave on $[1, x] \subseteq [n]$. Let $I = [x, y]$ be an interval of mass $P(I) = \tau$, and suppose that the interval $J = [1, x-1]$ has mass $p(J) = \sigma > 0$. Then*

$$p(y)/p(x) \leq 1 + \tau/\sigma.$$

Recall that any log-concave distribution is unimodal, and suppose the mode of p is at i_0. We will first focus on the intervals $I_1, \dots, I_{t_L}$ which lie entirely to the left of i_0 and M. We will refer to I_j as L_j for all $j \leq t_L$. Note that p is non-decreasing over these intervals.

The next steps to the analysis are as follows. First we show that the flattening of p over L_j is a multiplicative $(1 + O(1/j))$ estimate for each $p_i \in L_j$. Then, we show that flattening the empirical distribution $\hat{p}$ over L_j is a multiplicative $(1 + O(1/j))$ estimate of $p(i)$ for each $i \in L_j$. Finally, we exclude a small number of intervals (those corresponding to $O(\varepsilon)$ mass at the left and right side of the domain, as well as j^*) in order to get the χ^2-approximation we desire on an effective support.

- First, recall that $p(L_j) \leq \varepsilon^{3/2}$ for all j. Also, letting $J_j = [1, r_{j-1}]$, we have that $p(J_j) \geq (j-1) \cdot 3\varepsilon^{3/2}/5$. Thus by Lemma 19, $p(r_j) \leq p(l_j)(1 + 2/(j-1))$. Since the distribution is non-decreasing in L_j, the flattening $\bar{p}$ of p is such that $\bar{p}(i) \in p(i)(1 \pm \frac{2}{j-1})$ for all $i \in L_j$.

- We have that $p(L_j) \geq 3\varepsilon^{3/2}/5$, and $\hat{p}(L_j) \in p(L_j) \pm \varepsilon^{5/2}/10$, so $\hat{p}(L_j) \in p(L_j) \cdot (1 \pm \frac{\varepsilon}{6})$, and hence $\hat{p}(i) \in \bar{p}(i) \cdot (1 \pm \frac{\varepsilon}{6})$ for all $i \in L_j$. Combining with the previous point, we

have that

$$\hat{p}(i) \in p(i) \cdot \left(1 \pm \left(\frac{2\varepsilon}{3(j-1)} + \frac{\varepsilon}{6} + \frac{2}{j-1}\right)\right) \in p(i) \cdot \left(1 \pm \frac{11}{3(j-1)}\right).$$

A symmetric statement holds for the intervals that lie entirely to the right of i_0 and M. We will refer to I_j as R_{t-j} for all $j > t_L$.

To summarize, we have the following guarantees for the distribution f:

- For all $i \in M$, $f(i) \in p(i) \cdot (1 \pm \varepsilon)$;
- For all $i \in L_j$ (except L_1 and L_{j^*}), $f(i) \in p(i) \cdot \left(1 \pm \frac{22}{3j}\right)$;
- For all $i \in R_j$ (except R_1), $f(i) \in p(i) \cdot \left(1 \pm \frac{22}{3j}\right)$;

Note that, in particular, we have multiplicative estimates for all intervals, except those in L_1, L_{j^*}, R_1 and the interval containing i_0. Let S be the set of all intervals except L_{j^*}, L_j and R_j for $j \leq 1/\sqrt{\varepsilon}$, and the one containing i_0 Then, since each interval has probability mass at most $O(\varepsilon^{3/2})$ and we are excluding $O(1/\sqrt{\varepsilon})$ intervals, $p(S) > 1 - O(\varepsilon)$.

We now compute the χ^2-distance induced by this approximation for elements in S. For an element $i \in L_j \cap S$, we have

$$\frac{(f(i) - p(i))^2}{p(i)} \leq \frac{60p(i)}{j^2}.$$

Summing over all $i \in L_j \cap S$ gives

$$\frac{60\varepsilon^{3/2}}{j^2}$$

since the probability mass of L_j is at most $\varepsilon^{3/2}$. Summing this over all L_j for $j \geq 1/\sqrt{\varepsilon}$ and $j \neq j^*$ gives

$$\begin{aligned} 60\varepsilon^{3/2} \sum_{j=1/\sqrt{\varepsilon}}^{2/\varepsilon^{3/2}} \frac{1}{j^2} &\leq 60\varepsilon^{3/2} \int_{1/\sqrt{\varepsilon}}^{\infty} \frac{1}{x^2} dx \\ &= 60\varepsilon^{3/2}(\sqrt{\varepsilon}) \\ &= O(\varepsilon^2) \end{aligned}$$

as desired.

Part 2. To obtain a distribution $q \in \mathcal{LCD}_n$, we write a linear program. We will work in the log domain, so our variables will be Q_i, representing $\log q(i)$ for $i \in [n]$. We will use $F_i = \log f(i)$ as parameters in our LP. There will be no objective function, we simply search for a feasible point. Our constraints will be

$$Q_{i-1} + Q_{i+1} \leq 2Q_i \quad \forall i \in [n-1]$$

$$Q_i \leq 0 \quad \forall i \in [n]$$

$$\log(1+\varepsilon) \leq |Q_i - F_i| \leq \log(1+\varepsilon) \text{ for } i \in M$$

$$\log\left(1 - \frac{22}{3j}\right) \leq |Q_i - F_i| \leq \log\left(1 + \frac{22}{3j}\right) \text{ for } i \in L_j, j \geq 1/\sqrt{\varepsilon} \text{ and } j \neq j^*$$

$$\log\left(1 - \frac{22}{3j}\right) \leq |Q_i - F_i| \leq \log\left(1 + \frac{22}{3j}\right) \text{ for } i \in R_j, j \geq 1/\sqrt{\varepsilon}$$

If we run the linear program, then after a rescaling and summing the error over all the intervals in the LP gives us that the distance between p and q to be $O(\varepsilon^2)$ χ^2-distance in a set S which has measure $p(S) \geq 1 - 4\varepsilon$, as desired.

If the linear program finds a feasible point, then we obtain a $q \in \mathcal{LCD}_n$. Furthermore, if $p \in \mathcal{LCD}_n$, this also tells us that (after a rescaling of ε), summing the error over all intervals implies that $d_{\chi^2}(p_S, q_S) \leq \frac{\varepsilon^2}{500}$ for a known set S with $p(S) \geq 1 - O(\varepsilon)$, as desired. If $M \neq \emptyset$, this algorithm works as described. The issue is if $M = \emptyset$, then we don't know when the L intervals end and the R intervals begin. In this case, we run $O(1/\varepsilon)$ LPs, using each interval as the one containing i_0, and thus acting as the barrier between the L intervals (to its left) and the R intervals (to its right). If p truly was log-concave, then one of these guesses will be correct and the corresponding LP will find a feasible point. □

3.9 Testing Monotone Hazard Rate

In this section, we obtain our main result for testing for monotone hazard rate:

Theorem 19. *There exists an algorithm for testing monotone hazard rate over* $[n]$ *with sample complexity*

$$O\left(\frac{\sqrt{n}}{\varepsilon^2} + \frac{\log(n/\varepsilon)}{\varepsilon^4}\right)$$

and time complexity $\mathrm{poly}(n, 1/\varepsilon)$.

This implies the following optimal tester for the class:

Corollary 9. *Suppose* $\varepsilon > \sqrt{\log(n/\varepsilon)}/n^{1/4}$. *Then there exists an algorithm for testing monotone hazard rate over* $[n]$ *with sample complexity* $O\left(\sqrt{n}/\varepsilon^2\right)$.

We obey the same framework as before, first applying a χ^2-learner with the following guarantees:

Lemma 20. *Given* $\varepsilon > 0$ *and sample access to a distribution* p, *there exists an algorithm with the following guarantees:*

- *If* $p \in \mathcal{MHR}_n$, *the algorithm outputs a distribution* $q \in \mathcal{MHR}_n$ *and an* $O(\varepsilon)$*-effective support* S *of* p *such that* $d_{\chi^2}(p_S, q_S) \leq \frac{\varepsilon^2}{500}$ *with probability at least* $5/6$;
- *If* $d_{\mathrm{TV}}(p, \mathcal{MHR}_n) \geq \varepsilon$, *the algorithm either outputs a distribution* $q \in \mathcal{MHR}_n$ *and a set* $S \subseteq [n]$ *or* Reject.

The sample complexity is $O(\log(n/\varepsilon)/\varepsilon^4)$ *and the time complexity is* $\mathrm{poly}(n, 1/\varepsilon)$.

As with log-concave distributions, this implies the following proper learning result:

Corollary 10. *Given* $\varepsilon > 0$ *and sample access to a distribution* $p \in \mathcal{MHR}_n$, *there exists an algorithm which outputs a distribution* $q \in \mathcal{MHR}_n$ *such that* $d_{\mathrm{TV}}(p, q) \leq \varepsilon$. *The sample complexity is* $O(\log(n/\varepsilon)/\varepsilon^4)$ *and the time complexity is* $\mathrm{poly}(n, 1/\varepsilon)$.

Again, combining the learning guarantees of Lemma 20 with the appropriate variant of Theorem 13 (cf. Section 3.4.1), we obtain Theorem 19.

Proof of Lemma 20: As with log-concave distributions, our method for MHR distributions can be split into two parts. In the first step, if $p \in \mathcal{MHR}_n$, we obtain a distribution q which is $O(\varepsilon^2)$-close to p in χ^2-distance on a set $\mathcal{A}$ of intervals such that $p(\mathcal{A}) \geq 1 - O(\varepsilon)$. q

will achieve this by being a multiplicative $(1 + O(\varepsilon))$ approximation for each element within these intervals. This step is very similar to the decomposition used for unimodal distributions (described in Section 3.6), so we sketch the argument and highlight the key differences.

The second step will be to find a feasible point in a linear program. If $p \in \mathcal{MHR}_n$, there should always be a feasible point, indicating that q is close to a distribution in $\mathcal{MHR}_n$ (leveraging the particular guarantees for our algorithm for generating q). If $d_{\mathrm{TV}}(p, \mathcal{MHR}_n) \geq \varepsilon$, there may or may not be a feasible point, but when there is, it should imply the existence of a distribution $p^* \in \mathcal{MHR}_n$ such that $d_{\mathrm{TV}}(q, p^*) \leq \varepsilon/2$.

The analysis will rely on the following lemma from [CDSS14], which roughly states that an MHR distribution is "almost" non-decreasing.

Lemma 21 (Lemma 5.1 in [CDSS14])**.** *Let p be an MHR distribution over $[n]$. Let $I = [a, b] \subset [n]$ be an interval, and $R = [b + 1, n]$ be the elements to the right of I. Let $\eta = p(I)/p(R)$. Then $p(b + 1) \geq \frac{1}{1+\eta} p(a)$.*

Part 1. As before, with unimodal distributions, we start by taking $O(\frac{b \log b}{\varepsilon^2})$ samples, with the goal of partitioning the domain into intervals of mass approximately $\Theta(1/b)$. First, we will ignore the left and rightmost intervals of mass $\Theta(\varepsilon)$. For all "heavy" elements with mass $\geq \Theta(1/b)$, we consider them as singletons. We note that Lemma 21 implies that there will be at most $O(1/\varepsilon)$ contiguous intervals of such elements. The rest of the domain is greedily divided (from left to right) into intervals of mass $\Theta(1/b)$, cutting an interval short if we reach one of the heavy elements. This will result in the guarantee that all but potentially $O(1/\varepsilon)$ intervals have $\Theta(1/b)$ mass.

Next, similar to unimodal distributions, considering the flattened distribution, we discard all intervals for which the per-element probability is not within a $(1 \pm O(\varepsilon))$ multiplicative factor of the same value for both neighboring intervals. The claim is that all remaining intervals will have the property that the per-element probability is within a $(1 \pm O(\varepsilon))$ multiplicative factor of the true probability. This is implied by Lemma 21. If there were a point in an interval which was above this range, the distribution must decrease slowly, and the next interval would have a much larger per-element weight, thus leading to the removal of this interval. A similar argument forbids us from missing an interval which contains a

point that lies outside this range. Relying on the fact that truncating the left and rightmost intervals eliminates elements with low probability mass, similar to the unimodal case, one can show that we will remove at most $\log(n/\varepsilon)/\varepsilon$ intervals, and thus a $\log(n/\varepsilon)/b\varepsilon$ probability mass. Choosing $b = \Omega(\varepsilon^2/\log(n/\varepsilon))$ limits this to be $O(\varepsilon)$, as desired. At this point, if p is indeed MHR, the multiplicative estimates guarantee that the result is $O(\varepsilon^2)$-close in χ^2-distance among the remaining intervals.

Part 2. We note that an equivalent condition for distribution f being MHR is log-concavity of $\log(1-F)$, where F is the CDF of f. Therefore, our approach for this part will be similar to the approach used for log-concave distributions.

Given the output distribution q from the previous part of this algorithm, our goal will be check if there exists an MHR distribution f which is $O(\varepsilon)$-close to q. We will run a linear program with variables $\mathfrak{f}_i = \log(1 - F_i)$. First, we ensure that f is a distribution. This can be done with the following constraints:

$$\begin{aligned} \mathfrak{f}_i &\leq 0 \qquad \forall i \in [n] \\ \mathfrak{f}_i &\geq \mathfrak{f}_{i+1} \qquad \forall i \in [n-1] \\ \mathfrak{f}_n &= -\infty \end{aligned}$$

To ensure that f is MHR, we use the following constraint:

$$\mathfrak{f}_{i-1} + \mathfrak{f}_{i+1} \leq 2\mathfrak{f}_i \qquad \forall i \in [2, n-1]$$

Now, ideally, we would like to ensure f and q are ε-close in total variation distance by ensuring they are pointwise within a multiplicative $(1 \pm \varepsilon)$ factor of each other:

$$(1 - \varepsilon) \leq f_i/q_i \leq (1 + \varepsilon)$$

We note that this is a stronger condition than f and q being ε-close, but if $p \in \mathcal{MHR}_n$, the guarantees of the previous step would imply the existence of such an f.

We have a separate treatment for the identified singletons (i.e., those with probability

$\geq 1/b$) and the remainder of the support. For each element q_i identified to have $\geq 1/b$ mass, we add two constraints:

$$\log((1-\varepsilon/2b)(1-Q_i)) \leq \mathfrak{f}_i \leq \log((1+\varepsilon/2b)(1-Q_i))$$

$$\log((1-\varepsilon/2b)(1-Q_{i-1})) \leq \mathfrak{f}_{i-1} \leq \log((1+\varepsilon/2b)(1-Q_{i-1}))$$

If we satisfy these constraints, it implies that

$$q_i - \varepsilon/b \leq f_i \leq q_i + \varepsilon/b.$$

Since $q_i \geq 1/b$, this implies

$$(1-\varepsilon)q_i \leq f_i \leq (1+\varepsilon)q_i$$

as desired.

Now, the remaining elements each have $\leq 1/b$ mass. For each such element q_i, we create a constraint

$$(1-O(\varepsilon))\frac{q_i}{1-Q_{i-1}} \leq \mathfrak{f}_{i-1} - \mathfrak{f}_i \leq (1+O(\varepsilon))\frac{q_i}{1-Q_{i-1}}$$

Note that the middle term is

$$-\log\left(\frac{1-F_i}{1-F_{i-1}}\right) = -\log\left(1-\frac{f_i}{1-F_{i-1}}\right) \in \frac{f_i}{1-F_{i-1}}\left(1 \pm 2\varepsilon\right),$$

where the second equality uses the Taylor expansion and the facts that $f_i \leq 1/b$ and $1 - F_{i-1} \geq \varepsilon$ (since during the previous part, we ignored the rightmost $O(\varepsilon)$ probability mass). If we satisfy the desired constraints, it implies that

$$f_i \in \frac{1}{(1 \pm 2\varepsilon)}\frac{1-F_{i-1}}{1-Q_{i-1}}(1 \mp O(\varepsilon))q_i.$$

Since we are taking $\Omega(1/\varepsilon^4)$ samples and $1 - F_{i-1} \geq \Omega(\varepsilon)$, Lemma 1 implies that f_i is indeed a multiplicative $(1 \pm \varepsilon)$ approximation for these points as well.

We note that all points which do not fall into these two cases make up a total of $O(\varepsilon)$ probability mass. Therefore, f may be arbitrary at these points and only incur $O(\varepsilon)$ cost in

total variation distance.

If we find a feasible point for this linear program, it implies the existence of an MHR distribution within $O(\varepsilon)$ total variation distance. In this case, we continue to the testing portion of the algorithm. Furthermore, if $p \in \mathcal{MHR}_n$, our method for generating q certifies that such a distribution exists, and we continue on to the testing portion of the algorithm. □

3.10 Lower Bounds for Testing Classes

We now prove sharp lower bounds for the classes of distributions we consider. We show that the example studied by Paninski [Pan08] to prove lower bounds on testing uniformity can be used to prove lower bounds for the classes we consider. They consider a class $\mathcal{Q}$ consisting of $2^{n/2}$ distributions defined as follows. Without loss of generality assume that n is even. For each of the $2^{n/2}$ vectors $z_0 z_1 \dots z_{n/2-1} \in \{-1, 1\}^{n/2}$, define a distribution $q \in \mathcal{Q}$ over $[n]$ as follows.

$$q_i = \begin{cases} \frac{(1+z_\ell c\varepsilon)}{n} & \text{for } i = 2\ell + 1 \\ \frac{(1-z_\ell c\varepsilon)}{n} & \text{for } i = 2\ell. \end{cases} \tag{3.13}$$

Each distribution in $\mathcal{Q}$ has a total variation distance $c\varepsilon/2$ from $\mathcal{U}_n$, the uniform distribution over $[n]$. By choosing c to be an appropriate constant, Paninski [Pan08] showed that a distribution picked uniformly at random from $\mathcal{Q}$ cannot be distinguished from $\mathcal{U}_n$ with fewer than $\sqrt{n}/\varepsilon^2$ samples with probability at least $2/3$.

Suppose $\mathcal{C}$ is a class of distributions such that

- The uniform distribution $\mathcal{U}_n$ is in $\mathcal{C}$,
- For appropriately chosen c, $d_{\mathrm{TV}}(\mathcal{C}, \mathcal{Q}) \geq \varepsilon$,

then testing $\mathcal{C}$ is not easier than distinguishing $\mathcal{U}_n$ from $\mathcal{Q}$. Invoking [Pan08] immediately implies that testing the class $\mathcal{C}$ requires $\Omega(\sqrt{n}/\varepsilon^2)$ samples.

The lower bounds for all the one dimensional distributions will follow directly from this construction, and for testing monotonicity and independence in higher dimensions, we extend

this construction to $d \geq 1$, appropriately. These arguments are proved in the following subsections, leading to lower bounds for testing these classes:

Theorem 20.

- *For any $d \geq 1$, any algorithm for testing monotonicity over $[n]^d$ requires $\Omega(n^{d/2}/\varepsilon^2)$ samples.*
- *For $d \geq 1$, any algorithm for testing independence over $[n_1] \times \cdots \times [n_d]$ requires $\Omega\left(\frac{(n_1 \cdot n_2 \ldots \cdot n_d)^{1/2}}{\varepsilon^2}\right)$ samples.*
- *Any algorithm for testing unimodality, log-concavity, or monotone hazard rate over $[n]$ requires $\Omega(\sqrt{n}/\varepsilon^2)$ samples.*

3.10.1 Monotone Distributions

We first consider $d = 1$ and prove that for appropriately chosen c, any monotone distribution over $[n]$ is ε-far from all distributions in $\mathcal{Q}$. Consider any $q \in \mathcal{Q}$. For this distribution, we say that $i \in [n]$ is a *raise-point* if $q_i < q_{i+1}$. Let R_q be the set of raise points of q. For $q \in \mathcal{Q}$, (3.13) implies at least one in every four consecutive integers in $[n]$ is a raise point, and therefore, $|R_q| \geq n/4$. Moreover, note that if i is a raise-point, then $i+1$ is not a raise point. For any monotone (decreasing) distribution p, $p_i \geq p_{i+1}$. For any raise-point $i \in R_q$, by the triangle inequality,

$$|p_i - q_i| + |p_{i+1} - q_{i+1}| \geq |p_i - p_{i+1} + q_{i+1} - q_i| \geq q_{i+1} - q_i = \frac{2c\varepsilon}{n}. \tag{3.14}$$

Summing over the set R_q, we obtain $d_{\mathrm{TV}}(p, q) \geq \frac{1}{2}|R_q| \cdot \frac{2c\varepsilon}{n} \geq c\varepsilon/4$. Therefore, if $c \geq 4$, then $d_{\mathrm{TV}}(\mathcal{M}_n, q) \geq \varepsilon$. This proves the lower bound for $d = 1$.

This argument can be extended to $[n]^d$. Consider the following class of distributions on $[n]^d$. For each point $\mathbf{i} = (i_1, \ldots, i_d) \in [n]^d$, where i_1 is even, generate a random $z \in \{-1, 1\}$, and assign to $\mathbf{i}$ a probability of $(1 + zc\varepsilon)/n^d$. Let $\mathbf{e}_1 \triangleq (1, 0, \ldots, 0)$. Similar to $d = 1$, assign a probability $(1 - zc\varepsilon)/n^d$ to the point $\mathbf{i} + \mathbf{e}_1 = (i_1 + 1, i_2, \ldots, i_d)$. This class consists of $2^{\frac{n^{d/2}}{2}}$ distributions, and Paninski's arguments extend to give a lower bound of $\Omega(n^{d/2}/\varepsilon^2)$ samples

to distinguish this class from the uniform distribution over $[n]^d$. It remains to show that all these distributions are ε far from $\mathcal{M}_n^d$. Call a point $\mathbf{i}$ as a raise point if $p_{\mathbf{i}} < p_{\mathbf{i}+\mathbf{e}_1}$. For any $\mathbf{i}$, one of the points $\mathbf{i}$, $\mathbf{i}+\mathbf{e}_1$, $\mathbf{i}+2\mathbf{e}_1$, $\mathbf{i}+3\mathbf{e}_1$ is a raise point, and the number of raise points is at least $n^d/4$. Invoking the triangle inequality (identical to (3.14)) over the raise-points, in the first dimension shows that any monotone distribution over $[n]^d$ is at a distance $\frac{c\varepsilon}{4}$ from any distribution in this class. Choosing $c = 4$ yields a bound of ε.

3.10.2 Product Distributions

Our idea for testing independence is similar to the previous section. We sketch the construction of a class of distributions on $\mathcal{X} = [n_1] \times \cdots \times [n_d]$. Then $|\mathcal{X}| = n_1 \cdot n_2 \ldots \cdot n_d$. For each element in $\mathcal{X}$ assign a value $(1 \pm c\varepsilon)$ and then for each such assignment, normalize the values so that they add to 1, giving rise to a distribution. This gives us a class of $2^{|\mathcal{X}|}$ distributions. The key argument is to show that a *large* fraction of these distributions are far from being a product distribution. This follows since the degrees of freedom of a product distribution is exponentially smaller than the number of possible distributions. The second step is to simply apply Paninski's argument, now over the larger set of distributions, where we show that distinguishing the collection of distributions we constructed from the uniform distribution over $\mathcal{X}$ (which is a product distribution) requires $\sqrt{|\mathcal{X}|}/\varepsilon^2$ samples.

3.10.3 Log-concave and Unimodal Distributions

We will show that any log-concave or unimodal distribution is ε-far from all distributions in $\mathcal{Q}$. Since $\mathcal{LCD}_n \subset \mathcal{U}_n$, it will suffice to show this for every unimodal distribution. Consider any unimodal distribution p, with mode ℓ. Then, p is monotone non-decreasing over the interval $[\ell]$ and non-increasing over $\{\ell+1, \ldots, n\}$. By the argument for monotone distributions, the total variation distance between p and any distribution q over elements greater than ℓ is at least $\frac{n-\ell-1}{n}\frac{c\varepsilon}{4}$, and over elements less than ℓ is at least $\frac{\ell-1}{n}\frac{c\varepsilon}{4}$. Summing these two gives the desired bound.

3.10.4 Monotone Hazard Distributions

We will show that any monotone hazard rate distribution is ε-far from all distributions in $\mathcal{Q}$.

Let p be any monotone-hazard distribution. Any distribution $q \in \mathcal{Q}$ has mass at least $1/2$ over the interval $I = [n/4, 3n/4]$. Therefore, by Lemma 21, for any $i \in I$, $p_{i+1}\left(1 + \frac{p_i}{1/4}\right) \geq p_i$. As noted before, at least $n/8$ of the raise-points are in I.

For any $i \in I \cap R_q$, $q_i = (1 + c\varepsilon)/n$, $q_{i+1} = (1 - c\varepsilon)/n$

$$d_i = |p_i - q_i| + |p_{i+1} - q_{i+1}|. \tag{3.15}$$

If $p_i \geq (1 + 2c\varepsilon)/n$ or $p_i \leq 1/n$, then the first term, and therefore d_i is at least $c\varepsilon/n$. If $p_i \in (1/n, (1 + 2c\varepsilon)/n)$, then for $n > 5/(c\varepsilon)$

$$p_{i+1} \geq \frac{1}{n} \cdot \frac{1}{1 + \frac{4}{n}} \geq \frac{1 - c\varepsilon/2}{n}.$$

Therefore the second term of d_i is at $c\varepsilon/2n$. Since there are at least $n/8$ raise points in I,

$$d_{\mathrm{TV}}(p, q) \geq \frac{1}{2}\frac{n}{8} \cdot \frac{c\varepsilon}{2n} \geq \frac{c\varepsilon}{16}. \tag{3.16}$$

Thus any MHR distribution is ε-far from $\mathcal{Q}$ for $c \geq 16$.

Chapter 4

Testing High-Dimensional Ising Models

4.1 Introduction

Data analysis has become more prevalent in settings with multivariate data, which brings with it a host of new challenges. In particular, the dimensionality of modern datasets is much greater than what has been faced classically. We outline a few situation in which multivariate data arises naturally:

- In natural language processing and information retrieval, the bag-of-words model maps a text document to a vector, where each dimension corresponds to a word, and the projection of the vector in a dimension is the multiplicity of that word in the document. The dimensionality of the data corresponds to the number of unique words, which, in English, can correspond to hundreds of thousands of dimensions.

- Recently, online services like Netflix have turned to statistical data analysis to understand user behavior and optimize the user experience. For instance, recommendation engines use machine learning to determine what users will like, based on what they have liked in the past. Netflix takes this a step further, as it uses user data to decide which shows to green-light: famously, it chose to produce the hit show House of Cards based on the fact that there was a large number of users who simultaneously enjoyed the British version of "House of Cards," films featuring Kevin Spacey, and films directed by David Fincher. However, the dimensionality of the data is prohibitive for

the application of naive methods, as Netflix has tens of thousands of titles. Similar challenges are faced by online retailers like Amazon, which sells hundreds of millions of products.

- Data analysis is frequently performed on large-scale social networks, in order to understand the complex structure of user interactions and the spread of information. In this setting, the behavior of each user can be considered as a separate dimension. The challenge lies in the scale of the data: modern social networks can be incredibly large, with Facebook having over two billion active users.

Given the ubiquity of multivariate data, it is natural to ask whether distribution testing can be efficiently performed in high-dimensional settings. Unfortunately, in general, the answer is no: in an n-dimensional setting, the cost of most distribution testing problems necessarily scales exponentially with n.[1] One easy way to see this is to consider testing uniformity over the domain $[r]^n$. As shown by Paninski [Pan08] (Theorem 2), testing uniformity over a domain Σ requires $\Omega\left(\frac{\sqrt{|\Sigma|}}{\varepsilon^2}\right)$ samples. In this setting, this result implies a lower bound of $\Omega(r^{n/2})$. This intractability is not specific to testing uniformity; as we showed in Theorem 20, testing for monotonicity or independence also requires a number of samples which is exponential in the dimension. Very roughly speaking, this exponential dependence on the dimension arises because the size of the support also increases exponentially – this results in more space in which an adversary can pack exponentially many distributions which are difficult to distinguish. As a consequence, even in moderate dimensions, the cost of most standard distribution testing algorithms will be prohibitively expensive.

Given the importance of performing data analysis in high-dimensional settings, it is natural to ask whether we can develop tools which circumvent these statistical lower bounds. A common way of doing so is by assuming the underlying distribution possesses some additional *structure*. This can be seen as a way of going *beyond worst-case analysis*: phrased differently, assuming some underlying structure *limits* the type of distributions an adversary could ask an algorithm to distinguish. Stated yet another way, this assumption attempts to

[1] In this chapter alone, we use n to refer to the *dimension* of the data, rather than the size of the support. For instance, on the n-dimensional hypercube, the size of the support is 2^n, rather than n.

capture the notion that nature is not "evil" in the problems which it presents us with. Perhaps the most common type of structural restriction is that the data enjoys some notion of *sparsity*. This direction is far too broad to do justice, but perhaps the prototypical example of exploiting sparsity in data analysis is the Lasso for regression problems [Tib96]. In our work, we focus on a different type of structural restriction: graphical models.

Motivated by the above considerations and the ubiquity of Markov Random Fields (MRFs) in the modeling of high-dimensional distributions (see [Jor10] for the basics of MRFs and the references [STW10, KNS07] for a sample of applications), we initiate the study of distribution testing for the prototypical example of MRFs: *the Ising Model,* which captures all binary MRFs with node and edge potentials.[2] Recall that the Ising model is a distribution over $\{-1,1\}^n$, defined in terms of a graph $G = (V, E)$ with n nodes. It is parameterized by a scalar parameter $\theta_{u,v}$ for every edge $(u, v) \in E$, and a scalar parameter θ_v for every node $v \in V$, in terms of which it samples a vector $x \in \{\pm 1\}^V$ with probability:

$$p(x) = \exp\left(\sum_{v \in V} \theta_v x_v + \sum_{(u,v) \in E} \theta_{u,v} x_u x_v - \Phi(\vec{\theta})\right), \tag{4.1}$$

where $\vec{\theta}$ is the parameter vector and $\Phi(\vec{\theta})$ is the log-partition function, ensuring that the distribution is normalized. Intuitively, there is a random variable X_v sitting on every node of G, which may be in one of two states, or spins: up $(+1)$ or down (-1). The scalar parameter θ_v models a local (or "external") field at node v. The sign of θ_v represents whether this local field favors X_v taking the value $+1$, i.e. the up spin, when $\theta_v > 0$, or the value -1, i.e. the down spin, when $\theta_v < 0$, and its magnitude represents the strength of the local field. We will say a model is "without external field" when $\theta_v = 0$ for all $v \in V$. Similarly, $\theta_{u,v}$ represents the direct interaction between nodes u and v. Its sign represents whether it favors equal spins, when $\theta_{u,v} > 0$, or opposite spins, when $\theta_{u,v} < 0$, and its magnitude corresponds to the strength of the direct interaction. Of course, depending on the structure of the Ising model and the edge parameters, there may be indirect interactions between nodes, which may overwhelm local fields or direct interactions.

[2] This follows trivially by the definition of MRFs, and elementary Fourier analysis of Boolean functions.

The Ising model has a rich history, starting with its introduction by statistical physicists as a probabilistic model to study phase transitions in spin systems [Isi25]. Since then it has found a myriad of applications in diverse research disciplines, including probability theory, Markov chain Monte Carlo, computer vision, theoretical computer science, social network analysis, game theory, and computational biology [LPW09, Cha05, Fel04, DMR11, GG86, Ell93, MS10]. The ubiquity of these applications motivate the problem of inferring Ising models from samples, or inferring statistical properties of Ising models from samples. This type of problem has enjoyed much study in statistics, machine learning, and information theory, see, i.e., [CL68, AKN06, CT06b, RWL10, JJR11, SW12, BGS14, Bre15, VMLC16, BK16, Bha16, BM16, MdCCU16, HKM17, KM17]. Much of prior work has focused on *parameter learning*, where the goal is to determine the parameters of an Ising model to which sample access is given. In contrast to this type of work, which focuses on discerning *parametrically* distant Ising models, our goal is to discern *statistically* distant Ising models, in the hopes of dramatic improvements in the sample complexity. (We will come to a detailed comparison between the two inference goals shortly, after we have stated our results.) To be precise, we study the following problems:

Ising Model Goodness-of-fit (or Identity) Testing: Given sample access to an unknown Ising model p (with unknown parameters over an unknown graph) and a parameter $\varepsilon > 0$, the goal is to distinguish with probability at least $2/3$ between $p = q$ and $d_{\mathrm{SKL}}(p, q) > \varepsilon$, for some specific Ising model q.

Ising Model Independence Testing: Given sample access to an unknown Ising model p (with unknown parameters over an unknown graph) and a parameter $\varepsilon > 0$, the goal is to distinguish with probability at least $2/3$ between $p \in \mathcal{I}$ and $d_{\mathrm{SKL}}(p, \mathcal{I}) > \varepsilon$, where $\mathcal{I}$ are all product distributions over $\{\pm 1\}^n$.

We note that there are several potential notions of statistical distance one could consider — classically, total variation distance and the Kullback-Leibler (KL) divergence have seen the most study. As our focus here is on upper bounds, we consider the symmetrized KL divergence d_{SKL}, which is a "harder" notion of distance than both: in particular, testers for d_{SKL} immediately imply testers for both total variation distance and the KL divergence (cf. Proposition 1). Moreover, by virtue of the fact that d_{SKL} upper-bounds KL in both

directions, our tests offer useful information-theoretic interpretations of rejecting a model q, such as data differencing and large deviation bounds in both directions.

Sample Applications: As an instantiation of our proposed testing problems for the Ising model one may maintain the study of strategic behavior on a social network. To offer a little bit of background, a body of work in economics has modeled strategic behavior on a social network as the evolution of the Glauber dynamics of an Ising model, whose graph is the social network, and whose parameters are related to the payoffs of the nodes under different selections of actions by them and their neighbors. For example, [Ell93, MS10] employ this model to study the adoption of competing technologies with network effects, e.g. iPhone versus Android phones. Glauber dynamics, as described in Section 4.2, define the canonical Markov chain for sampling an Ising model. Hence an observation of the actions (e.g. technologies) used by the nodes of the social network should offer us a sample from the corresponding Ising model (at least if the Glauber dynamics have mixed). An analyst may not know the underlying social network or may know the social network but not the parameters of the underlying Ising model. In either case, how many independent observations would he need to test, e.g., whether the nodes are adopting technologies independently, or whether their adoptions conform to some conjectured parameters? Our results offer algorithms for testing such hypotheses in this stylized model of strategic behavior on a network.

As another application, we turn to the field of computer vision. In the Bayesian setting, it is assumed that images are generated according to some prior distribution. Often, practitioners take this prior to be an Ising model in the binary case, or, in general, a higher-order MRF [GG86]. As such, a dataset of images can be pictured as random samples from this prior. A natural question to ask is, given some distribution, does a set of images conform to this prior? This problem corresponds to goodness-of-fit testing for Ising models.

A third application comes up in the field of medicine and computational biology. In order to improve diagnosis, symptom prediction and classification, as well as to improve overall healthcare outcomes, graphical models are trained on data, often using heuristic methods [FLNP00] and surgeon intuition, thereby incorporating hard-wired expert knowledge; see, i.e., the pneumonia graphical model identified in [LAFH01]. Our methods give

efficient algorithms for testing the accuracy of such models. Furthermore, when the discrepancy is large, we expect that our algorithms could reveal the structural reasons for the discrepancy, i.e., blaming a large portion of the error on a misspecified edge.

Main Results and Techniques: Our main result is the following:

Theorem 21. *Both Ising Model Goodness-of-fit Testing and Ising Model Independence Testing can be solved from* $\text{poly}\left(n, \frac{1}{\varepsilon}\right)$ *samples in polynomial time.*

There are several variants of our testing problems, resulting from different knowledge that the analyst may have about the structure of the graph (connectivity, density), the nature of the interactions (attracting, repulsing, or mixed), as well as the temperature (low vs high). We proceed to discuss all these variants, instantiating the resulting polynomial sample complexity in the above theorem. We also illuminate the techniques involved to prove these theorems. This discussion should suffice in evaluating the merits of the results and techniques of this work.

A. Our Baseline Result. In the least favorable regime, i.e. when the analyst is oblivious to the structure of the Ising model p, the signs of the interactions, and their strength, the polynomial in Theorem 21 becomes $O\left(\frac{n^4\beta^2+n^2h^2}{\varepsilon^2}\right)$. In this expression, $\beta = \max\{|\theta^p_{u,v}|\}$ for independence testing, and $\beta = \max\{\max\{|\theta^p_{u,v}|\}, \max\{|\theta^q_{u,v}|\}\}$ for goodness-of-fit testing, while $h = 0$ for independence testing, and $h = \max\{\max\{|\theta^p_u|\}, \max\{|\theta^q_u|\}\}$ for goodness-of-fit testing; see Theorem 22. If the analyst has an upper bound on the maximum degree $\delta_{\max}$ (of all Ising models involved in the problem) the dependence improves to $O\left(\frac{n^2\delta^2_{\max}\beta^2+n\delta_{\max}h^2}{\varepsilon^2}\right)$, while if the analyst has an upper bound on the total number of edges m, then $\max\{m, n\}$ takes the role of $n\delta_{\max}$ in the previous bound; see Theorem 22.

Technical Discussion 1.0: "Testing via Localization." All the bounds mentioned so far are obtained via a simple localization argument showing that, whenever two Ising models p and q satisfy $d_{\text{SKL}}(p, q) > \varepsilon$, then "we can blame it on a node or an edge;" i.e. there exists a node with significantly different bias under p and q or a pair of nodes u, v whose covariance is significantly different under the two models. Pairwise correlation tests are a simple screening that is often employed in practice. For our setting, there is a straighforward and elegant way to show that pair-wise (and not higher-order) correlation tests suffice; see

Lemma 24.

For more details about our baseline localization tester see Section 4.3.

B. Anchoring Our Expectations. Our next results aim at improving the afore-described baseline bound. Before stating these improvements, however, it is worth comparing the sample complexity of our baseline results to the sample complexity of learning. Indeed, one might expect and it is often the case that testing problems can be solved in a two-step fashion, by first learning a hypothesis $\hat{p}$ that is statistically close to the true p and then using the learned hypothesis $\hat{p}$ as a proxy for p to determine whether it is close to or far from some q, or some set of distributions. Given that the KL divergence and its symmetrized version do not satisfy the triangle inequality, however, it is not clear how such an approach would work. Even if it could, the only algorithm that we are aware of for proper learning Ising models, which offers KL divergence guarantees but does not scale exponentially with the maximum degree and β, is a straightforward net-based algorithm. This algorithm, explained in Section 4.11, requires $\Omega\left(\frac{n^6\beta^2+n^4h^2}{\varepsilon^2}\right)$ samples and is time inefficient. In particular, our baseline algorithm already beats this sample complexity and is also time-efficient. Alternatively, one could aim to parameter-learn p; see, e.g., [VMLC16, KM17] and their references. However, these algorithms require sample complexity that is exponential in the maximum degree, and they typically use samples exponential in β as well. For instance, if we use [VMLC16], which is one of the state-of-the-art algorithms, to do parameter learning prior to testing, we would need $\tilde{O}(\frac{n^4.2^{\beta\cdot d_{\max}}}{\varepsilon^2})$ samples to learn p's parameters closely enough to be able to do the testing afterwards. Our baseline result beats this sample complexity, dramatically so if the degrees are unbounded.

The problem of learning the structure of Ising models (i.e., determining which edges are present in the graph) has enjoyed much study, especially in information theory – see [SW12, Bre15, VMLC16, HKM17, KM17] for some recent results. At a first glance, one may hope that these results have implications for testing Ising models. However, thematic similarities aside, the two problems are qualitatively very different – our problem focuses on statistical estimation, while theirs looks at structural estimation. To point out some qualitative differences for these two problems, the complexity of structure learning is exponential in the maximum degree and β, while only logarithmic in n. On the other hand, for testing Ising

models, the complexity has a polynomial dependence in all three parameters, which is both necessary and sufficient.

C. Trees and Ferromagnets. When p is a tree-structured (or forest-structured) Ising model, then independence testing can be performed computationally efficiently without any dependence on β, with an additional quadratic improvement with respect to the other parameters. In particular, without external fields, i.e. $\max\{|\theta^p_u|\} = 0$, independence can be solved from $O(\frac{n}{\varepsilon})$ samples, and this result is tight when $m = O(n)$; see Theorem 23 for an upper bound and Theorem 40 for a lower bound. Interestingly, we show the dependence on β cannot be avoided in the presence of external fields, or if we switch to the problem of identity testing; see Theorem 41. In the latter case, we can at least maintain the linear dependence on n; see Theorem 24. Similar results hold when p is a ferromagnet, i.e. $\theta^p_{u,v} \geq 0$, with no external fields, even if it is not a tree. In particular, the sample complexity becomes $O(\frac{\max\{m,n\}}{\varepsilon})$ (which is again tight when $m = O(n)$), see Theorem 25.

Technical Discussion 2.0: "Testing via Strong Localization." The improvements that we have just discussed are obtained via the same localization approach discussed earlier, which resulted into our baseline tester. That is, we are still going to "blame it on a node or an edge." The removal of the β dependence and the improved running times are due to the proof of a structural lemma, which relates the parameter $\theta_{u,v}$ on some edge (u, v) of the Ising model to the $\mathbf{E}[X_u X_v]$. We show that for forest-structured Ising models with no external fields, $\mathbf{E}[X_u X_v] = \tanh(\theta_{u,v})$, see Lemma 28. A similar statement holds for ferromagnets with no external field, i.e., $\mathbf{E}[X_u X_v] \geq \tanh(\theta_{u,v})$, see Lemma 31. The proof of the structural lemma for trees/forests is straightforward. Intuitively, the only source of correlation between the endpoints u and v of some edge (u, v) of the Ising model is the edge itself, as besides this edge there are no other paths between u and v that would provide alternative avenues for correlation. Significant more work is needed to prove the inequality for ferromagnets on arbitrary graphs. Now, there may be several paths between u and v besides the edge connecting them. Of course, because the model is a ferromagnet, these paths should intuitively only contribute to increase $\mathbf{E}[X_u X_v]$ beyond $\tanh(\theta_{u,v})$. But making this formal is not easy, as calculations involving the Ising model quickly become unwieldy

beyond trees.[3] Our argument uses a coupling between (an appropriate generalization of) the Fortuin-Kasteleyn random cluster model and the Ising model. The coupling provides an alternative way to sample the Ising model by first sampling a random clustering of the nodes, and then assigning uniformly random spins to the sampled clusters. Moreover, it turns out that the probability that two nodes u and v land in the same cluster increases as the vector of parameters $\vec{\theta}$ of the Ising model increases. Hence, we can work inductively. If only edge (u, v) were present, then $\mathbf{E}[X_u X_v] = \tanh(\theta_{u,v})$. As we start adding edges, the probability that u, v land in the same cluster increases, hence the probability that they receive the same spin increases, and therefore $\mathbf{E}[X_u X_v]$ increases.

A slightly more detailed discussion of the structural result for ferromagnets is in Section 4.1.1.1, and full details about our testers for trees and ferromagnets can be found in Sections 4.4.1 and 4.4.2, respectively.

D. Dobrushin's Uniqueness Condition and the High-Temperature Regime. Motivated by phenomena in the physical world, the study of Ising models has identified phase transitions in the behavior of the model as its parameters vary. A common transition occurs as the temperature of the model changes from low to high. As the parameters $\vec{\theta}$ correspond to inverse (individualistic) temperatures, this corresponds to a transition of these parameters from low values (high temperature) to high values (low temperature). Often the transition to high temperature is identified with the satisfaction of Dobrushin-type conditions [Geo11]. Under such conditions, the model enjoys a number of good properties, including rapid mixing of the Glauber dynamics, spatial mixing properties, and uniqueness of measure. The Ising model has been studied extensively in such high-temperature regimes [Dob56, Cha05, Hay06, DGJ08], and it is a regime that is often used in practice.

In the high-temperature regime, we show that we can improve our baseline result without making ferromagnetic or tree-structure assumptions, using a non-localization based argument, explained next. In particular, we show in Theorem 27 that under high temperature and with no external fields independence testing can be done computationally efficiently from $\tilde{O}\left(\frac{n^{10/3}}{\varepsilon^2 \delta_{\max}^2}\right)$ samples, which improves upon our baseline result if $\delta_{\max}$ is large enough. For instance, when $\delta_{\max} = \Omega(n)$, the sample complexity becomes $\tilde{O}\left(\frac{n^{4/3}}{\varepsilon^2}\right)$. Other tradeoffs

[3] We note that the partition function is #P-hard to compute[JS93].

between β, $\delta_{\max}$ and the sample complexity are explored in Theorem 26. Similar improvements hold when external fields are present (Theorem 29), as well as for identity testing, without and with external fields (Theorems 30 and 31).

We offer some intuition about the improvements in Figures 4-1 and 4-2 (appearing in Section 4.6), which are plotted for high temperature and no external fields. In Figure 4-1, we plot the number of samples required for testing Ising models with no external fields when $\beta = \Theta(\frac{1}{d_{\max}})$ as $d_{\max}$ varies. The horizontal axis is $\log_n \delta_{\max}$. We see that localization is the better algorithm for degrees smaller than $O(n^{2/3})$, above which its complexity can be improved. In particular, the sample complexity is $O(n^2/\varepsilon^2)$ until degree $\delta_{\max} = O(n^{2/3})$, beyond which it drops inverse quadratically in $\delta_{\max}$. In Figure 4-2, we consider a different tradeoff. We plot the number of samples required when $\beta = n^{-\alpha}$ and the degree of the graph varies. In particular, we see three regimes as a function of whether the Ising model is in high temperature ($d_{\max} = O(n^a)$) or low temperature ($d_{\max} = \omega(n^a)$), and also which of our techniques localization vs non-localization gives better sample complexity bounds.

We note that in the special case when the Ising model is both high-temperature *and* ferromagnetic, we can use a similar algorithm to achieve a sample complexity of $\tilde{O}(n/\varepsilon)$ (Theorem 36)[4]. This is nearly-tight for this case, as the lower bound instance of Theorem 40 (which requires $\Omega(n/\varepsilon)$ samples) is both high-temperature and ferromagnetic.

Technical Discussion 3.0: "Testing via a Global Statistic, and Variance Bounds." One way or another all our results up to this point had been obtained via localization, namely blaming the distance of p from independence, or from some distribution q to a node or an edge. Our improved bounds employ non-localized statistics that look at all the nodes of the Ising model simultaneously. Specifically, we employ statistics of the form $Z = \sum_{e=(u,v)\in E} c_e X_u X_v$ for some appropriately chosen signs c_e.

The first challenge we encounter here involves selecting the signs c_e in accordance with the sign of each edge marginal's expectation, $\mathbf{E}[X_u X_v]$. This is crucial to establish that the resulting statistic will be able to discern between the two cases. While the necessary estimates of these signs could be computed independently for each edge, this would incur

[4]We note that prior work of [GLP17] proves a qualitatively similar upper bound to ours, using a χ^2-style statistic. We show that our existing techniques suffice to give a near-optimal sample complexity.

an unnecessary overhead of $O(n^2)$ in the number of samples. Instead we try to learn signs that have a non-trivial agreement with the correct signs, from fewer samples. Despite the $X_u X_v$ terms potentially having nasty correlations with each other, a careful analysis using anti-concentration calculations allows us to sidestep this $O(n^2)$ cost and generate satisfactory estimates with a non-negligible probability, from fewer samples.

The second and more significant challenge involves bounding the variance of a statistic Z of the above form. Since Z's magnitude is at most $O(n^2)$, its variance can trivially be bounded by $O(n^4)$. However, applying this bound in our algorithm gives a vacuous sample complexity. As the X_u's will experience a complex correlation structure, it is not clear how one might arrive at non-trivial bounds for the variance of such statistics, leading to the following natural question:

Question 6. *How can one bound the variance of statistics over high-dimensional distributions?*

This meta-question is at the heart of many high-dimensional statistical tasks, and we believe it is important to develop general-purpose frameworks for such settings. In the context of the Ising model, in fairly general regimes, we can show the variance to be $\tilde{O}(n^2)$. We consider this to be surprising – stated another way, despite the complex correlations which may be present in the Ising model, the summands in Z behave roughly as if they were pairwise independent.

This question has been studied in-depth in three recent works [DDK17, GLP17, GSS18], which prove concentration of measure for d-linear statistics over the Ising model. We note that these results are stronger than what we require in this work – we need only variance bounds (which are implied by concentration of measure) for bilinear statistics. Despite these stronger bounds, for completeness, we present a proof of the variance bounds for bilinear statistics which we require[5]. This approach uses tools from [LPW09]. It requires a bound on the spectral gap of the Markov chain, and an expected Lipschitz property of the statistic when a step is taken at stationarity. The technique is described in Section 4.8, and the variance bounds are given in Theorems 37 and 38.

[5]We thank Yuval Peres for directing us towards the reference [LPW09] and the tools required to prove these bounds.

E. Our Main Lower Bound. The proof of our linear lower bound applies Le Cam's method [LC73]. Our construction is inspired by Paninski's lower bound for uniformity testing [Pan08], which involves pairing up domain elements and jointly perturbing their probabilities. This style of construction is ubiquitous in univariate testing lower bounds. A naive application of this approach would involve choosing a fixed matching of the nodes and randomly perturbing the weight of the edges, which leads to an $\Omega(\sqrt{n})$ lower bound. We analyze a construction of a similar nature as a warm-up for our main lower bound, while also proving a lower bound for uniformity testing on product distributions over a binary alphabet (which are a special case of the Ising model where no edges are present), see Theorem 39. To achieve the linear lower bound, we instead consider a *random* matching of the nodes. The analysis of this case turns out to be much more involved due to the complex structure of the probability function which corresponds to drawing k samples from an Ising model on a randomly chosen matching. Indeed, our proof turns out to have a significantly combinatorial flavor, and we believe that our techniques might be helpful for proving stronger lower bounds in combinatorial settings for multivariate distributions. Our analysis of this construction is tight, as uniformity testing on forests can be achieved with $O(n)$ samples. We believe that a super-linear lower bound would be very interesting, but also quite difficult to obtain. Proving our linear lower bound already required a very careful analysis for a relatively simple construction, and an improved lower bound would require analyzing a distribution over dense constructions, for which an improved structural understanding is needed. A further technical discussion of this lower bound is in Section 4.1.1.2, see Section 4.9 and Theorem 40 for a formal statement and full analysis of our main lower bound. As mentioned before, we also show that the sample complexity must depend on β and h in certain cases, see Theorem 41 for a formal statement.

Table 4.1 summarizes our algorithmic results.

The High-Dimensional Frontier and Related Work: We emphasize that we believe the study of high-dimensional distribution testing to be of significant importance, as real-world applications often involve multivariate data. As univariate distribution testing is now very well understood, with a thorough set of tools and techniques, this is the natural next frontier to attack. However, multivariate distributions pose several new technical challenges,

Testing Problem	No External Field	Arbitrary External Field
Independence using Localization	$\tilde{O}\left(\frac{n^2\delta_{\max}^2\beta^2}{\varepsilon^2}\right)$	$\tilde{O}\left(\frac{n^2\delta_{\max}^2\beta^2}{\varepsilon^2}\right)$
Identity using Localization	$\tilde{O}\left(\frac{n^2\delta_{\max}^2\beta^2}{\varepsilon^2}\right)$	$\tilde{O}\left(\frac{n^2\delta_{\max}^2\beta^2}{\varepsilon^2}+\frac{n^2h^2}{\varepsilon^2}\right)$
Independence under Dobrushin/high-temperature using Learn-Then-Test	$\tilde{O}\left(\frac{n^{10/3}\beta^2}{\varepsilon^2}\right)$	$\tilde{O}\left(\frac{n^{10/3}\beta^2}{\varepsilon^2}\right)$
Identity under Dobrushin/high-temperature using Learn-Then-Test	$\tilde{O}\left(\frac{n^{10/3}\beta^2}{\varepsilon^2}\right)$	$\tilde{O}\left(\frac{n^{11/3}\beta^2}{\varepsilon^2}+\frac{n^{5/3}h^2}{\varepsilon^2}\right)$
Independence on Forests using Improved Localization	$\tilde{O}\left(\frac{n}{\varepsilon}\right)$	$\tilde{O}\left(\frac{n^2\beta^2}{\varepsilon^2}\right)$
Identity on Forests using Improved Localization	$\tilde{O}\left(\frac{n\cdot c(\beta)}{\varepsilon}\right)$	$\tilde{O}\left(\frac{n^2\beta^2}{\varepsilon^2}+\frac{n^2h^2}{\varepsilon^2}\right)$
Independence on Ferromagnets using Improved Localization	$\tilde{O}\left(\frac{n\delta_{\max}}{\varepsilon}\right)$	$\tilde{O}\left(\frac{n^2\delta_{\max}^2\beta^2}{\varepsilon^2}\right)$
Independence on Ferromagnets under Dobrushin/high-temperature using Globalization	$\tilde{O}\left(\frac{n}{\varepsilon}\right)$	$\tilde{O}\left(\frac{n^{11/3}\beta^2}{\varepsilon^2}+\frac{n^{5/3}h^2}{\varepsilon^2}\right)$

Table 4.1: Summary of our results in terms of the sample complexity upper bounds for the various problems studied. n = number of nodes in the graph, $\delta_{\max}$ = maximum degree, β = maximum absolute value of edge parameters, h = maximum absolute value of node parameters (when applicable), and c is a function discussed in Theorem 24.

and many of these univariate tools are rendered obsolete – as such, we must extend these methods, or introduce new techniques entirely. It is important to develop approaches which may be applicable in much more general high-dimensional distribution testing settings, when there may be complex correlations between random variables. First, it is important to get a grasp on the concentration and variance of statistics in these settings, and we provide exposition of a technique for bounding the variance of some simple statistics. Additionally, our linear lower bound's construction and analysis give insight into which instances cause intractability to arise, and provide a recipe for the style of combinatorics required to analyze

them.

In further works, the authors and other groups have investigated more properties of multilinear functions over the Ising model [DDK17, GLP17, GSS18]. In the present work, we require and prove variance bounds for bilinear functions of the Ising model. These other works prove *concentration* bounds (which are qualitatively stronger than variance bounds) for multilinear functions of arbitrary degree d (rather than just bilinear functions, which are of degree $d = 2$).

High-dimensional distribution testing has recently attracted the interest of the theoretical computer science community, with work concurrent to ours on testing Bayesian networks[6] [CDKS17, DP17]. There is also a more recent work on learning and testing *causal* Bayesian networks with interventions [ABDK18]. One may also consider testing problems in settings involving Markov Chains, of which there has been interest in testing standard properties as well as domain specific ones (i.e., the mixing time) [BFF$^+$01, BV15, DDG18, HKS15, LP16, HKL$^+$17, BK18]. There have also been other recent works on learning and testing Ising models, in both the statistical and structural sense [GNS17, DMR18, BN18]. It remains to be seen which other multivariate distribution classes of interest allow us to bypass the curse of dimensionality.

We note that the paradigm of distribution testing under structural assumptions has been explored in the univariate setting, where we may assume the distribution satisfies some shape restriction: for example, we could assume the distribution is log-concave or k-modal. This often allows exponential savings in the sample complexity [BKR04, DDS$^+$13, DKN15b, DKN15a, DKN17, DKP18]. Note that this testing *with* structure is not to be confused with the problem of testing *for* structure (as was covered in Chapter 3).

4.1.1 Further Technical Discussion and Highlights

In this section, we give a slightly more in-depth discussion of some of the technical highlights of our work. For full details and more discussion, the interested reader can refer to the corresponding sections in the body.

[6] Bayes nets are another type of graphical model, and are in general incomparable to Ising models.

4.1.1.1 Structural Results for Ferromagnetic Ising Models

Our general-purpose testing algorithm is a localization-based algorithm – in particular, it operates based on the structural property that if two Ising models (with no external field) are far from each other, they will have a distant edge marginal. We convert this structural property to an algorithm by estimating each edge marginal and testing whether they match for the two models. However, the underlying structural property is quantitatively weak, and leads to sub-optimal testing bounds. In some cases of interest, we can derive quantitatively stronger versions of this structural result, giving us more efficient algorithms.

For instance, one can consider the ferromagnetic case, where one has all edge parameters $\theta_e \geq 0$. We would like to derive a relationship between an edge marginal (i.e., $\mathbf{E}[X_u X_v]$ for an edge $e = (u, v)$) and the parameter on that edge θ_e. For a tree-structured Ising model with no external field (ferromagnetic or not), it is not hard to show that $\mathbf{E}[X_u X_v] = \tanh(\theta_e)$ – for small edge parameters, this indicates a linear relationship between the edge marginal and the edge parameter. Intuitively, if a model is ferromagnetic and contains cycles, these cycles should only increase the correlation between adjacent nodes, i.e., we would expect that $\mathbf{E}[X_u X_v] \geq \tanh(\theta_e)$. While this is true, it proves surprisingly difficult to prove directly, and we must instead view the Ising model through the Fortuin-Kastelyn random cluster model.

At a high level, the Fortuin-Kastelyn random cluster model is defined for a graph $G = (V, E)$ with a probability parameter $0 < r_e < 1$ on each edge. This parameter indicates the probability of a bond being present on edge e (i.e., the distribution gives a measure over $\{0, 1\}^E$), placing this model into the space of bond percolation models (see Section 4.4.2.1 and (4.20) for the formal definition). It turns out that an alternative way to draw a sample from the Ising model is through this random cluster model. Namely, we first draw a sample from the Fortuin-Kastelyn model (defined with appropriate parameters), and for each connected component in the resulting graph, we flip a fair coin to determine whether all the nodes in the component should be -1 or $+1$.

With this correspondence in hand, we can apply results for the Fortuin-Kastelyn model – crucial for our purposes is that the fact that the FK model's measure is stochastically increasing. Roughly, this means that if we increase the values of the r_e's, the probability of

an edge having a 1 can only increase. Intuitively, this leads to an increase in $\mathbf{E}[X_u X_v]$ in the Ising model, since it increases the probability that the nodes are connected in the FK model, and thus the expectation of any edge can only increase as we increase the ferromagnetic edge parameters. Careful work is needed to carry through the implications of this correspondence, but it allows us to conclude the nearly-optimal sample complexity of $\tilde{O}(m/\varepsilon)$, and under the additional constraint that the Ising model is in the high-temperature regime, $\tilde{O}(n/\varepsilon)$.

Full details are provided in Sections 4.4.2 and 4.7.

4.1.1.2 A Linear Lower Bound for Testing Ising Models

As a starting point for our lower bound, we use Le Cam's classical two-point method. This is the textbook method for proving lower bounds in distribution testing. It involves defining two families of distributions $\mathcal{P}$ and $\mathcal{Q}$, such that every distribution $p \in \mathcal{P}$ is ε-far from every distribution $q \in \mathcal{Q}$. We consider selecting a uniformly random pair $(p, q) \in (\mathcal{P}, \mathcal{Q})$ and then drawing k independent samples from each of p and q. If we can show that the resulting two transcripts of k samples are close in total variation distance, then k samples are insufficient to distinguish these two cases.

While this method is fairly well-understood in the univariate setting, it proves more difficult to apply in some multivariate settings. This difficulty arises in the definition of the set $\mathcal{Q}$[7]. In the univariate setting, we often decompose the domain into several disjoint sets, and define $\mathcal{Q}$ by applying perturbations to each of these sets independently. This style of construction allows us to analyze each subset locally and compose the results. In the multivariate setting, constructions of this local nature are still possible and are not too hard to analyze – see Theorem 39. In this construction, we consider an Ising model defined by taking a fixed perfect matching on the graph and selecting a distribution from $\mathcal{Q}$ by applying a random sign vector to the edge potentials of this matching. This allows us to prove an $\Omega(\sqrt{n})$ lower bound on the complexity of uniformity testing.

However, such local constructions prove to be limited in the multivariate setting. In order to prove stronger lower bounds, we instead must consider an Ising model generated by taking a *random* perfect matching on the graph. This construction is more global in nature,

[7]We note that for simplicity, $\mathcal{P}$ is often chosen to be a singleton.

since the presence of an edge gives us information about the presence of other edges in the graph. As a result, the calculations no longer decompose elegantly over the (known) edges in the matching. While at a first glance, the structure of such a construction may seem too complex to analyze, we reduce it to analyzing the structure of a random pair of matchings by exploiting combinatorial symmetries. An important step in the proof requires us to understand the random variable representing the number of edges shared by two random perfect matchings. This analysis allows us to prove a quadratically-better lower bound of $\Omega(n)$. We believe our analysis may be useful in proving lower bounds for such global constructions in other multivariate settings.

Full details are provided in Section 4.9.

4.1.2 Organization

In Section 4.2, we discuss preliminaries and the notation that we use throughout the chapter. In Section 4.3, we give a simple localization-based algorithm for independence testing and its corresponding variant for goodness-of-fit testing. In Section 4.4, we present improvements to our localization-based algorithms for forest-structured and ferromagnetic Ising models. In Section 4.5, we describe our main algorithm for the high-temperature regime which uses a global statistic on the Ising model. In Section 4.6, we compare our algorithms from Sections 4.3 and 4.5. In Section 4.7, we give a global statistic for testing models which are both ferromagnetic and high-tempearture. In Section 4.8, we discuss the bounds in [LPW09] and apply them to bounding the variance of bilinear statistics over the Ising model. In Section 4.9, we describe our lower bounds.

4.2 Preliminaries

In this chapter, we have a few differences in notation from the rest of the thesis, which we outline here. Rather than considering distributions over the domain $[n]$, we instead consider distributions over the domain $\{\pm 1\}^n$. While before, our goal was to design algorithms which were sublinear in the size of the domain, we now desire algorithms which are *(poly)-logarithmic*: this will correspond to algorithms with sample complexity $\mathrm{poly}(n)$. As the

symbol m is used to refer to the number of edges in a graph, we instead use k for the number of samples used by an algorithm.

Recall the definition of the Ising model from Eq. (4.1). We will abuse notation, referring to both the probability distribution p and the random vector X that it samples in $\{\pm 1\}^V$ as the Ising model. That is, $X \sim p$. We will use X_u to denote the variable corresponding to node u in the Ising model X. When considering multiple samples from an Ising model X, we will use $X^{(l)}$ to denote the l^{th} sample. We will use h to denote the largest node parameter in absolute value and β to denote the largest edge parameter in absolute value. That is, $|\theta_v| \leq h$ for all $v \in V$ and $|\theta_e| \leq \beta$ for all $e \in E$. Depending on the setting, our results will depend on h and β. Furthermore, in this chapter we will use the convention that $E = \{(u, v) \mid u, v \in V, u \neq v\}$ and θ_e may be equal to 0, indicating that edge e is not present in the graph. We use m to denote the number of edges with non-zero parameters in the graph, and $\delta_{\max}$ to denote the maximum degree of a node.

Throughout this chapter, we will use the notation $\mu_v \triangleq \mathbf{E}[X_v]$ for the marginal expectation of a node $v \in V$ (also called node marginal), and similarly $\mu_{uv} \triangleq \mathbf{E}[X_u X_v]$ for the marginal expectation of an edge $e = (u, v) \in E$ (also called edge marginal). In case a context includes multiple Ising models, we will use μ_e^p to refer to the marginal expectation of an edge e under the model p.

We will use $\mathcal{U}_n$ to denote the uniform distribution over $\{\pm 1\}^n$, which also corresponds to the Ising model with $\vec{\theta} = \vec{0}$. Similarly, we use $\mathcal{I}$ for the set of all product distributions over $\{\pm 1\}^n$.

We will consider *Rademacher* random variables, where $Rademacher(p)$ takes value 1 with probability p, and -1 otherwise.

When $\vec{p}$ and $\vec{q}$ are vectors, we will write $\vec{p} \leq \vec{q}$ to mean that $p_i \leq q_i$ for all i.

Definition 10. *In the setting with* no external field*, $\theta_v = 0$ for all $v \in V$.*

Definition 11. *In the* ferromagnetic *setting, $\theta_e \geq 0$ for all $e \in E$.*

Definition 12 (Dobrushin's Uniqueness Condition)**.** *Consider an Ising model p defined on a graph $G = (V, E)$ with $|V| = n$ and parameter vector $\vec{\theta}$. Suppose $\max_{v \in V} \sum_{u \neq v} \tanh(|\theta_{uv}|) \leq 1 - \eta$ for some constant $\eta > 0$. Then p is said to satisfy Dobrushin's uniqueness condition,*

or be in the high temperature regime. Note that since $\tanh(|x|) \leq |x|$ *for all* x*, the above condition follows from more simplified conditions which avoid having to deal with hyperbolic functions. For instance, either of the following two conditions:*

$$\max_{v \in V} \sum_{u \neq v} |\theta_{uv}| \leq 1 - \eta \text{ or}$$

$$\beta \delta_{\max} \leq 1 - \eta$$

are sufficient to imply Dobrushin's condition (where $\beta = \max_{u,v} |\theta_{uv}|$ *and* $\delta_{\max}$ *is the maximum degree of* G*).*

In general, when one refers to the temperature of an Ising model, a high temperature corresponds to small θ_e values, and a low temperature corresponds to large θ_e values. In this work, we will only use the precise definition as given in Definition 12.

Remark 2. *We note that high-temperature is not strictly needed for our results to hold – we only need Hamming contraction of the "greedy coupling." This condition implies rapid mixing of the Glauber dynamics (in* $O(n \log n)$ *steps) via path coupling (Theorem 15.1 of [LPW09]). See [DDK17, GLP17, GSS18] for further discussion of this weaker condition.*

Lipschitz functions of the Ising model have the following variance bound, which is in Chatterjee's thesis [Cha05]:

Lemma 22 (Lipschitz Concentration Lemma)**.** *Suppose that* $f(X_1, \ldots, X_n)$ *is a function of an Ising model in the high-temperature regime. Suppose the Lipschitz constants of* f *are* $l_1, l_2, \ldots, l_n$ *respectively. That is,*

$$|f(X_1, \ldots, X_i, \ldots, X_n) - f(X_1, \ldots, X_i', \ldots, X_n)| \leq l_i$$

for all values of $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ *and for any* X_i *and* X_i'*. Then for some absolute constant* c_1*,*

$$\Pr\left[|f(X) - \mathbf{E}[f(X)]| > t\right] \leq 2 \exp\left(-\frac{c_1 t^2}{2 \sum_{i=1}^n l_i^2}\right).$$

In particular, for some absolute constant c_2,

$$\mathbf{Var}(f(X)) \leq c_2 \sum_i l_i^2.$$

We will use the following folklore result on estimating the parameter of a Rademacher random variable.

Lemma 23. *Given i.i.d. random variables $X_1, \ldots, X_k \sim Rademacher(p)$ for $k = O(\log(1/\delta)/\varepsilon^2)$, there exists an algorithm which obtains an estimate $\hat{p}$ such that $|\hat{p} - p| \leq \varepsilon$ with probability $1 - \delta$.*

In Section 4.8 we use the Glauber dynamics on the Ising model. Glauber dynamics is the canonical Markov chain for sampling from an Ising model. We consider the basic variant known as single-site Glauber dynamics. The dynamics are a Markov chain defined on the set Σ^n where $\Sigma = \{\pm 1\}$. They proceed as follows:

1. Start at any state $X^{(0)} \in \Sigma^n$. Let $X^{(t)}$ denote the state of the dynamics at time t.

2. Let $N(u)$ denote the set of neighbors of node u. Pick a node u uniformly at random and update X as follows:

$$X_u^{(t+1)} = 1 \quad \text{w.p.} \quad \frac{\exp\left(\theta_u + \sum_{v \in N(u)} \theta_{uv} X_v^{(t)}\right)}{\exp\left(\theta_u + \sum_{v \in N(u)} \theta_{uv} X_v^{(t)}\right) + \exp\left(-\theta_u - \sum_{v \in N(u)} \theta_{uv} X_v^{(t)}\right)}$$

$$X_u^{(t+1)} = -1 \quad \text{w.p.} \quad \frac{\exp\left(-\theta_u - \sum_{v \in N(u)} \theta_{uv} X_v^{(t)}\right)}{\exp\left(\theta_u + \sum_{v \in N(u)} \theta_{uv} X_v^{(t)}\right) + \exp\left(-\theta_u - \sum_{v \in N(u)} \theta_{uv} X_v^{(t)}\right)}$$

$$X_v^{(t+1)} = X_v^{(t)} \quad \forall \quad v \neq u.$$

Glauber dynamics define a reversible, ergodic Markov chain whose stationary distribution is identical to the corresponding Ising model. In many relevant settings, such as, for instance, the high-temperature regime, the dynamics are fast mixing, i.e., they mix in time $O(n \log n)$ and hence offer an efficient way to sample from Ising models.

Input to Goodness-of-Fit Testing Algorithms. To solve the goodness-of-fit testing or identity testing problem with respect to a discrete distribution q, a description of q is given as part of the input along with sample access to the distribution p which we are testing. In case q is an Ising model, its support has exponential size and specifying the vector of probability values at each point in its support is inefficient. Since q is characterized by the edge parameters between every pair of nodes and the node parameters associated with the nodes, a succinct description would be to specify the parameters vectors $\{\theta_{uv}\}, \{\theta_u\}$. In many cases, we are also interested in knowing the edge and node marginals of the model. Although these quantities can be computed from the parameter vectors, there is no efficient method known to compute the marginals exactly for general regimes. A common approach is to use MCMC sampling to generate samples from the Ising model. However, for this technique to be efficient we require that the mixing time of the Markov chain be small which is not true in general. Estimating and exact computation of the marginals of an Ising model is a well-studied problem but is not the focus of this work. Hence, to avoid such computational complications we will assume that for the identity testing problem the description of the Ising model q includes both the parameter vectors $\{\theta_{uv}\}, \{\theta_u\}$ as well as the edge and node marginal vectors $\{\mu_{uv} = \mathbf{E}[X_u X_v]\}, \{\mu_u = \mathbf{E}[X_u]\}$.

Symmetric KL Divergence Between Two Ising Models. We note that the symmetric KL divergence between two Ising models p and q admits a very convenient expression [SW12]:

$$d_{\mathrm{SKL}}(p, q) = \sum_{v \in V} (\theta_v^p - \theta_v^q)(\mu_v^p - \mu_v^q) + \sum_{e=(u,v) \in E} (\theta_e^p - \theta_e^q)(\mu_e^p - \mu_e^q). \tag{4.2}$$

This expression will form the basis for all our algorithms.

4.3 Testing via Localization

Our first algorithm is a general purpose "localization" algorithm. While extremely simple, this serves as a proof-of-concept that testing on Ising models can avoid the curse of dimensionality, while simultaneously giving a very efficient algorithm for certain parameter

regimes. The main observation which enables us to do a localization based approach is stated in the following Lemma, which allows us to "blame" a difference between models p and q on a discrepant node or edge.

Lemma 24. *Given two Ising models p and q, if $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$, then either*

- *There exists an edge $e = (u,v)$ such that $(\theta^p_{uv} - \theta^q_{uv})(\mu^p_{uv} - \mu^q_{uv}) \geq \frac{\varepsilon}{2m}$; or*
- *There exists a node u such that $(\theta^p_u - \theta^q_u)(\mu^p_u - \mu^q_u) \geq \frac{\varepsilon}{2n}$.*

Proof of Lemma 24: We have,

$$d_{\mathrm{SKL}}(p,q) = \sum_{e=(u,v)\in E} (\theta^p_e - \theta^q_e)(\mu^p_e - \mu^q_e) + \sum_{v\in V} (\theta^p_v - \theta^q_v)(\mu^p_v - \mu^q_v) \geq \varepsilon$$

$$\implies \sum_{e=(u,v)\in E} (\theta^p_e - \theta^q_e)(\mu^p_e - \mu^q_e) \geq \varepsilon/2 \quad \text{or} \quad \sum_{v\in V} (\theta^p_v - \theta^q_v)(\mu^p_v - \mu^q_v) \geq \varepsilon/2$$

In the first case, there has to exist an edge $e = (u,v)$ such that $(\theta^p_{uv} - \theta^q_{uv})(\mu^p_{uv} - \mu^q_{uv}) \geq \frac{\varepsilon}{2m}$ and in the second case there has to exist a node u such that $(\theta^p_u - \theta^q_u)(\mu^p_u - \mu^q_u) \geq \frac{\varepsilon}{2n}$ thereby proving the lemma. □

Before giving a description of the localization algorithm, we state its guarantees.

Theorem 22. *Given $\tilde{O}\left(\frac{m^2\beta^2}{\varepsilon^2}\right)$ samples from an Ising model p, there exists a polynomial-time algorithm which distinguishes between the cases $p \in \mathcal{I}$ and $d_{\mathrm{SKL}}(p,\mathcal{I}) \geq \varepsilon$ with probability at least $2/3$. Furthermore, given $\tilde{O}\left(\frac{m^2\beta^2}{\varepsilon^2} + \frac{n^2h^2}{\varepsilon^2}\right)$ samples from an Ising model p and a description of an Ising model q, there exists a polynomial-time algorithm which distinguishes between the cases $p = q$ and $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$ with probability at least $2/3$ where $\beta = \max\{|\theta_{uv}|\}$ and $h = \max\{|\theta_u|\}$. The above algorithms assume that m, an upper bound on the number of edges, is known. If no upper bound is known, we may use the trivial upper bound of $\binom{n}{2}$. If we are given as input the maximum degree of nodes in the graph $\delta_{\max}$, m in the above bounds is substituted by $n\delta_{\max}$.*

Note that the sample complexity achieved by the localization algorithm gets worse as the graph becomes denser. This is because as the number of possible edges in the graph grows,

the contribution to the distance by any single edge grows smaller thereby making it harder to detect.

We describe the algorithm for independence testing in Section 4.3.1. The algorithm for testing identity is similar, its description and correctness proofs are given in Section 4.3.2.

4.3.1 Testing Independence via Localization

We start with a high-level description of the algorithm. Given sample access to Ising model $X \sim p$ it will first obtain empirical estimates of the node marginals μ_u for each node $u \in V$ and edge marginals μ_{uv} for each pair of nodes (u, v). Denote these empirical estimates by $\hat{\mu}_u$ and $\hat{\mu}_{uv}$ respectively. Using these empirical estimates, the algorithm computes the empirical estimate for the covariance of each pair of variables in the Ising model. That is, it computes an empirical estimate of $\lambda_{uv} = \mathbf{E}[X_u X_v] - \mathbf{E}[X_u]\mathbf{E}[X_v]$ for all pairs (u, v). If they are all close to zero, then we can conclude that $p \in \mathcal{I}$. If there exists an edge for which λ_{uv} is far from 0, this indicates that p is far from $\mathcal{I}$. The reason for this follows from the expression Lemma 24 and is described in further detail in the proof of Lemma 26. A precise description of the test is given in in Algorithm 2 and its correctness is proven via Lemmas 25 and 26. We note that this algorithm is phrased as if an upper bound on the number of edges m is known. If we instead know an upper bound on the maximum degree $\delta_{\max}$, then we can replace m by $n\delta_{\max}$.

Algorithm 2 Test if an Ising model p is product

1: **function** LocalizationTest(sample access to Ising model p, accuracy parameter ε, β, m)
2: Draw $k = O\left(\frac{m^2\beta^2 \log n}{\varepsilon^2}\right)$ samples from p. Denote the samples by $X^{(1)}, \ldots, X^{(k)}$
3: Compute empirical estimates $\hat{\mu}_u = \frac{1}{k}\sum_i X_u^{(i)}$ for each node $u \in V$ and $\hat{\mu}_{uv} = \frac{1}{k}\sum_i X_u^{(i)} X_v^{(i)}$ for each pair of nodes (u, v)
4: Using the above estimates compute the covariance estimates $\hat{\lambda}_{uv} = \hat{\mu}_{uv} - \hat{\mu}_u\hat{\mu}_v$ for each pair of nodes (u, v)
5: If for any pair of nodes (u, v), $\left|\hat{\lambda}_{uv}\right| \geq \frac{\varepsilon}{4m\beta}$ return that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$
6: Otherwise, return that $p \in \mathcal{I}$
7: **end function**

To prove correctness of Algorithm 2, we will require the following lemma, which allows

us to detect pairs u, v for which λ_{uv} is far from 0.

Lemma 25. *Given $O\left(\frac{\log n}{\varepsilon^2}\right)$ samples from an Ising model $X \sim p$, there exists a polynomial-time algorithm which, with probability at least $9/10$, can identify all pairs of nodes $(u, v) \in V^2$ such that $|\lambda_{uv}| \geq \varepsilon$, where $\lambda_{uv} = \mathbf{E}[X_u X_v] - \mathbf{E}[X_u]\mathbf{E}[X_v]$. Namely, the algorithm computes the empirical value of $|\lambda_{uv}|$ for each pair of nodes and identifies pairs such that this value is sufficiently far from 0.*

Proof. This lemma is a direct consequence of Lemma 23. Note that for any edge $e = (u, v) \in E$, $X_u X_v \sim Rademacher((1 + \mu_e)/2)$. Also $X_u \sim Rademacher((1 + \mu_u)/2)$ and $X_v \sim Rademacher((1 + \mu_v)/2)$. We will use Lemma 23 to show that $O(\log n/\varepsilon^2)$ samples suffice to detect whether $\lambda_e = 0$ or $|\lambda_e| \geq \varepsilon$ with probability at least $1 - 1/10n^2$. With $O(\log n/\varepsilon^2)$ samples, Lemma 23 implies we can obtain estimates $\hat{\mu}_{uv}$, $\hat{\mu}_u$ and $\hat{\mu}_v$ for μ_{uv}, μ_u and μ_v respectively such that $|\hat{\mu}_{uv} - \mu_{uv}| \leq \frac{\varepsilon}{10}$, $|\hat{\mu}_u - \mu_u| \leq \frac{\varepsilon}{10}$ and $|\hat{\mu}_v - \mu_v| \leq \frac{\varepsilon}{10}$ with probability at least $1 - 1/10n^2$. Let $\hat{\lambda}_{uv} = \hat{\mu}_{uv} - \hat{\mu}_u\hat{\mu}_v$. Then from the above, it follows by triangle inequality that $|\lambda_{uv} - \hat{\lambda}_{uv}| \leq \frac{3\varepsilon}{10} + \frac{\varepsilon^2}{100}$. It can be seen that in the case when the latter term in the previous inequality dominates the first, ε is large enough that $O(\log n)$ samples suffice to distinguish the two cases. In the more interesting case, $\frac{\varepsilon^2}{100} \leq \frac{\varepsilon}{10}$, and $|\lambda_{uv} - \hat{\lambda}_{uv}| \leq \frac{4\varepsilon}{10}$. Therefore if $|\lambda_{uv}| \geq \varepsilon$, then $\left|\hat{\lambda}_{uv}\right| \geq \frac{6\varepsilon}{10}$, and if $|\lambda_{uv}| = 0$, then $\left|\hat{\lambda}_{uv}\right| \leq \frac{4\varepsilon}{10}$ thereby implying that with probability at least $1 - 1/10n^2$ we can detect whether $\lambda_{uv} = 0$ or $|\lambda_{uv}| \geq \varepsilon$. Taking a union bound over all edges, the probability that we correctly identify all such edges is at least $9/10$. □

With this lemma in hand, we now prove the first part of Theorem 22.

Lemma 26. *Given $\tilde{O}\left(\frac{m^2\beta^2}{\varepsilon^2}\right)$ samples from an Ising model $X \sim p$, Algorithm 2 distinguishes between the cases $p \in \mathcal{I}$ and $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$ with probability at least $2/3$.*

Proof. We will run Algorithm 2 on all pairs X_u, X_v to identify any pair such that $|\lambda_{uv}|$ is large. This will involve using the algorithm of Lemma 25 with parameter "ε" as $\varepsilon/2\beta m$. If no such pair is identified, output that $p \in \mathcal{I}$, and otherwise, output that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$. If $p \in \mathcal{I}$, we know that $\mathbf{E}[X_u X_v] = \mathbf{E}[X_u]\mathbf{E}[X_v]$ for all edges (u, v), and therefore, with probability $9/10$, there will be no edges for which the empirical estimate of $|\lambda_e| \geq \frac{\varepsilon}{2\beta m}$. On

the other hand, if $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$, then $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ for every $q \in \mathcal{I}$. In particular, consider the product distribution q on n nodes such that $\mu_u^q = \mu_u^p$ for all $u \in V$. For this particular product distribution q, by (4.2), there must exist some e^* such that $|\lambda_{e^*}| \geq \frac{\varepsilon}{2\beta m}$, and the algorithm will identify this edge. This is because

$$\sum_{v \in V} (\theta_v^p - \theta_v^q)(\mu_v^p - \mu_v^q) = 0 \tag{4.3}$$

$$\therefore d_{\mathrm{SKL}}(p, q) \geq \varepsilon$$

$$\implies \exists e^* = (u, v) \text{ s.t } (\theta_e^p - \theta_e^q)(\mu_e^p - \mu_e^q) \geq \frac{\varepsilon}{m} \tag{4.4}$$

$$\implies \exists e^* = (u, v) \text{ s.t } |(\mu_e^p - \mu_e^q)| \geq \frac{\varepsilon}{2\beta m} \tag{4.5}$$

$$\implies \exists e^* = (u, v) \text{ s.t } |\lambda_{e^*}| \geq \frac{\varepsilon}{2\beta m}.$$

where (4.3) follows because $\mu_v^p = \mu_v^q$ for all $v \in V$, (4.4) follows from Lemma 24 and (4.5) follows because $|\theta_e^p - \theta_e^q| \leq 2\beta$. This completes the proof of the first part of Theorem 22. □

4.3.2 Testing Identity via Localization

If one wishes to test for identity of p to an Ising model q, the quantities whose absolute values indicate that p is far from q are $\mu_{uv}^p - \mu_{uv}^q$ for all pairs u, v, and $\mu_u^p - \mu_u^q$ for all u, instead of λ_{uv}. Since μ_{uv}^q and μ_u^q are given as part of the description of q, we only have to identify whether $\mathbf{E}[X_u X_v] \geq c$ and $\mathbf{E}[X_u] \geq c$ for any constant $c \in [-1, 1]$. A variant of Lemma 25 as stated in Lemma 27 achieves this goal. Algorithm 3 describes the localization based identity test. Its correctness proof will imply the second part of Theorem 22 and is similar in vein to that of Algorithm 2. It is omitted here.

Lemma 27. *Given $O\left(\frac{\log n}{\varepsilon^2}\right)$ samples from an Ising model p, there exists a polynomial-time algorithm which, with probability at least $9/10$, can identify all pairs of nodes $(u, v) \in V^2$ such that $|\mu_{uv}^p - c| \geq \varepsilon$ for any constant $c \in [-1, 1]$. There exists a similar algorithm, with sample complexity $O\left(\frac{\log n}{\varepsilon^2}\right)$ which instead identifies all $v \in V$ such that $|\mu_v^p - c| \geq \varepsilon$, where $\mu_v^p = \mathbf{E}[X_v]$ for any constant $c \in [-1, 1]$.*

Proof of Lemma 27: The proof follows along the same lines as Lemma 25. Let $X \sim p$. Then,

for any pair of nodes (u, v), $X_u X_v \sim Rademacher((1+\mu_e^p)/2)$. Also $X_u \sim Rademacher((1+\mu_u^p)/2)$ for any node u. For any pair of nodes u, v, with $O(\log n/\varepsilon^2)$ samples, Lemma 23 implies we that the empirical estimate $\hat{\mu}_{uv}^p$ is such that $|\hat{\mu}_{uv}^p - \mu_{uv}^p| \leq \frac{\varepsilon}{10}$ with probability at least $1 - 1/10n^2$. By triangle inequality, we get $|\mu_{uv}^p - c| - \frac{\varepsilon}{10} \leq |\hat{\mu}_{uv}^p - c| \leq |\mu_{uv}^p - c| + \frac{\varepsilon}{10}$. Therefore if $|\mu_{uv}^p - c| = 0$, then $|\hat{\mu}_{uv}^p - c| \leq \frac{\varepsilon}{10}$ w.p. $\geq 1 - 1/10n^2$ and if $|\mu_{uv}^p - c| \geq \varepsilon$, then $|\hat{\mu}_{uv}^p - c| \geq \frac{9\varepsilon}{10}$ w.p. $\geq 1 - 1/10n^2$. Hence by comparing whether $|\hat{\mu}_{uv}^p - c|$ to $\varepsilon/2$ we can distinguish between the cases $|\mu_{uv}^p - c| = 0$ and $|\mu_{uv}^p - c| \geq \varepsilon$ w.p. $\geq 1 - 1/10n^2$. Taking a union bound over all edges, the probability that we correctly identify all such edges is at least 9/10. The second statement of the Lemma about the nodes follows similarly. □

Algorithm 3 Test if an Ising model p is identical to q

function LocalizationTestIdentity(sample access to Ising model $X \sim p$, description of Ising model q, accuracy parameter ε,β,h,m)
  Draw $k = O\left(\frac{\left(m^2\beta^2 + n^2h^2\right)\log n}{\varepsilon^2}\right)$ samples from p. Denote the samples by $X^{(1)}, \ldots, X^{(k)}$
  Compute empirical estimates $\hat{\mu}_u^p = \frac{1}{k}\sum_i X_u^{(i)}$ for each node $u \in V$ and $\hat{\mu}_{uv}^p = \frac{1}{k}\sum_i X_u^{(i)} X_v^{(i)}$ for each pair of nodes (u, v)
  If for any pair of nodes (u, v), $|\hat{\mu}_{uv}^p - \mu_{uv}^q| \geq \frac{\varepsilon}{8m\beta}$ return that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$
  If for any node u, if $|\hat{\mu}_u^p - \mu_u^q| \geq \frac{\varepsilon}{8nh}$ return that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$
  Otherwise, return that $p = q$
end function

The proof of correctness of Algorithm 3 follows along the same lines as that of Algorithm 2 and uses Lemma 27. We omit the proof here.

4.4 Improved Testing on Forests and Ferromagnets

In this section we will describe testing algorithms for two commonly studied classes of Ising models, namely forests and ferromagnets. In these cases, the sample complexity improves compared to the baseline result when in the regime of no external field. The testers are still localization based (like those of Section 4.3), but we can now leverage structural properties to obtain more efficient testers.

First, we consider the class of all forest structured Ising models, where the underlying graph $G = (V, E)$ is a forest. Such models exhibit nice structural properties which can

be exploited to obtain more efficient tests. In particular, under no external field, the edge marginals μ_e, which, in general are hard to compute, have a simple closed form expression. This structural information enables us to improve our testing algorithms from Section 4.3 on forest graphs. We state the improved sample complexities here and defer a detailed description of the algorithms to Section 4.4.1.

Theorem 23 (Independence testing of Forest-Structured Ising Models)**.** *Algorithm 4 takes in $\tilde{O}\left(\frac{n}{\varepsilon}\right)$ samples from an Ising model $X \sim p$ whose underlying graph is a forest and which is under no external field and outputs whether $p \in \mathcal{I}$ or $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$ with probability $\geq 9/10$.*

Remark 3. *Note that Theorem 23 together with our lower bound described in Theorem 40 indicate a tight sample complexity up to logarithmic factors for independence testing on forest-structured Ising models under no external field.*

Theorem 24 (Identity Testing of Forest-Structured Ising Models)**.** *Algorithm 5 takes in the edge parameters of an Ising model q on a forest graph and under no external field as input, and draws $\tilde{O}\left(c(\beta)\frac{n}{\varepsilon}\right)$ samples from an Ising model $X \sim p$ (where $c(\beta)$ is a function of the parameter β) whose underlying graph is a forest and under no external field, and outputs whether $p = q$ or $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ with probability $\geq 9/10$.*

Note that for identity testing, any algorithm necessarily has to have at least a β dependence due to the lower bound we show in Theorem 41.

The second class of Ising models we consider this section are ferromagnets. For a ferromagnetic Ising model, $\theta_{uv} \geq 0$ for every pair of nodes u, v. Ferromagnets may potentially contain cycles but since all interactions are ferromagnetic, the marginal of every edge is at least what it would have been if it was a solo edge. This intuitive property turns out to be surprisingly difficult to prove in a direct way. We prove this structural property using an alternative view of the Ising model density which comes from the Fortuin-Kasteleyn random cluster model. Using this structural property, we give a quadratic improvement in the dependence on parameter m for testing independence under no external field. We state our main result in this regime here and a full description of the algorithm and the structural lemma are provided in Section 4.4.2.

Theorem 25 (Independence Testing of Ferromagnetic Ising Models)**.** *Algorithm 6 takes in* $\tilde{O}\left(\frac{n\delta_{\max}}{\varepsilon}\right)$ *samples from a ferromagnetic Ising model* $X \sim p$ *which is under no external field and outputs whether* $p \in \mathcal{I}$ *or* $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$ *with probability* $\geq 9/10$*.*

4.4.1 Testing on Forests

Before we present the improved algorithms, we will prove the following fact about the edge marginals of an arbitrary Ising model with no external field where the underlying graph is a forest. This result was known prior to this work by the community but we couldn't find a proof of the same, hence we provide our own proof of the lemma.

Lemma 28 (Structural Lemma for Forest-Structured Ising Models)**.** *If* p *is an Ising model on a forest graph with no external field, and* $X \sim p$*, then for any* $(u, v) \in E$*, we have*

$$\mathbf{E}\left[X_u X_v\right] = \tanh(\theta_{uv}).$$

Proof. Consider any edge $e = (u, v) \in E$. Consider the tree (T, E_T) which contains e. Let n_T be the number of nodes in the tree. We partition the vertex set T into U and V as follows. Remove edge e from the graph and let U denote all the vertices which lie in the connected component of node u except u itself. Similarly, let V denote all the vertices which lie in the connected component of node v except node v itself. Hence, $T = U \cup V \cup \{u\} \cup \{v\}$. Let X_U be the vector random variable which denotes the assignment of values in $\{\pm 1\}^{|U|}$ to the nodes in U. X_V is defined similarly. We will also denote a specific value assignment to a set of nodes S by x_S and $-x_S$ denotes the assignment which corresponds to multiplying each coordinate of x_S by -1. Now we state the following claim which follows from the tree structure of the Ising model.

Claim 2. $\Pr\left[X_U = x_U, X_u = 1, X_v = 1, X_V = x_V\right] = \exp(2\theta_{uv}) \Pr\left[X_U = x_U, X_u = 1, X_v = -1, X_V = -x_V\right]$

In particular the above claim implies the following corollary which is obtained by marginalization of the probability to nodes u and v.

Corollary 11. *If* X *is an Ising model on a forest graph* $G = (V, E)$ *with no external field, then for any edge* $e = (u, v) \in E$*,* $\Pr\left[X_u = 1, X_v = 1\right] = \exp(2\theta_{uv}) \Pr\left[X_u = 1, X_v = -1\right]$.

Now,

$$\mathbf{E}\left[X_u X_v\right] = \Pr\left[X_u X_v = 1\right] - \Pr\left[X_u X_v = -1\right] \tag{4.6}$$

$$= 2Pr\left[X_u = 1, X_v = 1\right] - 2\Pr\left[X_u = 1, X_v = -1\right] \tag{4.7}$$

$$= \frac{2Pr\left[X_u = 1, X_v = 1\right] - 2\Pr\left[X_u = 1, X_v = -1\right]}{2Pr\left[X_u = 1, X_v = 1\right] + 2\Pr\left[X_u = 1, X_v = -1\right]} \tag{4.8}$$

$$= \frac{Pr\left[X_u = 1, X_v = 1\right] - \Pr\left[X_u = 1, X_v = -1\right]}{Pr\left[X_u = 1, X_v = 1\right] + \Pr\left[X_u = 1, X_v = -1\right]} \tag{4.9}$$

$$= \left(\frac{\exp(2\theta_{uv}) - 1}{\exp(2\theta_{uv}) + 1}\right) \frac{\Pr\left[X_u = 1, X_v = -1\right]}{\Pr\left[X_u = 1, X_v = -1\right]} \tag{4.10}$$

$$= \tanh(\theta_{uv}) \tag{4.11}$$

where (4.7) follows because $\Pr\left[X_u = 1, X_v = 1\right] = \Pr\left[X_u = -1, X_v = -1\right]$ and $\Pr\left[X_u = -1, X_v = 1\right] = \Pr\left[X_u = 1, X_v = -1\right]$ by symmetry. Line (4.8) divides the expression by the total probability which is 1 and (4.10) follows from Corollary 11.

□

Given the above structural lemma, we give the following simple algorithm for testing independence on forest Ising models under no external field.

Algorithm 4 Test if a forest Ising model p under no external field is product

1: **function** TESTFORESTISING-PRODUCT(sample access to Ising model p)
2: Run the algorithm of Lemma 25 to identify all edges $e = (u, v)$ such that $|\mathbf{E}[X_u X_v]| \geq \sqrt{\frac{\varepsilon}{n}}$ using $\tilde{O}\left(\frac{n}{\varepsilon}\right)$ samples. If it identifies any edges, return that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$
3: Otherwise, return that p is product.
4: **end function**

Algorithm 4, at a high level, works as follows. If there is an edge parameter whose absolute value is larger than a certain threshold, it will be easy to detect due to the structural information about the edge marginals. In case all edges have parameters smaller in absolute value than this threshold, the expression for $d_{\mathrm{SKL}}(.,.)$ between two Ising models tells us that there still has to be at least one edge with a significantly large value of μ_e in case the model is far from uniform, and hence will still be detectable by the algorithm of Lemma 25. The proof of Theorem 23 shows this formally.

Proof of Theorem 23: Firstly, note that under no external field, the only product Ising model is the uniform distribution $\mathcal{U}_n$. Therefore the problem reduces to testing whether p is uniform or not. Consider the case when p is indeed uniform. That is, there are no edges in the underlying graph of the Ising model. In this case with probability at least 9/10 the localization algorithm of Lemma 25 will output no edges. Hence Algorithm 4 will output that p is uniform.

In case $d_{\mathrm{SKL}}(p,\mathcal{U}_n) \geq \varepsilon$, we split the analysis into two cases.

- *Case 1:* There exists an edge $e = (u, v)$ such that $|\theta_{uv}| \geq \sqrt{\frac{\varepsilon}{n}}$. In this case, $\mathbf{E}[X_u X_v] = \tanh(\theta_{uv})$ and in the regime where $|\theta| = o(1)$, $|\tanh(\theta)| \geq |\theta/2|$. Hence implying that $|\mathbf{E}[X_u X_v]| \geq |\theta_{uv}/2| \geq \left|\sqrt{\frac{\varepsilon}{n}}/2\right|$. Therefore the localization algorithm of Lemma 25 would identify such an edge with probability at least 9/10. Note that the regime where the inequality $|\tanh(\theta)| \geq |\theta/2|$ isn't valid is easily detectable using $\tilde{O}(\frac{n}{\varepsilon})$ samples, as this would imply that $|\theta| \geq 1.9$ and $|\mathbf{E}[X_u X_v]| \geq 0.95$.

- *Case 2:* All edges $e = (u, v)$ are such that $|\theta_{uv}| \leq \left|\sqrt{\frac{\varepsilon}{n}}\right|$. In this case we have,

$$
\begin{aligned}
d_{\mathrm{SKL}}(p,\mathcal{U}_n) &\geq \varepsilon && (4.12)\\
\implies \exists \text{ edge } e = (u, v) \text{ s.t } \theta_{uv}\mathbf{E}[X_u X_v] &\geq \frac{\varepsilon}{n} && (4.13)\\
\implies \exists \text{ edge } e = (u, v) \text{ s.t } |\mathbf{E}[X_u X_v]| &\geq \left|\frac{\varepsilon}{n} \times \sqrt{\frac{n}{\varepsilon}}\right| && (4.14)\\
&= \sqrt{\frac{\varepsilon}{n}} && (4.15)
\end{aligned}
$$

  Hence, the localization algorithm of Lemma 25 would identify such an edge with probability at least 9/10.

□

Next, we will present an algorithm for identity testing on forest Ising models under no external field.

Proof of Theorem 24: Consider the case when p is indeed q. In this case with probability at least 9/10 the localization algorithm of Lemma 25 will output no edges. Hence Algorithm 5

Algorithm 5 Test if a forest Ising model p under no external field is identical to a given Ising model q

function TESTFORESTISING-IDENTITY(Ising model q,sample access to Ising model p)
    If the Ising model q is not a forest, or has a non-zero external field on some node,. return $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$
    Run the algorithm of Lemma 25 to identify all edges $e = (u,v)$ such that. $|\mathbf{E}[X_u X_v] - \tanh(\theta^q_{uv})| \geq \sqrt{\frac{\varepsilon}{n}}$ using $\tilde{O}\left(\frac{n}{\varepsilon}\right)$ samples. If it identifies any edges, return that $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$
    Otherwise, return that $p = q$.
end function

will output that p is uniform.

In case $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$, we split the analysis into two cases.

- *Case 1:* There exists an edge $e = (u,v)$ such that $|\theta^p_{uv} - \theta^q_{uv}| \geq \sqrt{\frac{\varepsilon}{n}}$. In this case, $\mathbf{E}[X_u X_v] - \mu^q_{uv} = \tanh(\theta^p_{uv}) - \tanh(\theta^q_{uv})$ and hence has the same sign as $\theta^p_{uv} - \theta^q_{uv}$. Assume that $\theta^p_{uv} \geq \theta^q_{uv}$. The argument for the case $\theta^q_{uv} > \theta^p_{uv}$ will follow similarly. If $\theta^p_{uv} - \theta^q_{uv} \leq 1/2 \tanh(\beta)$, then the following inequality holds from Taylor's theorem.

$$\tanh(\theta^p_{uv}) - \tanh(\theta^q_{uv}) \geq \frac{\operatorname{sech}^2(\beta)\,(\theta^p_{uv} - \theta^q_{uv})}{2}$$

  which would imply $\tanh(\theta^p_{uv}) - \tanh(\theta^q_{uv}) \geq \frac{\operatorname{sech}^2(\beta)}{2}\sqrt{\frac{\varepsilon}{n}}$ and hence the localization algorithm of Lemma 25 would identify edge e with probability at least 9/10 using $\tilde{O}\left(\frac{c_1(\beta) n}{\varepsilon}\right)$ samples (where $c_1(\beta) = \cosh^4(\beta)$). If $\theta^p_{uv} - \theta^q_{uv} > 1/2 \tanh(\beta)$, then $\tanh(\theta^p_{uv}) - \tanh(\theta^q_{uv}) \geq \tanh(\beta) - \tanh\left(\beta - \frac{1}{2\tanh(\beta)}\right)$ and hence the localization algorithm of Lemma 25 would identify edge e with probability at least 9/10 using $\tilde{O}\left(c_2(\beta)\right)$ samples where $c_2(\beta) = \frac{1}{(\tanh(\beta) - \tanh(\beta - 1/2\tanh(\beta)))^2}$. Note that as β grows small, $c_2(\beta)$ gets worse. However it cannot grow unbounded as we also have to satisfy the constraint that $\theta^p_{uv} - \theta^q_{uv} \leq 2\beta$. This implies that

$$c_2(\beta) = \min\left\{\beta^2, \frac{1}{(\tanh(\beta) - \tanh(\beta - 1/2\tanh(\beta)))^2}\right\}$$

  samples suffice in this case. Therefore the algorithm will give the correct output with probability $> 9/10$ using $\tilde{O}\left(c(\beta)\frac{n}{\varepsilon}\right)$ samples where $c(\beta) = \max\{c_1(\beta), c_2(\beta)\}$.

- *Case 2:* All edges $e = (u, v)$ are such that $|\theta^q_{uv} - \theta^q_{uv}| \leq \sqrt{\frac{\varepsilon}{n}}$. In this case we have,

$$\begin{aligned} d_{\mathrm{SKL}}(p, q) &\geq \varepsilon && (4.16)\\ \implies \exists \text{ edge } e = (u, v) \text{ s.t } (\theta^p_{uv} - \theta^q_{uv})\left(\mathbf{E}[X_u X_v] - \mu^q_{uv}\right) &\geq \frac{\varepsilon}{n} && (4.17)\\ \implies \exists \text{ edge } e = (u, v) \text{ s.t } |\mathbf{E}[X_u X_v] - \mu^q_{uv}| &\geq \left|\frac{\varepsilon}{n} \times \sqrt{\frac{n}{\varepsilon}}\right| && (4.18)\\ &= \sqrt{\frac{\varepsilon}{n}} && (4.19) \end{aligned}$$

  Hence, the localization algorithm of Lemma 25 would identify such an edge with probability at least 9/10.

□

4.4.2 Testing on Ferromagnets

In this section we will describe an algorithm for testing independence of ferromagnetic Ising models under no external field. The tester follows the localization based recipe of Section 4.3 but leverages additional structural information about ferromagnets to obtain an improved sample complexity.

At a high level, the algorithm is as follows: if there exists an edge with a large edge parameter, then we lower bound its marginal by $\tanh(\theta_{uv})$ where uv is the edge under consideration. This implies that its marginal sticks out and is easy to catch via performing local tests on all edges. If all the edge parameters were small, then Algorithm 2 is already efficient.

We first prove a structural lemma about ferromagnetic Ising models. We will use the Fortuin-Kasteleyn random cluster model and its coupling with the Ising model (described in Chapter 10 of [RAS15]) to argue that in any ferromagnetic Ising model $\mu_{uv} \geq \tanh(\theta_{uv})$ for all pairs u, v.

4.4.2.1 Random Cluster Model

Let $G = (V, E)$ be a finite graph. The random cluster measure is a probability distribution on the space $\Omega = \{0, 1\}^E$ of bond configurations denoted by $\eta = (\eta(e))_{e \in E} \in \{0, 1\}^E$. Each edge has an associated bond $\eta(e)$. $\eta(e) = 1$ denotes that bond e is open or present and $\eta(e) = 0$ implies that bond e is closed or unavailable. A random cluster measure is parameterized by an edge probability $0 < r < 1$ and by a second parameter $0 < s < \infty$. Let $k(\eta)$ denote the number of connected components in the graph (V, η). The random cluster measure is defined by

$$\rho_{r,s}(\eta) = \frac{1}{Z_{r,s}} \left(\prod_{e \in E} r^{\eta(e)} (1 - r)^{1 - \eta(e)} \right) s^{k(\eta)}$$

where $Z_{r,s}$ is a normalizing factor to make ρ a probability density. We consider a generalization of the random cluster model where each edge is allowed to have its own parameter $0 < r_e < 1$. Under this generalization, the measure becomes

$$\rho_{\vec{r},s}(\eta) = \frac{1}{Z_{\vec{r},s}} \left(\prod_{e \in E} r_e^{\eta(e)} (1 - r_e)^{1 - \eta(e)} \right) s^{k(\eta)}. \tag{4.20}$$

The random cluster measure is stochastically increasing in $\vec{r}$ when $s \geq 1$. This property is formally stated in Lemma 10.3 of [RAS15]. We state a generalized version of the Lemma here which holds when each edge is allowed its own probability parameter r_e.

Lemma 29. *[Lemma 10.3 from [RAS15]] For $s \geq 1$, and $\vec{r_1} \leq \vec{r_2}$ coordinate-wise, $\rho_{\vec{r_1},s} \leq \rho_{\vec{r_2},s}$ where given two bond configurations η_1 and η_2, $\eta_1 \geq \eta_2$ iff $\eta_1(e) = 1$ for all e such that $\eta_2(e) = 1$.*

4.4.2.2 Coupling between the Random Cluster Model and the Ising model

We will now describe a coupling between the random cluster measure and the probability density function for a ferromagnetic Ising model. In particular, the edge probability r_e under the random cluster measure and the edge parameters θ_e of the Ising model are related by

$$r_e = 1 - \exp(-2\theta_e)$$

and the parameter $s = 2$ because the Ising model has two spins ± 1. The coupling Q will be a joint distribution on the spin variables $X = (X_1 \dots X_n)$ of the Ising model and the bond variables $\eta = (\eta(e))_{e \in E}$. The measure Q is defined as

$$Q(X, \eta) = \frac{1}{Z} \prod_{e=(u,v) \in E} r_e^{\eta(e)} (1 - r_e)^{1-\eta(e)} \left(\mathbb{1}_{X_u = X_v} + (1 - \eta(e)) \mathbb{1}_{X_u \neq X_v} \right)$$

where Z is a normalizing constant so as to make Q a probability measure. Under the relation stated above between r_e and θ_e, the following properties regarding the marginal distributions of Q hold.

$$\begin{aligned} \sum_{\eta \in \{0,1\}^E} Q(X, \eta) &= \frac{1}{Z'} \exp\left(\sum_{u \neq v} \theta_{uv} X_u X_v \right) \\ \sum_{X \in \{\pm 1\}^n} Q(X, \eta) &= \frac{1}{Z''} \left(\prod_{e \in E} r_e^{\eta(e)} (1 - r_e)^{1-\eta(e)} \right) 2^{k(\eta)} = \rho_{\vec{r},2}(\eta) \end{aligned} \tag{4.21}$$

where Z', Z'' are normalizing constants to make the marginals probability densities. The above equations imply that the measure Q is a valid coupling and more importantly they yield an alternative way to sample from the Ising model as follows:

First sample a bond configuration η according to $\rho_{\vec{r},2}(\eta)$. For each connected component in the bond graph, flip a fair coin to determine if the variables in that component will be all $+1$ or all -1.

In addition to the above information about the marginals of Q, we will need the following simple observations.

1. $Q(X, \eta) = 0$ if $\eta(e) = 1$ for any $e \notin E$.
2. $Q(X, \eta) = 0$ if for any $e = (u, v) \in E$, $\eta(e) = 1$ and $X_u \neq X_v$.

Next we state another property of the coupling $Q(.,.)$ which says that if two nodes u and v are in different connected components in the bond graph specified by η, then the

probability that $X_u = X_v$ is the same as the probability that $X_u \neq X_v$.

Claim 3. *Let $C_\eta(u, v)$ denote the predicate that under the bond configuration η, u and v are connected with a path of open bonds. Then,*

$$\sum_{\substack{\eta \ s.t \\ C_\eta(u,v)=0}} \sum_{\substack{X \ s.t. \\ X_u = X_v}} Q(X, \eta) = \sum_{\substack{\eta \ s.t \\ C_\eta(u,v)=0}} \sum_{\substack{X \ s.t. \\ X_u \neq X_v}} Q(X, \eta)$$

The proof of the above claim is quite simple and follows by matching the appropriate terms in the probability density Q when u and v lie in different connected components. The proof is omitted here.

Armed with the coupling Q and its properties stated above, we are now ready to state the main structural lemma we show for ferromagnetic Ising models.

Lemma 30. *Consider two ferromagnetic Ising models p and q under no external field defined on $G_p = (V, E_p)$ and $G_q = (V, E_q)$. Denote the parameter vector of p model by $\vec{\theta^p}$ and that of q model by $\vec{\theta^q}$. If $\vec{\theta^p} \geq \vec{\theta^q}$ coordinate-wise, then for any two nodes $u, v \in V$, $\mu^p_{uv} \geq \mu^q_{uv}$.*

Proof. Since

$$\begin{aligned} \mu^p_{uv} &= \Pr_p [X_u = X_v] - \Pr_p [X_u \neq X_v] \\ \implies \mu^p_{uv} &= 2 \Pr_p [X_u = X_v] - 1 \end{aligned}$$

to show that $\mu^p_{uv} \geq \mu^q_{uv}$ it suffices to show that $\Pr_p [X_u = X_v] \geq \Pr_q [X_u = X_v]$. Consider the coupling $Q(X, \eta)$ described above between the random cluster measure and the Ising model probability. $\Pr_p [X_u = X_v]$ can be expressed in terms of $Q_p(X, \eta)$ as follows:

$$\Pr_p [X_u = X_v] = \sum_{\substack{X \text{ s.t.} \\ X_u = X_v}} \sum_{\eta} Q_p(X, \eta)$$

Denote the sum on the right in the above equation by S_p. It suffices to show that $S_p \geq S_q$.

Lemma 10.3 of [RAS15] gives that for any bond configuration η_0,

$$\sum_{\eta \geq \eta_0} \rho_p^{E_b}(\eta) \geq \sum_{\eta \geq \eta_0} \rho_q^{E_b}(\eta).$$

This follows because the parameter vectors of p and q satisfy the condition of the lemma that $\vec{\theta}^p \geq \vec{\theta}^q$. Again, let $C_\eta(u, v)$ denote the predicate that under the bond configuration η, u and v are connected. Let H be the set of all bond configurations such that u and v are connected by a single distinct path. Therefore $C_{\eta_0}(u, v) = 1$ for all $\eta_0 \in H$. Then the set

$$C = \{\eta | \eta \geq \eta_0 \text{ for some } \eta_0 \in H\}$$

represents precisely the bond configurations in which u and v are connected. Applying Lemma 10.3 of [RAS15] on each $\eta_0 \in H$ and summing up the inequalities obtained, we get

$$\begin{aligned}
&\sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \rho_p^{E_b}(\eta) \geq \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \rho_q^{E_b}(\eta) \\
\implies &\sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_X Q_p(X, \eta) \geq \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_X Q_q(X, \eta) \\
\implies &\sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u = X_v}} Q_p(X, \eta) \geq \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u = X_v}} Q_q(X, \eta)
\end{aligned} \tag{4.22}$$

where the last inequality follows because $Q(X, \eta) = 0$ if for any pair u, v, $\eta(uv) = 1$ but $X_u \neq X_v$.

Also, from Claim 3, we have that for any Ising model,

$$\sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u = X_v}} Q(X, \eta) = \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u \neq X_v}} Q(X, \eta) \tag{4.23}$$

And since $Q(.,.)$ is a probability measure we have that for any Ising model,

$$\sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q(X,\eta) + \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q(X,\eta) + \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u\neq X_v}} Q(X,\eta) + \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u\neq X_v}} Q(X,\eta) = 1 \tag{4.24}$$

$$\implies \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q(X,\eta) + \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q(X,\eta) + \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u\neq X_v}} Q(X,\eta) = 1 \tag{4.25}$$

$$\implies \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q(X,\eta) + 2 \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q(X,\eta) = 1 \tag{4.26}$$

where (4.25) follows because the last term in (4.24) is 0 and (4.26) follows from (4.23).

Equation (4.26) implies that

$$\begin{aligned} S_p &= \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q_p(X,\eta) + \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=0}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q_p(X,\eta) \\ &= \frac{1}{2} \sum_{\substack{\eta \text{ s.t} \\ C_\eta(u,v)=1}} \sum_{\substack{X \text{ s.t.} \\ X_u=X_v}} Q_p(X,\eta) + \frac{1}{2} \end{aligned}$$

Therefore from (4.22), we get

$$S_p \geq S_q$$

□

Using the above lemma, we now prove the main structural lemma for ferromagnets which will be crucial to our algorithm for testing ferromagnetic Ising models.

Lemma 31 (Structural Lemma about Ferromagnetic Ising Models)**.** *If $X \sim p$ is a ferromagnetic Ising model on a graph $G = (V, E)$ under zero external field, then $\mu_{uv} \geq \tanh(\theta_{uv})$ for all edges $(u, v) \in E$.*

Proof. Fix the edge of concern $e = (u, v)$. If the graph doesn't contain cycles, then from Lemma 28 $\mu_{uv} = \tanh(\theta_{uv})$ and the statement is true. To show that the statement holds for general graphs we will use induction on the structure of the graph. Graph G can be constructed as follows. Start with the single edge $e = (u, v)$ and then add the remaining edges in $E\backslash\{e\}$ one by one in some order. Denote the intermediate graphs obtained during

this process as $G_0, G_1, \ldots, G_m = G$ where G_0 is the graph consisting of just a single edge. For each graph G_i we can associate the corresponding Ising model p_i to be the model which has $\theta_e^{p_i} = \theta_e$ for $e \in E_{G_i}$ and $\theta_e^{p_i} = 0$ otherwise. For each graph G_i in the sequence, we will use $\mu_{uv}^{p_i}$ to denote $\mathbf{E}[X_u X_v]$ for the Ising model corresponding to graph G_i. We will prove that $\mu_{uv}^{p} \geq \tanh(\theta_{uv})$ by induction on this sequence of graphs. The statement can be easily verified to be true for G_0. In fact, $\mu_{uv}^{p_0} = \tanh(\theta_{uv})$. Suppose the statement was true for some G_i in the sequence. By Lemma 30, we have that $\mu_{uv}^{p_{i+1}} \geq \mu_{uv}^{p_i}$. This implies that $\mu_{uv}^{Gpi+1} \geq \tanh(\theta_{uv})$ hence showing the statement to be true for all graphs G_i in the sequence. □

Given the above structural lemma about ferromagnetic Ising models under no external field, we present the following algorithm for testing whether a ferromagnetic Ising model is product or not.

Algorithm 6 Test if a ferromagnetic Ising model p under no external field is product

1: **function** TESTFERROISING-INDEPENDENCE(sample access to an Ising model p)
2: Run the algorithm of Lemma 25 to identify if all edges $e = (u, v)$ such that $\mathbf{E}[X_u X_v] \geq \sqrt{\varepsilon}/n$ using $\tilde{O}\left(\frac{n^2}{\varepsilon}\right)$ samples. If it identifies any edges, return that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$
3: Otherwise, return that p is product.
4: **end function**

Proof of Theorem 25: Firstly, note that under no external field, the only product Ising model is the uniform distribution $\mathcal{U}_n$. To the problem reduces to testing whether p is uniform or not. Consider the case when p is indeed uniform. That is, there are no edges in the underlying graph of the Ising model. In this case with probability at least 9/10 the localization algorithm of Lemma 25 with output no edges. Hence Algorithm 6 will output that p is product.
In case $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$, we split the analysis into two cases.

- *Case 1:* There exists an edge $e = (u, v)$ such that $|\theta_{uv}| \geq \sqrt{\frac{\varepsilon}{n^2}}$. In this case, $|\mathbf{E}[X_u X_v]| \geq |\tanh(\theta_{uv})|$ and in the regime where ε is a fixed constant, $|\tanh(\theta)| \geq |\theta/2|$. Hence implying that $|\mathbf{E}[X_u X_v]| \geq |\theta_{uv}/2| \geq \sqrt{\frac{\varepsilon}{n^2}}/2$. Therefore the localization algorithm of Lemma 25 would identify such an edge with probability at least 9/10. (The regime where the inequality $|\tanh(\theta)| \geq |\theta/2|$ isn't valid would be easily detectable using $\tilde{O}(\frac{n^2}{\varepsilon})$ samples.)

- *Case 2:* All edges $e = (u, v)$ are such that $\theta_{uv} \leq \sqrt{\frac{\varepsilon}{n^2}}$. In this case we have,

$$\begin{aligned} d_{\mathrm{SKL}}(X, \mathcal{I}) &\geq \varepsilon && (4.27) \\ \implies \exists \text{ edge } e = (u, v) \text{ s.t } \theta_{uv}\mathbf{E}[X_u X_v] &\geq \frac{\varepsilon}{n^2} && (4.28) \\ \implies \exists \text{ edge } e = (u, v) \text{ s.t } \mathbf{E}[X_u X_v] &\geq \frac{\varepsilon}{n^2} \times \sqrt{\frac{n^2}{\varepsilon}} && (4.29) \\ &= \sqrt{\frac{\varepsilon}{n^2}} && (4.30) \end{aligned}$$

  Hence, the localization algorithm of Lemma 25 would identify such an edge with probability at least 9/10.

□

4.5 Improved Testing in High-Temperature

In this section, we describe a framework for testing Ising models in the high-temperature regime which results in algorithms which are more efficient than our baseline localization algorithm of Section 4.3 for dense graphs. This is the more technically involved part of our result and we modularize the description and analysis into different parts. We will give a high level overview of our approach here.

The main approach we take in this section is to consider a global test statistic over all the variables on the Ising model in contrast to the localized statistics of Section 4.3. For ease of exposition, we first describe the approach for testing independence under no external field. We then describe the changes that need to be made to obtain tests for independence under an external field and goodness-of-fit in Section 4.5.5.

Note that testing independence under no external field boils down to testing uniformity as the only independent Ising model when there is no external field is the one corresponding to the uniform distribution. The intuition for the core of the algorithm is as follows. Suppose we are interested in testing uniformity of Ising model p with parameter vector $\vec{\theta}$. Note that for the uniform Ising model, $\theta_{uv} = \theta_u = 0$ for all $u, v \in V$. We start by obtaining an upper bound on the SKL between p and $\mathcal{U}_n$ which can be captured via a statistic that does not

depend on $\vec{\theta}$. From (4.2), we have that under no external field ($\theta_u = 0$ for all $u \in V$),

$$d_{\mathrm{SKL}}(p, \mathcal{U}_n) = \sum_{e=(u,v)\in E} \theta_{uv}\mu_{uv}$$

$$\implies d_{\mathrm{SKL}}(p, \mathcal{U}_n) \leq \sum_{u\neq v} \beta \left|\mu_{uv}\right| \tag{4.31}$$

$$\implies \frac{d_{\mathrm{SKL}}(p, \mathcal{U}_n)}{\beta} \leq \sum_{u\neq v} \left|\mu_{uv}\right|. \tag{4.32}$$

where (4.31) holds because $|\theta_{uv}| \leq \beta$.

Given the above upper bound, we consider the statistic $Z = \sum_{u\neq v} \mathbf{sign}(\mu_{uv}) \cdot (X_u X_v)$, where $X \sim p$ and $\mathbf{sign}(\mu_{uv})$ is chosen arbitrarily if $\mu_{uv} = 0$.

$$\mathbf{E}[Z] = \sum_{u\neq v} \left|\mu_{uv}\right|.$$

If $X \in \mathcal{I}$, then $\mathbf{E}[Z] = 0$. On the other hand, by (4.32), we know that if $d_{\mathrm{SKL}}(X, \mathcal{I}) \geq \varepsilon$, then $\mathbf{E}[Z] \geq \varepsilon/\beta$. If the $\mathbf{sign}(\mu_e)$ parameters were known, we could simply plug them into Z, and using Chebyshev's inequality, distinguish these two cases using $\mathbf{Var}(Z)\beta^2/\varepsilon^2$ samples.

There are two main challenges here.

- First, the sign parameters, $\mathbf{sign}(\mu_{uv})$, are *not* known.
- Second, it is not obvious how to get a non-trivial bound for $\mathbf{Var}(Z)$.

One can quickly see that learning all the sign parameters might be prohibitively expensive. For example, if there is an edge e such that $|\mu_e| = 1/2^n$, there would be no hope of correctly estimating its sign with a polynomial number of samples. Instead, we perform a process we call *weak learning* – rather than trying to correctly estimate all the signs, we instead aim to obtain a $\vec{\Gamma}$ which is *correlated* with the vector $\mathbf{sign}(\mu_e)$. In particular, we aim to obtain $\vec{\Gamma}$ such that, in the case where $d_{\mathrm{SKL}}(p, \mathcal{U}_n) \geq \varepsilon$, $\mathbf{E}[\sum_{e=(u,v)\in E} \Gamma_e (X_u X_v)] \geq \varepsilon/\zeta\beta$, where $\zeta = \mathrm{poly}(n)$. That is we learn a sign vector $\vec{\Gamma}$ which is correlated enough with the true sign vector such that a sufficient portion of the signal from the d_{SKL} expression is still preserved. The main difficulty of analyzing this process is due to correlations between random variables $(X_u X_v)$. Naively, we could get an appropriate Γ_e for $(X_u X_v)$ by running

a weak learning process independently for each edge. However, this incurs a prohibitive cost of $O(n^2)$ by iterating over all edges. We manage to sidestep this cost by showing that, despite these correlations, learning all Γ_e simultaneously succeeds with a probability which is $\geq 1/\operatorname{poly}(n)$, for a moderate polynomial in n. Thus, repeating this process several times, we can obtain a $\vec{\Gamma}$ which has the appropriate guarantee with sufficient constant probability.

At this point, we are in the setting as described above – we have a statistic Z' of the form:

$$Z' = \sum_{u \neq v} c_{uv} X_u X_v \tag{4.33}$$

where $c \in \{\pm 1\}^{\binom{V}{2}}$ represent the signs obtained from the weak learning procedure. $\mathbf{E}[Z'] = 0$ if $X \in \mathcal{I}$, and $\mathbf{E}[Z'] \geq \varepsilon/\zeta\beta$ if $d_{\mathrm{SKL}}(X, \mathcal{I}) \geq \varepsilon$. These two cases can be distinguished using $\mathbf{Var}(Z')\zeta^2\beta^2/\varepsilon^2$ samples, by Chebyshev's inequality. At this point, we run into the second issue mentioned above. Since the range of Z' is $\Omega(n^2)$, a crude bound for $\mathbf{Var}(Z')$ is $O(n^4)$, granting us no savings over the localization algorithm of Theorem 22. However, in the high temperature regime, we show the following bound on the variance of Z' (Theorem 37).

$$\mathbf{Var}(Z') = \tilde{O}(n^2).$$

In other words, despite the potentially complex structure of the Ising model and potential correlations, the variables $X_u X_v$ contribute to the variance of Z' roughly as if they were all independent! We describe the result and techniques involved in the analysis of the variance bound in Section 4.8. Given the tighter bound on the variance of our statistic, we run the Chebyshev-based test on all the hypotheses obtained in the previous learning step (with appropriate failure probability) to conclude our algorithm. Further details about the algorithm are provided in Sections 4.5.1-4.5.4.

We state the sample complexity achieved via our learn-then-test framework for independence testing under no external field here. The corresponding statements for independence testing under external fields and identity testing are given in Section 4.5.5.

Theorem 26 (Independence Testing using Learn-Then-Test, No External Field)**.** *Suppose*

p is an Ising model in the high temperature regime under no external field. Then, given $\tilde{O}\left(\frac{n^{10/3}\beta^2}{\varepsilon^2}\right)$ i.i.d samples from p, the learn-then-test algorithm runs in polynomial time and distinguishes between the cases $p \in \mathcal{I}$ and $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$ with probability at least $9/10$.

Next, we state a corollary of Theorem 26 with sample complexities we obtain when β is close to the high temperature threshold.

Theorem 27 (Independence Testing with β near the Threshold of High Temperature, No External Field)**.** *Suppose that p is an Ising model in the high temperature regime and suppose that $\beta = \frac{1}{4\delta_{\max}}$. That is, β is close to the high temperature threshold. Then:*

- *Given $\tilde{O}\left(\frac{n^{10/3}}{\varepsilon^2\delta_{\max}^2}\right)$ i.i.d samples from p* ***with no external field****, the learn-then-test algorithm runs in polynomial time and distinguishes between the cases $p \in \mathcal{I}$ and $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$ with probability at least $2/3$. For testing identity of p to an Ising model q in the high temperature regime, we obtain the same sample complexity as above.*

Figure 4-1 shows the dependence of sample complexity of testing as $\delta_{\max}$ is varied in the regime of Theorem 27 for the case of no external field.

The description of our algorithm is presented in Algorithm 7. It contains a parameter τ, which we choose to be the value achieving the minimum in the sample complexity of Theorem 28. The algorithm follows a learn-then-test framework, which we outline here.

Algorithm 7 Test if an Ising model p under no external field is product using Learn-Then-Test

function LEARN-THEN-TEST-ISING(sample access to an Ising model $p, \beta, \delta_{\max}, \varepsilon, \tau$)
  Run the localization Algorithm 2 on p with accuracy parameter $\frac{\varepsilon}{n^\tau}$. If it identifies. any edges, return that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$
  for $\ell = 1$ to $O(n^{2-\tau})$ **do**
    Run the weak learning Algorithm 8 on $S = \{X_u X_v\}_{u \neq v}$ with parameters τ and ε/β. to generate a sign vector $\vec{\Gamma}^{(\ell)}$ where $\Gamma_{uv}^{(\ell)}$ is weakly correlated with $\mathbf{sign}\left(\mathbf{E}\left[X_{uv}\right]\right)$
  end for
  Using the *same set of samples for all ℓ*, run the testing algorithm of Lemma 34 on each. of the $\vec{\Gamma}^{(\ell)}$ with parameters $\tau_2 = \tau, \delta = O(1/n^{2-\tau})$. If any output that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$, return that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$. Otherwise, return that $p \in \mathcal{I}$
end function

Note: The first step in Algorithm 7 is to perform a localization test to check if $|\mu_e|$ is not too far away from 0 for all e. It is added to help simplify the analysis of the algorithm

and is not necessary in principle. In particular, we use the first part of Algorithm 2, which checks if any edge looks far from uniform, to perform this first step, albeit with a smaller value of the accuracy parameter ε than before. Similar to before, if we find a single non-uniform edge, this is sufficient evidence to output $d_{\mathrm{SKL}}(X, \mathcal{I}) \geq \varepsilon$. If we do not find any edges which are verifiably far from uniform, we proceed onward, with the additional guarantee that $|\mu_e|$ is small for all $e \in E$.

A statement of the exact sample complexity achieved by our algorithm is given in Theorem 28. When optimized for the parameter τ, this yields Theorem 26.

Theorem 28. *Given $\tilde{O}\left(\min_{\tau>0}\left(n^{2+\tau} + n^{6-2\tau}\right)\frac{\beta^2}{\varepsilon^2}\right)$ i.i.d samples from an Ising model p in the high-temperature regime with no external field, there exists a polynomial-time algorithm which distinguishes between the cases $p \in \mathcal{I}$ and $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$ with probability at least $2/3$.*

The organization of the rest of the section is as follows. We describe and analyze our weak learning procedure in Section 4.5.1. Given a vector with the appropriate weak learning guarantees, we describe and analyze the testing procedure in Section 4.5.2. In Section 4.5.3, we describe how to combine all these ideas – in particular, our various steps have several parameters, and we describe how to balance the complexities to obtain the sample complexity stated in Theorem 28. Finally, in Section 4.5.4, we optimize the sample complexities from Theorem 28 for the parameter τ and filter out cleaner statement of Theorem 26. We compare the performance of our localization and learn-then-test algorithms and describe the best sample complexity achieved in different regimes in Section 4.6.

4.5.1 Learn ...

Our overall goal of this section is "weakly learn" the sign of $\mu_e = \mathbf{E}[X_u X_v]$ for all edges $e = (u, v)$. More specifically, we wish to output a vector $\vec{\Gamma}$ with the following guarantee:

$$\mathbf{E}_X\left[\sum_{e=(u,v)\in E} \Gamma_e X_u X_v\right] \geq \frac{c\varepsilon}{2\beta n^{2-\tau_2}},$$

for some constant $c > 0$ and parameter τ_2 to be specified later. Note that the "best" Γ, for which $\Gamma_e = \mathbf{sign}(\mu_e)$, has this guarantee with $\tau_2 = 2$ – by relaxing our required learning

guarantee, we can reduce the sample complexity in this stage.

The first step will be to prove a simple but crucial lemma answering the following question: Given k samples from a Rademacher random variable with parameter p, how well can we estimate the sign of its expectation? This type of problem is well studied in the regime where $k = \Omega(1/p^2)$, in which we have a constant probability of success (see, i.e. Lemma 23), but we analyze the case when $k \ll 1/p^2$ and prove how much better one can do versus randomly guessing the sign. See Lemma 40 in Section 4.10 for more details.

With this lemma in hand, we proceed to describe the weak learning procedure. Given parameters τ,ε and sample access to a set S of 'Rademacher-like' random variables which may be *arbitrarily correlated* with each other, the algorithm draws $\tilde{O}\left(\frac{n^{2\tau}}{\varepsilon^2}\right)$ samples from each random variable in the set and computes their empirical expected values and outputs a signs of thus obtained empirical expectations. The procedure is described in Algorithm 8.

Algorithm 8 Weakly Learn Signs of the Expectations of a set of Rademacher-like random variables

function WeakLearning(sample access to set $S = \{Z_i\}_i$ of random variables where $|S| = O(n^s)$ and where $Z_i \in \{-1, 0, +1\}$ and can be arbitrarily correlated,ε, τ,).
    Draw $k = \tilde{O}\left(\frac{n^{2\tau}}{\varepsilon^2}\right)$ samples from each Z_i. Denote the samples by $Z_i^{(1)}, \ldots, Z_i^{(k)}$.
    Compute the empirical expectation for each Z_i: $\hat{Z}_i = \frac{1}{k}\sum_{l=1}^{k} Z_i^{(l)}$.
    Output $\vec{\Gamma}$ where $\Gamma_i = \mathbf{sign}(\hat{Z}_i)$.
end function

We now turn to the setting of the Ising model, discussed in Section 4.5.1.1. We invoke the weak-learning procedure of Algorithm 8 on the set $S = \{X_u X_v\}_{u \neq v}$ with parameters ε/β and $0 \leq \tau \leq 2$. By linearity of expectations and Cauchy-Schwarz, it is not hard to see that we can get a guarantee of the form we want in expectation (see Lemma 32). However, the challenge remains to obtain this guarantee with constant probability. Carefully analyzing the range of the random variable and using this guarantee on the expectation allows us to output an appropriate vector $\vec{\Gamma}$ with probability inversely polynomial in n (see Lemma 33). Repeating this process several times will allow us to generate a collection of candidates $\{\vec{\Gamma}^{(\ell)}\}$, at least one of which has our desired guarantees with constant probability.

4.5.1.1 Weak Learning the Edges of an Ising Model

We now turn our attention to weakly learning the edge correlations in the Ising model. To recall, our overall goal is to obtain a vector $\vec{\Gamma}$ such that

$$\mathbf{E}_{X\sim p}\left[\sum_{e=(u,v)\in E}\Gamma_e X_u X_v\right] \geq \frac{c\varepsilon}{2\beta n^{2-\tau_2}}.$$

We start by proving that the weak learning algorithm 8 yields a $\vec{\Gamma}$ for which such a bound holds in expectation. The following is fairly straightforward from Lemma 40 and linearity of expectations.

Lemma 32. *Given $k = O\left(\frac{n^{2\tau_2}\beta^2}{\varepsilon^2}\right)$ samples from an Ising model $X \sim p$ such that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$ and $|\mu_e| \leq \frac{\varepsilon}{\beta n^{\tau_2}}$ for all $e \in E$, Algorithm 8 outputs $\vec{\Gamma} = \{\Gamma_e\} \in \{\pm 1\}^{|E|}$ such that*

$$\mathbf{E}_{\vec{\Gamma}}\left[\mathbf{E}_{X\sim p}\left[\sum_{e=(u,v)\in E}\Gamma_e X_u X_v\right]\right] \geq \frac{c\beta}{\varepsilon n^{2-\tau_2}}\left(\sum_{e\in E}|\mu_e|\right)^2,$$

for some constant $c > 0$.

Proof. Since for all $e = (u, v) \in E$, $|\mu_e| \leq \frac{\varepsilon}{\beta n^{\tau_2}}$, and by our upper bound on k, all of the random variables $X_u X_v$ fall into the first case of Lemma 40 (the "small k" regime). Hence, we get that

$$\Pr\left[\Gamma_e = \mathbf{sign}(\mu_e)\right] \geq \frac{1}{2} + \frac{c_1|\mu_e|\sqrt{k}}{2}$$

which implies that

$$\begin{aligned}\mathbf{E}_{\Gamma_e}\left[\Gamma_e\mu_e\right] &\geq \left(\frac{1}{2} + \frac{c_1|\mu_e|\sqrt{k}}{2}\right)|\mu_e| + \left(\frac{1}{2} - \frac{c_1|\mu_e|\sqrt{k}}{2}\right)(-|\mu_e|) \\ &= c_1|\mu_e|^2\sqrt{k}\end{aligned}$$

Summing up the above bound over all edges, we get

$$\begin{aligned}\mathbf{E}_{\vec{\Gamma}}\left[\sum_{e\in E}\Gamma_e\mu_e\right] &\geq c_1\sqrt{k}\sum_{e\in E}|\mu_e|^2 \\ &\geq \frac{c_1' n^{\tau_2}\beta}{\varepsilon}\sum_{e\in E}|\mu_e|^2,\end{aligned}$$

for some constant $c_1' > 0$. Applying the Cauchy-Schwarz inequality gives us

$$\mathbf{E}_{\vec{\Gamma}}\left[\sum_{e\in E}\Gamma_e\mu_e\right] \geq \frac{c\beta}{\varepsilon n^{2-\tau_2}}\left(\sum_{e\in E}|\mu_e|\right)^2,$$

as desired. □

Next, we prove that the desired bound holds with sufficiently high probability. The following lemma follows by a careful analysis of the extreme points of the random variable's range.

Lemma 33. *Given $k = O\left(\frac{n^{2\tau_2}\beta^2}{\varepsilon^2}\right)$ i.i.d. samples from an Ising model p such that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$ and $|\mu_e| \leq \frac{\varepsilon}{\beta n^{\tau_2}}$ for all $e \in E$, Algorithm 8 outputs $\vec{\Gamma} = \{\Gamma_e\} \in \{\pm 1\}^{|E|}$ where: Define χ_{τ_2} to be the event that*

$$\mathbf{E}_{X\sim p}\left[\sum_{e=(u,v)\in E}\Gamma_e X_u X_v\right] \geq \frac{c\varepsilon}{2\beta n^{2-\tau_2}},$$

for some constant $c > 0$. We have that

$$\mathbf{Pr}_{\Gamma}\left[\chi_{\tau_2}\right] \geq \frac{c}{4n^{2-\tau_2}}.$$

Proof. We introduce some notation which will help in the elucidation of the argument which follows. Let $r = \mathbf{Pr}_{\Gamma}\left[\chi_{\tau_2}\right]$. Let

$$T = \frac{c\beta}{2\varepsilon n^{2-\tau_2}}\left(\sum_{e\in E}|\mu_e|\right)^2.$$

Let Y be the random variable defined as follows

$$Y = \mathbf{E}_{X\sim p}\left[\sum_{e=(u,v)\in E} \Gamma_e X_u X_v\right],$$

$$U = \mathbf{E}_{\vec{\Gamma}}\left[Y | Y > T\right] \text{ and}$$

$$L = \mathbf{E}_{\vec{\Gamma}}\left[Y | Y \leq T\right]$$

Then we have

$$\begin{aligned} rU + (1-r)L &\geq 2T \text{ (From Lemma 32)} \\ \implies r &\geq \frac{2T - L}{U - L} \end{aligned}$$

Since $U \leq \sum_{e\in E} |\mu_e|$, we have

$$r \geq \frac{2T - L}{\left(\sum_{e\in E} |\mu_e|\right) - L}$$

Since $L \geq -\sum_{e\in E} |\mu_e|$,

$$r \geq \frac{2T - L}{2\left(\sum_{e\in E} |\mu_e|\right)}$$

Since $L \leq T$, we get

$$r \geq \frac{T}{2\left(\sum_{e\in E} |\mu_e|\right)}$$

Substituting in the value for T we get

$$r \geq \frac{c\beta\left(\sum_{e\in E} |\mu_e|\right)^2}{4\varepsilon n^{2-\tau_2}\left(\sum_{e\in E} |\mu_e|\right)}$$

$$\implies r \geq \frac{c\beta\left(\sum_{e\in E} |\mu_e|\right)}{4\varepsilon n^{2-\tau_2}}$$

Since $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$, this implies $\left(\sum_{e\in E} |\mu_e|\right) \geq \varepsilon/\beta$ and thus

$$r \geq \frac{c}{4n^{2-\tau_2}},$$

as desired. □

4.5.2 . . . Then Test!

In this section, we assume that we were successful in weakly learning a vector $\vec{\Gamma}$ which is "good" (i.e., it satisfies χ_{τ_2}, which says that the expectation the statistic with this vector is sufficiently large). With such a $\vec{\Gamma}$, we show that we can distinguish between $p \in \mathcal{I}$ and $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$.

Lemma 34. *Let p be an Ising model, let $X \sim p$, and let σ^2 be such that, for any $\vec{\gamma} = \{\gamma_e\} \in \{\pm 1\}^{|E|}$,*

$$\mathbf{Var}\left(\sum_{e=(u,v)\in E} \gamma_e X_u X_v\right) \leq \sigma^2.$$

Given $k = O\left(\sigma^2 \cdot \frac{n^{4-2\tau_2}\beta^2 \log(1/\delta)}{\varepsilon^2}\right)$ i.i.d samples from p, which satisfies either $p \in \mathcal{I}$ or $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$, and $\vec{\Gamma} = \{\Gamma_e\} \in \{\pm 1\}^{|E|}$ which satisfies χ_{τ_2} (as defined in Lemma 33) in the case that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$, then there exists an algorithm which distinguishes these two cases with probability $\geq 1 - \delta$.

Proof. We prove this lemma with failure probability $1/3$ – by standard boosting arguments, this can be lowered to δ by repeating the test $O(\log(1/\delta))$ times and taking the majority result.

Denote the ith sample as $X^{(i)}$. The algorithm will compute the statistic

$$Z = \frac{1}{k}\left(\sum_{i=1}^{k} \sum_{e=(u,v)\in E} \Gamma_e X_u^{(i)} X_v^{(i)}\right).$$

If $Z \leq \frac{c\varepsilon}{4\beta n^{2-\tau_2}}$, then the algorithm will output that $p \in \mathcal{I}$. Otherwise, it will output that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$.

By our assumptions in the lemma statement, in either case,

$$\mathbf{Var}\,(Z) \leq \frac{\sigma^2}{k}.$$

If $p \in \mathcal{I}$, then we have that

$$\mathbf{E}[Z] = 0.$$

By Chebyshev's inequality, this implies that

$$\Pr\left[Z \geq \frac{\varepsilon}{4\beta n^{2-\tau_2}}\right] \leq \frac{16\sigma^2\beta^2 n^{4-2\tau_2}}{kc^2\varepsilon^2}.$$

Substituting the value of k gives the desired bound in this case. The case where $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$ follows similarly, but additionally using the fact that χ_{τ_2} implies that

$$\mathbf{E}[Z] \geq \frac{c\varepsilon}{2\beta n^{2-\tau_2}}.$$

□

4.5.3 Putting Them Together

In this section, we combine lemmas from the previous sections to complete the proof of Theorem 28. Lemma 33 gives us that a single iteration of the weak learning step gives a "good" $\vec{\Gamma}$ with probability at least $\Omega\left(\frac{1}{n^{2-\tau_2}}\right)$. We repeat this step $O(n^{2-\tau_2})$ times, generating $O(n^{2-\tau_2})$ hypotheses $\vec{\Gamma}^{(\ell)}$. By standard tail bounds on geometric random variables, this will imply that at least one hypothesis is good (i.e. satisfying χ_{τ_2}) with probability at least 9/10. We then run the algorithm of Lemma 34 on each of these hypotheses, with failure probability $\delta = O(1/n^{2-\tau_2})$. If $p \in \mathcal{I}$, all the tests will output that $p \in \mathcal{I}$ with probability at least 9/10. Similarly, if $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$, conditioned on at least one hypothesis $\vec{\Gamma}^{(\ell^*)}$ being good, the test will output that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$ for this hypothesis with probability at least 9/10. This proves correctness of our algorithm.

To conclude our proof, we analyze its sample complexity. Combining the complexities of Lemmas 25, 33, and 34, the overall sample complexity is

$$O\left(\frac{n^{2\tau_1}\beta^2 \log n}{\varepsilon^2}\right) + O\left(\frac{n^{2+\tau_2}\beta^2}{\varepsilon^2}\right) + O\left(\sigma^2 \frac{n^{4-2\tau_2}\beta^2}{\varepsilon^2} \log n\right).$$

Noting that the first term is always dominated by the second term we can simplify the

complexity to the following expression.

$$O\left(\frac{n^{2+\tau_2}\beta^2}{\varepsilon^2}\right) + O\left(\sigma^2\frac{n^{4-2\tau_2}\beta^2}{\varepsilon^2}\log n\right). \tag{4.34}$$

Plugging in the variance bounds from Section 4.8, Theorems 37 and 38 gives Theorem 28.

4.5.4 Balancing Learning and Testing

The sample complexities in the statement of Theorem 28 arise from a combination of two separate algorithms and from a variance bound for our multi-linear statistic which depends on β and $\delta_{\max}$. To balance for the optimal value of τ in various regimes of β and $\delta_{\max}$ we use Claim 4 which can be easily verified and arrive at Theorem 26.

Claim 4. *Let* $S = \tilde{O}\left((n^{2+\tau} + n^{4-2\tau}\cdot\sigma^2)\frac{\beta^2}{\varepsilon^2}\right)$. *Let* $\sigma^2 = O(n^s)$. *The value of* τ *which minimizes* S *is* $\frac{2+s}{3}$.

Claim 4 together with the variance bound (Theorem 37) implies Theorem 26.

Theorem 26 (Independence Testing using Learn-Then-Test, No External Field)**.** *Suppose* p *is an Ising model in the high temperature regime under no external field. Then, given* $\tilde{O}\left(\frac{n^{10/3}\beta^2}{\varepsilon^2}\right)$ *i.i.d samples from* p*, the learn-then-test algorithm runs in polynomial time and distinguishes between the cases* $p \in \mathcal{I}$ *and* $d_{\mathrm{SKL}}(p,\mathcal{I}) \geq \varepsilon$ *with probability at least* $9/10$.

4.5.5 Modifications for Testing Independence and Identity

We describe the modifications that need to be done to the learn-then-test approach described in Sections 4.5.1-4.5.4 to obtain testers for independence under an arbitrary external field (Section 4.5.5.1), identity without an external field (Section 4.5.5.2), and identity under an external field (Section 4.5.5.3).

4.5.5.1 Independence Testing under an External Field

Under an external field, the statistic we considered in Section 4.5 needs to be modified. Suppose we are interested in testing independence of an Ising model p defined on a graph $G =$

(V, E) with a parameter vector $\vec{\theta^p}$. Let $X \sim p$. We have that $d_{\mathrm{SKL}}(p, \mathcal{I}) = \min_{q \in \mathcal{I}} d_{\mathrm{SKL}}(p, q)$. In particular, we consider q to be the independent Ising model on graph $G' = (V, E')$ with parameter vector $\vec{\theta^q}$ such that $E' = \phi$ and θ^q_u is such that $\mu^q_u = \mu^p_u$ for all $u \in V$. Then,

$$\begin{aligned} d_{\mathrm{SKL}}(p, \mathcal{I}) &\leq d_{\mathrm{SKL}}(p, q) \\ &= \sum_{e=(u,v) \in E} \theta^p_{uv} \left(\mu^p_{uv} - \mu^q_{uv}\right) \\ &= \sum_{e=(u,v) \in E} \theta^p_{uv} \left(\mu^p_{uv} - \mu^p_u \mu^p_v\right) \\ &\leq \sum_{e=(u,v) \in E} \beta \left|\mu^p_{uv} - \mu^p_u \mu^p_v\right| \\ \implies \frac{d_{\mathrm{SKL}}(p, \mathcal{I})}{\beta} &\leq \sum_{e=(u,v) \in E} \left|\mu^p_{uv} - \mu^p_u \mu^p_v\right|. \end{aligned} \tag{4.35}$$

The above inequality suggests a statistic Z such that $\mathbf{E}[Z] = \sum_{e=(u,v) \in E} |\lambda^p_{uv}|$ where $\lambda^p_{uv} = \mu^p_{uv} - \mu^p_u \mu^p_v$. We consider $Z = \sum_{u \neq v} \mathbf{sign}(\lambda_{uv}) \left(X^{(1)}_u - X^{(2)}_u\right) \left(X^{(1)}_v - X^{(2)}_v\right)$ where $X^{(1)}, X^{(2)} \sim p$ are two independent samples from p. It can be seen that Z has the desired expectation. However, we have the same issue as before that we don't know the $\mathbf{sign}(\lambda_{uv})$ parameters. Luckily, it turns out that our weak learning procedure is general enough to handle this case as well. Consider the following random variable: $Z_{uv} = \frac{1}{4} \left(X^{(1)}_u - X^{(2)}_u\right) \left(X^{(1)}_v - X^{(2)}_v\right)$. Z_{uv} takes on values in $\{-1, 0, +1\}$. Consider an associated Rademacher variable Z'_{uv} defined as follows: $\Pr[Z_{uv} = -1] = \Pr[Z_{uv} = -1] + 1/2 \Pr[Z_{uv} = 0]$. It is easy to simulate a sample from Z'_{uv} given access to a sample from Z_{uv}. If $Z_{uv} = 0$, toss a fair coin to decide whether $Z'_{uv} = -1$ or $+1$. $\mathbf{E}[Z'_{uv}] = \mathbf{E}[Z_{uv}] = \frac{\lambda_{uv}}{2}$. Hence $Z'_{uv} \sim \mathit{Rademacher}\left(\frac{1}{2} + \frac{\lambda_{uv}}{4}\right)$ and by Lemma 40 with k copies of the random variable Z_{uv} we get a success probability of $1/2 + c_1 \sqrt{k} \left|\lambda_{uv}\right|$ of estimating $\mathbf{sign}(\lambda_{uv})$ correctly. Given this guarantee, the rest of the weak learning argument of Lemmas 32 and 33 follows analogously by replacing μ_e with λ_e. After we have *weakly learnt* the signs, we are left with a statistic Z'_{cen} of the form:

$$Z'_{cen} = \sum_{u \neq v} c_{uv} \left(X^{(1)}_u - X^{(2)}_u\right) \left(X^{(1)}_v - X^{(2)}_v\right) \tag{4.36}$$

where the subscript *cen* denotes that the statistic is a centered one and $c \in \{\pm 1\}^{\binom{V}{2}}$. We need to obtain a bound on $\mathbf{Var}(Z'_{cen})$. We again employ the techniques described in Section 4.8 to obtain a non-trivial bound on $\mathbf{Var}(Z'_{cen})$ in the high-temperature regime. The statement of the variance result is given in Theorem 38 and the details are in Section 4.8.3. Combining the weak learning part and the variance bound gives us the following sample complexity for independence testing under an external field:

$$\begin{aligned}&\tilde{O}\left(\frac{(n^{2+\tau}+n^{4-2\tau}\sigma^2)\beta^2}{\varepsilon^2}\right)\\ =&\tilde{O}\left(\frac{(n^{2+\tau}+n^{4-2\tau}n^2)\beta^2}{\varepsilon^2}\right)\end{aligned}$$

Balancing for the optimal value of the τ parameter gives Theorem 29.

Theorem 29 (Independence Testing using Learn-Then-Test, Arbitrary External Field)**.** *Suppose p is an Ising model in the high temperature regime under an arbitrary external field. The learn-then-test algorithm takes in $\tilde{O}\left(\frac{n^{10/3}\beta^2}{\varepsilon^2}\right)$ i.i.d. samples from p and distinguishes between the cases $p \in \mathcal{I}$ and $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$ with probability $\geq 9/10$.*

The tester is formally described in Algorithm 9.

Algorithm 9 Test if an Ising model p under arbitrary external field is product

function LEARN-THEN-TEST-ISING(sample access to an Ising model $p, \beta, \delta_{\max}, \varepsilon, \tau$)
  Run the localization Algorithm 2 with accuracy parameter $\frac{\varepsilon}{n^\tau}$. If it identifies any edges, return that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$
  for $\ell = 1$ to $O(n^{2-\tau})$ **do**
    Run the weak learning Algorithm 8 on $S = \{(X_u^{(1)} - X_u^{(2)})(X_v^{(1)} - X_v^{(2)})\}_{u \neq v}$ with parameters $\tau_2 = \tau$ and ε/β to generate a sign vector $\vec{\Gamma}^{(\ell)}$ where $\Gamma_{uv}^{(\ell)}$ is weakly correlated with $\mathbf{sign}\left(\mathbf{E}\left[(X_u^{(1)} - X_u^{(2)})(X_v^{(1)} - X_v^{(2)})\right]\right)$
  end for
  Using the *same set of samples for all* ℓ, run the testing algorithm of Lemma 34 on each of the $\vec{\Gamma}^{(\ell)}$ with parameters $\tau_2 = \tau, \delta = O(1/n^{2-\tau})$. If any output that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$, return that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$. Otherwise, return that $p \in \mathcal{I}$
end function

4.5.5.2 Identity Testing under No External Field

We first look at the changes needed for identity testing under no external field. Similar to before, we start by obtaining an upper bound on the SKL between the Ising models p and q. We get that,

$$d_{\mathrm{SKL}}(p,q) = \sum_{(u,v)\in E} \left(\theta^p_{uv} - \theta^q_{uv}\right)\left(\mu^p_{uv} - \mu^q_{uv}\right)$$
$$\implies \frac{d_{\mathrm{SKL}}(p,q)}{2\beta} \leq \sum_{u\neq v} \left|\left(\mu^p_{uv} - \mu^q_{uv}\right)\right|$$

Since we know μ^q_{uv} for all pairs u, v, the above upper bound suggests the statistic Z of the form

$$Z = \sum_{u\neq v} \mathbf{sign}\left(\mu^p_{uv} - \mu^q_{uv}\right)\left(X_u X_v - \mu^q_{uv}\right)$$

If $p = q$, $\mathbf{E}[Z] = 0$ and if $d_{\mathrm{SKL}}(p,q) \geq \varepsilon$, $\mathbf{E}[Z] \geq \varepsilon/2\beta$. As before, there are two things we need to do: learn a sign vector which is weakly correlated with the right sign vector and obtain a bound on $\mathbf{Var}(Z)$. By separating out the part of the statistic which is just a constant, we obtain that

$$\mathbf{Var}(Z) \leq \mathbf{Var}\left(\sum_{u\neq v} c_{uv} X_u X_v\right)$$

where $c \in \{\pm 1\}^{\binom{V}{2}}$. Hence, the variance bound of Theorem 37 holds for $\mathbf{Var}(Z)$.

As for the weakly learning the signs, using Corollary 12 of Lemma 40 we get that for each pair u, v, with k samples, we can achieve a success probability $1/2 + c_1\sqrt{k}\left|\mu^p_{uv} - \mu^q_{uv}\right|$ of correctly estimating $\mathbf{sign}(\mu^p_{uv} - \mu^q_{uv})$. Following this up with analogous proofs of Lemmas 32 and 33 where μ_e is replaced by $\mu^p_e - \mu^q_e$, we achieve our goal of weakly learning the signs with a sufficient success probability.

By making these changes we arrive at the following theorem for testing identity to an Ising model under no external field.

Theorem 30 (Identity Testing using Learn-Then-Test, No External Field)**.** *Suppose p and q are Ising models in the high temperature regime under no external field. The learn-then-test algorithm takes in $\tilde{O}\left(\frac{n^{10/3}\beta^2}{\varepsilon^2}\right)$ i.i.d. samples from p and distinguishes between the cases $p = q$ and $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ with probability $\geq 9/10$.*

The tester is formally described in Algorithm 10.

Algorithm 10 Test if an Ising model p under no external field is identical to q

1: **function** TESTISING(sample access to an Ising model $p, \beta, \delta_{\max}, \varepsilon, \tau$, description of Ising model q under no external field)
2: Run the localization Algorithm 3 with accuracy parameter $\frac{\varepsilon}{n^\tau}$. If it identifies any. edges, return that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$
3: **for** $\ell = 1$ to $O(n^{2-\tau})$ **do**
4: Run the weak learning Algorithm 8 on $S = \{X_u X_v - \mu^q_{uv}\}_{u \neq v}$ with parameters. $\tau_2 = \tau$ and ε/β to generate a sign vector $\vec{\Gamma}^{(\ell)}$ where $\Gamma^{(\ell)}_{uv}$ is weakly correlated with $\mathbf{sign}\left(\mathbf{E}\left[X_{uv} - \mu^q_{uv}\right]\right)$
5: **end for**
6: Using the *same set of samples for all* ℓ, run the testing algorithm of Lemma 34 on each. of the $\vec{\Gamma}^{(\ell)}$ with parameters $\tau_2 = \tau, \delta = O(1/n^{2-\tau})$. If any output that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$, return that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$. Otherwise, return that $p = q$
7: **end function**

4.5.5.3 Identity Testing under an External Field

When an external field is present, two things change. Firstly, the terms corresponding to nodes of the Ising model in the SKL expression no longer vanish and have to be accounted for. Secondly, it is unclear how to define an appropriately centered statistic which has a variance of $O(n^2)$ in this setting, and we consider this an interesting open question. Instead, we use the uncentered statistic which has variance $\Theta(n^3)$.

We now describe the first change in more detail now. Again, we start by considering an upper bound on the SKL between Ising models p and q.

$$d_{\mathrm{SKL}}(p, q) = \sum_{v \in V} \left(\theta^p_v - \theta^q_v\right)\left(\mu^p_v - \mu^q_v\right) + \sum_{(u,v) \in E} \left(\theta^p_{uv} - \theta^q_{uv}\right)\left(\mu^p_{uv} - \mu^q_{uv}\right)$$

$$\implies d_{\mathrm{SKL}}(p, q) \leq 2h \sum_{v \in V} \left|\left(\mu^p_v - \mu^q_v\right)\right| + 2\beta \sum_{u \neq v} \left|\left(\mu^p_{uv} - \mu^q_{uv}\right)\right|$$

Hence if $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$, then either

- $2h \sum_{v \in V} |(\mu_v^p - \mu_v^q)| \geq \varepsilon/2$ or

- $2\beta \sum_{u \neq v} |(\mu_{uv}^p - \mu_{uv}^q)| \geq \varepsilon/2$.

Moreover, if $p = q$, then both $2h \sum_{v \in V} |(\mu_v^p - \mu_v^q)| = 0$ and $2\beta \sum_{u \neq v} |(\mu_{uv}^p - \mu_{uv}^q)| = 0$. Our tester will first test for case (i) and if that test doesn't declare that the two Ising models are far, then proceeds to test whether case (ii) holds.

We will first describe the test to detect whether $\sum_{v \in V} |(\mu_v^p - \mu_v^q)| = 0$ or is $\geq \varepsilon/2h$. We observe that the random variables X_v are Rademachers and hence we can use the weak-learning framework we developed so far to accomplish this goal. The statistic we consider is $Z = \sum_{v \in V} \mathbf{sign}(\mu_v^p)\,(X_v - \mu_v^q)$. Again, as before, we face two challenges: we don't know the signs of the node expectations μ_v^p and we need a bound on $\mathbf{Var}(Z)$.

We employ the weak-learning framework described in Sections 4.5.1-4.5.4 to weakly learn a sign vector correlated with the true sign vector. In particular, since $X_v \sim Rademacher(1/2+\mu_v/2)$, from Corollary 12, we have that with k samples we can correctly estimate $\mathbf{sign}(\mu_v^p - \mu_v^q)$ with probability $1/2 + c_1\sqrt{k}\,|\mu_v^p - \mu_v^q|$. The rest of the argument for obtaining a sign vector which, with sufficient probability, preserves a sufficient amount of signal from the expected value of the statistic, proceeds in a similar way as before. However since the total number of terms we have in our expression is only linear we get some savings in the sample complexity.

And from Lemma 22, we have the following bound on functions $f_c(.)$ of the form $f_c(X) = \sum_{v \in V} c_v X_v$ (where $c \in \{\pm 1\}^V$) on the Ising model:

$$\mathbf{Var}(f_c(X)) = O(n).$$

By performing calculations analogous to the ones in Sections 4.5.3 and 4.5.4, we obtain that by using $\tilde{O}\left(\frac{n^{5/3}h^2}{\varepsilon^2}\right)$ samples we can test whether $\sum_{v \in V} |(\mu_v^p - \mu_v^q)| = 0$ or is $\geq \varepsilon/4h$ with probability $\geq 19/20$. If the tester outputs that $\sum_{v \in V} |(\mu_v^p - \mu_v^q)| = 0$, then we proceed to test whether $\sum_{u \neq v} |(\mu_{uv}^p - \mu_{uv}^q)| = 0$ or $\geq \varepsilon/4\beta$.

To perform this step, we begin by looking at the statistic Z used in Section 4.5.5.2:

$$Z = \sum_{u \neq v} \mathbf{sign}\left(\mu^p_{uv} - \mu^q_{uv}\right)\left(X_u X_v - \mu^q_{uv}\right)$$

as Z has the right expected value. We learn a sign vector which is weakly correlated with the true sign vector. However we need to obtain a variance bound on functions of the form $f_c(X) = \sum_{u \neq v} c_{uv}(X_u X_v - \mu^q_{uv})$ where $c \in \{\pm 1\}^{\binom{V}{2}}$. By ignoring the constant term in $f_c(X)$, we get that,

$$\mathbf{Var}(f_c(X)) = \mathbf{Var}\left(\sum_{u \neq v} c_{uv} X_u X_v\right)$$

which can be $\Omega(n^3)$ as it is not appropriately centered. We employ Lemma 22 to get a variance bound of $O(n^3)$ which yields a sample complexity of $\tilde{O}\left(\frac{n^{11/3}\beta^2}{\varepsilon^2}\right)$ for this setting. Theorem 31 captures the total sample complexity of our identity tester under the presence of external fields.

Theorem 31 (Identity Testing using Learn-Then-Test, Arbitrary External Field)**.** *Suppose p and q are Ising models in the high temperature regime under arbitrary external fields. The learn-then-test algorithm takes in $\tilde{O}\left(\frac{n^{5/3}h^2 + n^{11/3}\beta^2}{\varepsilon^2}\right)$ i.i.d. samples from p and distinguishes between the cases $p = q$ and $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ with probability $\geq 9/10$.*

The tester is formally described in Algorithm 11.

4.6 Localization Versus Learn-then-Test

At this point, we now have two algorithms: the localization algorithm of Section 4.3 and the learn-then-test algorithm of Section 4.5. Both algorithms are applicable in all temperature regimes but learn-then-test beats localization's sample complexity in high temperature under some degree regimes. We note that their sample complexities differ in their dependence on β and $\delta_{\max}$. In this section, we offer some intuition as to why the difference arises and state the best sample complexities we achieve for our testing problems by combining these two approaches.

Algorithm 11 Test if an Ising model p under an external field is identical to Ising model q

function TESTISING(sample access to an Ising model $p, \beta, \delta_{\max}, \varepsilon, \tau_1, \tau_2$, description of Ising model q)
  Run the localization Algorithm 3 on the nodes with accuracy parameter $\frac{\varepsilon}{2n^{\tau_1}}$. If it. identifies any nodes, return that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$
  for $\ell = 1$ to $O(n^{1-\tau_1})$ **do**
    Run the weak learning Algorithm 8 on $S = \{(X_u - Y_u\}_{u \in V}$, where $Y_u \sim$. *Rademacher*$(1/2 + \mu_u^q/2)$, with parameters τ_1 and $\varepsilon/2h$ to generate a sign vector $\vec{\Gamma}^{(\ell)}$ where $\Gamma_u^{(\ell)}$ is weakly correlated with $\mathbf{sign}\left(\mathbf{E}\left[X_u - \mu_u^q\right]\right)$
  end for
  Using the *same set of samples for all* ℓ, run the testing algorithm of Lemma 34. on each of the $\vec{\Gamma}^{(\ell)}$ with parameters $\tau_3 = \tau_1, \delta = O(1/n^{1-\tau_1})$. If any output that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$, return that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$

  Run the localization Algorithm 3 on the edges with accuracy parameter $\frac{\varepsilon}{2n^{\tau_2}}$. If it. identifies any edges, return that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$
  for $\ell = 1$ to $O(n^{2-\tau_2})$ **do**
    Run the weak learning Algorithm 8 on $S = \{(X_u X_v - Y_{uv}\}_{u \neq v}$, where $Y_{uv} \sim$. *Rademacher*$(1/2 + \mu_{uv}^q/2)$, with parameters τ_2 and $\varepsilon/2\beta$ to generate a sign vector $\vec{\Gamma}^{(\ell)}$ where $\Gamma_{uv}^{(\ell)}$ is weakly correlated with $\mathbf{sign}\left(\mathbf{E}\left[X_u X_v - \mu_{uv}^q\right]\right)$
  end for
  Using the *same set of samples for all* ℓ, run the testing algorithm of Lemma 34. on each of the $\vec{\Gamma}^{(\ell)}$ with parameters $\tau_4 = \tau_2, \delta = O(1/n^{2-\tau_2})$. If any output that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$, return that $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$. Otherwise, return that $p = q$
end function

First, we note that if the algorithm is agnostic of the maximum degree $\delta_{\max}$, then learn-then-test always outperforms localization in the high temperature regime. This leads to Theorem 32.

Theorem 32. *Suppose p is an Ising model in the high temperature regime. To test either independence or identity agnostic of the maximum degree of the graph $\delta_{\max}$, localization requires $\tilde{O}\left(\frac{n^4\beta^2}{\varepsilon^2}\right)$ samples from p for a success probability $> 2/3$. Learn-then-test, on the other hand, requires $\tilde{O}\left(\frac{n^{10/3}\beta^2}{\varepsilon^2}\right)$ for independence testing and identity testing under no external field. It requires $\tilde{O}\left(\frac{n^{11/3}\beta^2}{\varepsilon^2}\right)$ for identity testing under an external field.*

When knowledge of $\delta_{\max}$ is available to the tester, he can improve his sample complexities of localization approach. Now the sample complexity of localization gets worse as $\delta_{\max}$ increases. As noted in Section 4.3, the reason for this worsening is that the contribution to the distance by any single edge grows smaller thereby making it harder to detect. However,

when we are in the high-temperature regime a larger $\delta_{\max}$ implies a tighter bound on the strength of the edge interactions β and the variance bound of Section 4.8 exploits this tighter bound to get savings in sample complexities when the degree is large enough.

We combine the sample complexities obtained by the localization and the learn-then-test algorithms and summarize in the following theorems the best sample complexities we can achieve for testing independence and identity by noting the parameter regimes in which of the above two algorithms gives better sample complexity. In both of the following theorems we fix β to be $n^{-\alpha}$ for some α and present which algorithm dominates as $\delta_{\max}$ ranges from a constant to n.

Theorem 33 (Best Sample Complexity Achieved, No External Field)**.** *Suppose p is an Ising model under* ***no external field****.*

- *if $\beta = O(n^{-2/3})$, then for the range $\delta_{\max} \leq n^{2/3}$, localization performs better, for both independence and identity testing. For the range $n^{2/3} \leq \delta_{\max} \leq \frac{1}{4\beta}$, learn-then-test performs better than localization for both independence and identity testing yielding a sample complexity which is independent of $\delta_{\max}$. If $\delta_{\max} \geq \frac{1}{4\beta}$, then we are no longer in the high temperature regime.*

- *if $\beta = \omega(n^{-2/3})$, then for the entire range of $\delta_{\max}$ localization performs at least as well as the learn-then-test algorithm for both independence and identity testing.*

The theorem stated above is summarized in Figure 4-2 for the regime when $\beta = O(n^{-2/3})$.

The comparison for independence testing under the presence of an external field is similar and is presented in Theorem 34.

Theorem 34 (Best Sample Complexity Achieved for Independence Testing, Arbitrary External Field)**.** *Suppose p is an Ising model under* ***an arbitrary external field****.*

- *if $\beta = O(n^{-2/3})$, then for the range $\delta_{\max} \leq n^{2/3}$, localization performs better, for independence testing. For the range $n^{2/3} \leq \delta_{\max} \leq \frac{1}{4\beta}$, learn-then-test performs better than localization for independence testing yielding a sample complexity which is independent of $\delta_{\max}$. If $\delta_{\max} \geq \frac{1}{4\beta}$, then we are no longer in the high temperature regime.*

- *if $\beta = \omega(n^{-2/3})$, then for the entire range of $\delta_{\max}$ localization performs at least as well as the learn-then-test algorithm for independence testing.*

Finally, we note in Theorem 35, the parameter regimes when learn-then-test performs better for identity testing under an external field. Here our learn-then-test approach suffers worse bounds due to a weaker bound on the variance of our statistic.

Theorem 35 (Best Sample Complexity Achieved for Identity Testing, Arbitrary External Field)**.** *Suppose p is an Ising model under* ***an arbitrary external field****.*

- *if $\beta = O(n^{-5/6})$, then for the range $n^{2/3} \leq \delta_{\max} \leq \frac{1}{4\beta}$, learn-then-test performs better than localization for identity testing yielding a sample complexity which is independent of $\delta_{\max}$. If $\delta_{\max} \geq \frac{1}{4\beta}$, then we are no longer in the high temperature regime.*

- *if $\beta = \omega(n^{-5/6})$, then for the entire range of $\delta_{\max}$ localization performs at least as well as the learn-then-test algorithm for identity.*

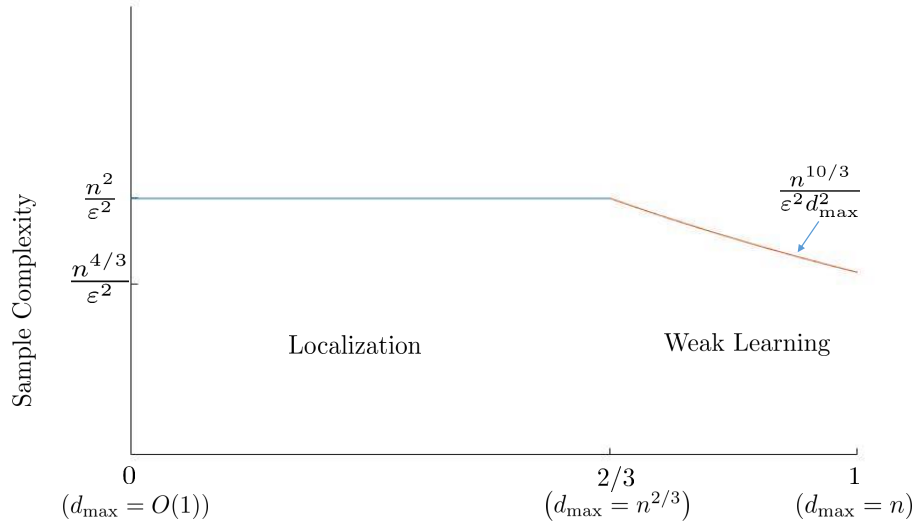


Figure 4-1: Localization versus Learn-Then-Test: A plot of the sample complexity of testing identity under no external field when $\beta = \frac{1}{4\delta_{\max}}$ is close to the threshold of high temperature. Note that throughout the range of values of $\delta_{\max}$ we are in high temperature regime in this plot.

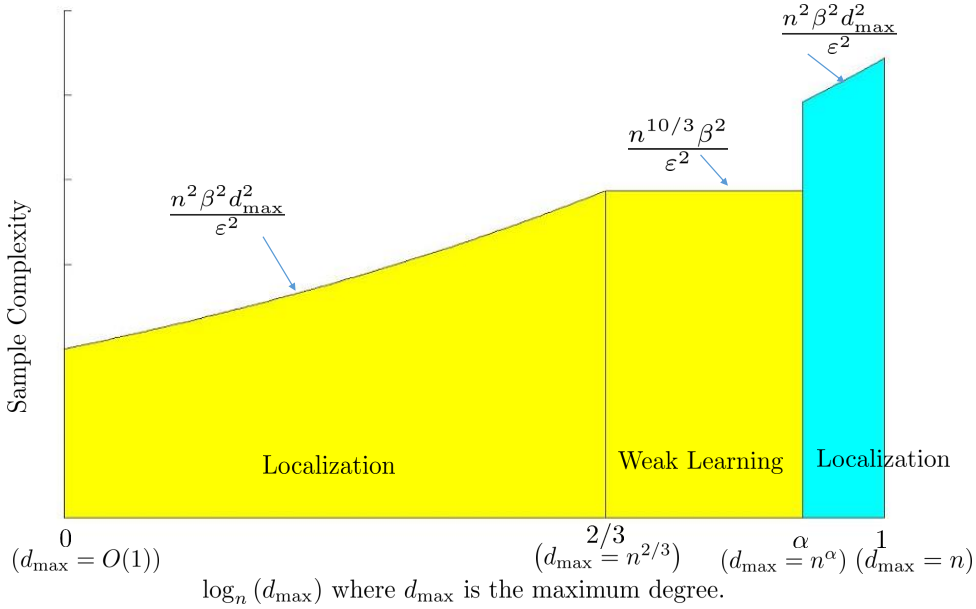


Figure 4-2: Localization versus Learn-Then-Test: A plot of the sample complexity of testing identity under no external field when $\beta \leq n^{-2/3}$. The regions shaded yellow denote the high temperature regime while the region shaded blue denotes the low temperature regime. The algorithm which achieves the better sample complexity is marked on the corresponding region.

4.7 Improved Testing on High-Temperature Ferromagnets

In this section, we present an improved upper bound for testing uniformity of Ising models which are both high-temperature and ferromagnetic. Similar to the algorithms of Section 4.5, we use a *global* statistic, in comparison to the local statistic which is employed for general ferromagnets in Section 4.4.2.

Our result is the following:

Theorem 36 (Independence Testing of High-Temperature Ferromagnetic Ising Models)**.** *Algorithm 12 takes in $\tilde{O}\left(\frac{n}{\varepsilon}\right)$ samples from a high-temperature ferromagnetic Ising model $X \sim p$ which is under no external field and outputs whether $p \in \mathcal{I}$ or $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$ with probability $\geq 9/10$.*

We note that a qualitatively similar result was previously shown in [GLP17], using a χ^2-style statistic. Our algorithm is extremely similar to our test for general high-temperature Ising models. The additional observation is that, since the model is ferromagnetic, we know that all edge marginals have non-negative expectation, and thus we can skip the "weak learning" stage by simply examining the global statistic with the all-ones coefficient vector. The test is described precisely in Algorithm 12.

Algorithm 12 Test if a high-temperature ferromagnetic Ising model p under no external field is product

function TestHighTemperatureFerroIsing-Independence(sample access to an Ising model p)
  Run the algorithm of Lemma 25 to identify if all edges $e = (u, v)$ such that $\mathbf{E}[X_u X_v] \geq \sqrt{\varepsilon/n}$ using $\tilde{O}\left(\frac{n}{\varepsilon}\right)$ samples. If it identifies any edges, return that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$.
  Draw $k = \tilde{O}\left(\frac{n}{\varepsilon}\right)$ samples from p, denote them by $X^{(1)}, \ldots, X^{(k)}$.
  Compute the statistic $Z = \frac{1}{k} \sum_{i=1}^{k} \sum_{(u,v) \in E} X_u^{(i)} X_v^{(i)}$.
  If $Z \geq \frac{1}{4}\sqrt{\varepsilon n}$, return that $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$.
  Otherwise, return that p is product.
end function

Proof of Theorem 36: First, note that under no external field, the only product Ising model is the uniform distribution $\mathcal{U}_n$, and the problem reduces to testing whether p is uniform or

not. Consider first the filtering in Step 2. By the correctness of Lemma 25, this will not wrongfully reject any uniform Ising models. Furthermore, for the remainder of the algorithm, we have that $\mathbf{E}[X_u X_v] \leq \sqrt{\varepsilon/n}$.

Now, we consider the statistic Z. By Theorem 37, we know that the variance of Z is at most $\tilde{O}\left(n^2/k\right)$ (since we are in high-temperature). It remains to consider the expectation of the statistic. When p is indeed uniform, it is clear that $\mathbf{E}[Z] = 0$. When $d_{\mathrm{SKL}}(p, \mathcal{U}_n) \geq \varepsilon$, we have that

$$\begin{aligned}
\varepsilon &\leq \sum_{(u,v)\in E} \theta_{uv} \mathbf{E}[X_u X_v] && (4.37)\\
&\leq \sum_{(u,v)\in E} \tanh^{-1}(\mathbf{E}[X_u X_v]) \mathbf{E}[X_u X_v] && (4.38)\\
&\leq \sum_{(u,v)\in E} 2\mathbf{E}[X_u X_v]^2 && (4.39)\\
&\leq 2\sqrt{\frac{\varepsilon}{n}} \sum_{(u,v)\in E} \mathbf{E}[X_u X_v] && (4.40)
\end{aligned}$$

(4.37) follows by (4.2), (4.38) is due to Lemma 31 (since the model is ferromagnetic), (4.39) is because $\tanh^{-1}(x) \leq 2x$ for $x \leq 0.95$, and (4.40) is since after Step 2, we know that $\mathbf{E}[X_u X_v] \leq \sqrt{\varepsilon/n}$. This implies that $\mathbf{E}[Z] \geq \sqrt{\varepsilon n/4}$.

At this point, we have that $\mathbf{E}[Z] = 0$ when p is uniform, and $\mathbf{E}[Z] \geq \sqrt{\varepsilon n/4}$ when $d_{\mathrm{SKL}}(p, \mathcal{U}_n) \geq \varepsilon$. Since the standard deviation of Z is $\tilde{O}\left(n/\sqrt{k}\right)$, by Chebyshev's inequality, choosing $k = \tilde{\Omega}(n/\varepsilon)$ suffices to distinguish the two cases with probability $\geq 9/10$. □

4.8 Variance Bounds in High-Temperature

In this section, we describe a technique for bounding the variance of our statistics on the Ising model in high temperature. As the structure of Ising models can be quite complex, it can be challenging to obtain non-trivial bounds on the variance of even relatively simple statistics. In particular, to apply our learn-then-test framework of Section 4.5, we must bound the variance of statistics of the form $Z' = \sum_{u \neq v} c_{uv} X_u X_v$ (under no external field, see (4.33)) and $Z'_{cen} = \sum_{u \neq v} c_{uv} \left(X_u^{(1)} - X_u^{(2)}\right)\left(X_v^{(1)} - X_v^{(2)}\right)$ (under an external field, see

(4.36)). While the variance for both the statistics is easily seen to be $O(n^2)$ if the graph has no edges, to prove variance bounds better than the trivial $O(n^4)$ for general graphs requires some work. We show the following two theorems in this section.

The first result, Theorem 37, bounds the variance of functions of the form $\sum_{u \neq v} c_{uv} X_u X_v$ under no external field which captures the statistic used for testing independence and identity by the learn-then-test framework of Section 4.5 in the absence of an external field.

Theorem 37 (High Temperature Variance Bound, No External Field)**.** *Let $c \in [-1,1]^{\binom{V}{2}}$ and define $f_c : \{\pm 1\}^V \to \mathbb{R}$ as follows: $f_c(x) = \sum_{i \neq j} c_{\{i,j\}} x_i x_j$. Let also X be distributed according to an Ising model, without node potentials (i.e. $\theta_v = 0$, for all v), in the high temperature regime of Definition 12. Then*

$$\mathbf{Var}\left(f_c(X)\right) = \tilde{O}(n^2).$$

The second result of this section, Theorem 38, bounds the variance of functions of the form $\sum_{u \neq v} c_{uv}(X_u^{(1)} - X_u^{(2)})(X_v^{(1)} - X_v^{(2)})$ which captures the statistic of interest for independence testing using the learn-then-test framework of Section 4.5 under an external field. Intuitively, this modification is required to "recenter" the random variables. Here, we view the two samples from Ising model p over graph $G = (V, E)$ as coming from a single Ising model $p^{\otimes 2}$ over a graph $G^{(1)} \cup G^{(2)}$ where $G^{(1)}$ and $G^{(2)}$ are identical copies of G.

Theorem 38 (High Temperature Variance Bound, Arbitrary External Field)**.** *Let $c \in [-1,1]^{\binom{V}{2}}$ and let X be distributed according to Ising model $p^{\otimes 2}$ over graph $G^{(1)} \cup G^{(2)}$ in the high temperature regime of Definition 12 and define $g_c : \{\pm 1\}^{V \cup V'} \to \mathbb{R}$ as follows: $g_c(x) = \sum_{\substack{u,v \in V \\ s.t.\ u \neq v}} c_{uv}(x_{u^{(1)}} - x_{u^{(2)}})(x_{v^{(1)}} - x_{v^{(2)}})$. Then*

$$\mathbf{Var}(g_c(X)) = \tilde{O}\left(n^2\right).$$

4.8.1 Overview

We will use tools from Chapter 13 of [LPW09] to obtain the variance bounds of this section. At a high level the technique to bound the variance of a function f on a distribution μ

involves first defining a reversible Markov chain with μ as its stationary distribution. By studying the mixing time properties (via the spectral gap) of this Markov chain along with the second moment of the variation of f when a single step is taken under this Markov chain we obtain bounds on the second moment of f which consequently yield the desired variance bounds.

The Markov chain in consideration here will be the Glauber dynamics chain on the Ising model p. As stated in Section 4.2, the Glauber dynamics are reversible and ergodic for Ising models. Let M be the reversible transition matrix for the Glauber dynamics on some Ising model p. Let γ_* be the absolute spectral gap for this Markov chain. The first step is to obtain a lower bound on γ_*.

Claim 5. *In the high-temperature regime/under Dobrushin conditions,* $\gamma_* \geq \Omega\left(\frac{1}{n \log n}\right)$ *under an arbitrary external field.*

Proof. From Theorem 15.1 of [LPW09], we have that the mixing time of the Glauber dynamics is $O(n \log n)$. Since the Glauber dynamics on an Ising model is ergodic and reversible, using the relation between mixing and relaxation times (Theorem 12.4 of [LPW09]) we get that

$$t_{mix} \geq \left(\frac{1}{\gamma_*} - 1\right) \log(2) \tag{4.41}$$

$$\implies \frac{1}{\gamma_*} \leq \frac{n \log n}{\log(2)} + 1 \tag{4.42}$$

$$\implies \gamma_* \geq \Omega\left(\frac{1}{n \log n}\right). \tag{4.43}$$

□

For a function f, define

$$\mathcal{E}(f) = \frac{1}{2} \sum_{x,y \in \{\pm 1\}^n} [f(x) - f(y)]^2 \pi(x) M(x,y).$$

This can be interpreted as the expected square of the difference in the function, when a step

is taken at stationarity. That is,

$$\mathcal{E}(f) = \frac{1}{2}\mathbf{E}\left[(f(x) - f(y))^2\right] \tag{4.44}$$

where x is drawn from the Ising distribution and y is obtained by taking a step in the Glauber dynamics starting from x. We now state a slight variant of Remark 13.13 which we will use as Lemma 35.

Lemma 35. *For a reversible transition matrix P on state space Ω with stationary distribution π, let*

$$\mathcal{E}(f) := \frac{1}{2}\sum_{x,y\in\Omega}(f(x) - f(y))^2\pi(x)P(x,y),$$

where f is a function on Ω such that $\mathbf{Var}_\pi(f) > 0$. Also let γ_* be the absolute spectral gap of P. Then

$$\gamma_* \leq \frac{\mathcal{E}(f)}{\mathbf{Var}_\pi(f)}.$$

Note: Remark 13.13 in [LPW09] states a bound on the spectral gap as opposed to the absolute spectral gap bound which we use here. However, the proof of Remark 13.13 also works for obtaining a bound on the absolute spectral gap γ_*.

4.8.2 No External Field

We prove Theorem 37 now. Consider the function $f_c(x) = \sum_{u,v} c_{uv}x_u x_v$ where $c \in [-1,1]^{\binom{|V|}{2}}$.

Claim 6. *For an Ising model under no external field, $\mathcal{E}(f_c) = \widetilde{O}(n)$.*

Proof. Since y is obtained by taking a single step on the Glauber dynamics from x, $f_c(x) - f_c(y)$ is a function of the form $\sum_v c_v x_v$ where $c_v \in [-1,1]$ for all $v \in V$. The coefficients $\{c_v\}_v$ depend on which node $v_0 \in V$ was updated by the Glauber dynamics. Since there are n choices for nodes to update, and since the update might also leave x unchanged, i.e. $y = x$, $f_c(x) - f_c(y)$ is one of at most $n+1$ linear functions of the form $\sum_v c_v x_v$. Denote, by $E(x,y)$, the event that $|f_c(x) - f_c(y)| \geq c\sqrt{n}\log n$. We have, from the concentration of linear functions on the Ising model around their expected value (Lemma 22) and a union

bound over the $n+1$ possible linear functions, that for a sufficiently large c, under no external field, $\Pr[E(x,y)] \leq \frac{1}{10n^2}$. Now,

$$\begin{aligned}\mathbf{E}\left[(f_c(x) - f_c(y))^2\right] &= \mathbf{E}\left[(f_c(x) - f_c(y))^2 | E(x,y)\right] \Pr[E(x,y)] \\ &+ \mathbf{E}\left[(f_c(x) - f_c(y))^2 | \neg E(x,y)\right] \Pr[\neg E(x,y)] \\ &\leq n^2 \times \tfrac{1}{10n^2} + c^2 n \log^2 n \left(1 - \tfrac{1}{10n^2}\right) \\ &= \widetilde{O}(n)\end{aligned}$$

where we used the fact that the absolute maximum value of $(f_c(x) - f_c(y))^2$ is n^2. □

Claim 5 together with Claim 6 are sufficient to conclude an upper bound on the variance of f_c, by using Lemma 35, thus giving us Theorem 37.

4.8.3 Arbitrary External Field

Under the presence of an external field, we saw that we need to appropriately center our statistics to achieve low variance. The function $g_c(x)$ of interest now is defined over the 2-sample Ising model $p^{\otimes 2}$ and is of the form

$$g_c(x) = \sum_{u,v} c_{uv}(x_u^{(1)} - x_u^{(2)})(x_v^{(1)} - x_v^{(2)})$$

where now $x, y \in \{\pm 1\}^{2|V|}$. First, note that the absolute spectral gap for $p^{\otimes 2}$ is also at least $\widetilde{\Omega}(1/n)$. Now we bound $\mathcal{E}(g_c)$.

Claim 7. *For an Ising model under an arbitrary external field, $\mathcal{E}(g_c) = \widetilde{O}(n)$.*

Proof. Since y is obtained by taking a single step on the Glauber dynamics from x, it can be seen that $g_c(x) - g_c(y)$ is a function of the form $\sum_v c_v \left(x_v^{(1)} - x_v^{(2)}\right)$ where $c_v \in [-1, 1]$ for all $v \in V$. The coefficients $\{c_v\}_v$ depend on which node $v_0 \in V$ was updated by the Glauber dynamics. Since there are n choices for nodes to update, and since the update might also leave x unchanged, i.e. $y = x$, $g_c(x) - g_c(y)$ is one of at most $n + 1$ linear functions of the form $\sum_v c_v \left(x_v^{(1)} - x_v^{(2)}\right)$. Denote, by $E(x,y)$, the event that $|g_c(x) - g_c(y)| \geq c\sqrt{n} \log n$. We

have, from Lemma 22 and a union bound, that for a sufficiently large c, $\Pr[E(x,y)] \leq \frac{1}{10n^2}$. Now,

$$\begin{aligned}
\mathbf{E}\left[(g_c(x) - g_c(y))^2\right] &= \mathbf{E}\left[(g_c(x) - g_c(y))^2 | E(x,y)\right] \Pr[E(x,y)] && (4.45)\\
&+ \mathbf{E}\left[(g_c(x) - g_c(y))^2 | E(x,y)^c\right] \Pr[E(x,y)^c] && (4.46)\\
&\leq 4n^2 \times \tfrac{1}{10n^2} + c^2 n \log^2 n \left(1 - \tfrac{1}{10n^2}\right) && (4.47)\\
&= \widetilde{O}(n) && (4.48)
\end{aligned}$$

where we used the fact that the absolute maximum value of $(g_c(x) - g_c(y))^2$ is $4n^2$. □

Similar to before, Claim 5 together with Claim 7 are sufficient to conclude an upper bound on the variance of f_c, by using Lemma 35, thus giving us Theorem 38.

4.9 Lower Bounds for Testing Ising Models

In this section we describe our lower bound constructions and state the main results.

4.9.1 Dependences on n

Our first lower bounds show dependences on n, the number of nodes, in the complexity of testing Ising models.

To start, we prove that uniformity testing on product measures over a binary alphabet requires $\Omega(\sqrt{n}/\varepsilon)$ samples. Note that a binary product measure corresponds to the case of an Ising model with no edges. This implies the same lower bound for identity testing, but (not) independence testing, as a product measure always has independent marginals, so the answer is trivial.

Theorem 39. *There exists a constant $c > 0$ such that any algorithm, given sample access to an Ising model p with no edges (i.e., a product measure over a binary alphabet), which distinguishes between the cases $p = \mathcal{U}_n$ and $d_{\mathrm{SKL}}(p, \mathcal{U}_n) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq c\sqrt{n}/\varepsilon$ samples.*

Next, we show that any algorithm which tests uniformity of an Ising model requires $\Omega(n/\varepsilon)$ samples. In this case, it implies the same lower bounds for independence and identity testing.

Theorem 40. *There exists a constant $c > 0$ such that any algorithm, given sample access to an Ising model p, which distinguishes between the cases $p = \mathcal{U}_n$ and $d_{\mathrm{SKL}}(p, \mathcal{U}_n) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq cn/\varepsilon$ samples. This remains the case even if p is known to have a tree structure and only ferromagnetic edges.*

The lower bounds use Le Cam's two point method which constructs a family of distributions $\mathcal{P}$ such that the distance between any $P \in \mathcal{P}$ and a particular distribution Q is large (at least ε). But given a $P \in \mathcal{P}$ chosen uniformly at random, it is hard to distinguish between P and Q with at least $2/3$ success probability unless we have sufficiently many samples.

Our construction for product measures is inspired by Paninski's lower bound for uniformity testing [Pan08]. We start with the uniform Ising model and perturb each node positively or negatively by $\sqrt{\varepsilon/n}$, resulting in a model which is ε-far in d_{SKL} from $\mathcal{U}_n$. The proof appears in Section 4.9.3.1.

Our construction for the linear lower bound builds upon this style of perturbation. In the previous construction, instead of perturbing the node potentials, we could have left the node marginals to be uniform and perturbed the edges of some fixed, known matching to obtain the same lower bound. To get a linear lower bound, we instead choose a *random* perfect matching, which turns out to require quadratically more samples to test. Interestingly, we only need ferromagnetic edges (i.e., positive perturbations), as the randomness in the choice of matching is sufficient to make the problem harder. Our proof is significantly more complicated for this case, and it uses a careful combinatorial analysis involving graphs which are unions of two perfect matchings. The lower bound is described in detail in Section 4.9.3.2.

Remark 4. *Similar lower bound constructions to those of Theorems 39 and 40 also yield $\Omega(\sqrt{n}/\varepsilon^2)$ and $\Omega(n/\varepsilon^2)$ for the corresponding testing problems when d_{SKL} is replaced with d_{TV}. In our constructions, we describe families of distributions which are ε-far in d_{SKL}. This is done by perturbing certain parameters by a magnitude of $\Theta(\sqrt{\varepsilon/n})$. We can instead describe families of distributions which are ε-far in d_{TV} by performing perturbations of $\Theta(\varepsilon/\sqrt{n})$, and*

the rest of the proofs follow similarly.

4.9.2 Dependences on h, β

Finally, we show that dependences on the h and β parameters are, in general, necessary for independence and identity testing. Recall that h and β are upper bounds on the absolute values of the node and edge parameters, respectively. Our constructions are fairly simple, involving just one or two nodes, and the results are stated in Theorem 41.

Theorem 41. *There is a linear lower bound on the parameters h and β for testing problems on Ising models. More specifically,*

- *There exists a constant $c > 0$ such that, for all $\varepsilon < 1$ and $\beta \geq 0$, any algorithm, given sample access to an Ising model p, which distinguishes between the cases $p \in \mathcal{I}$ and $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq c\beta/\varepsilon$ samples.*

- *There exists constants $c_1, c_2 > 0$ such that, for all $\varepsilon < 1$ and $\beta \geq c_1 \log(1/\varepsilon)$, any algorithm, given a description of an Ising model q with no external field (i.e., $h = 0$) and has sample access to an Ising model p, and which distinguishes between the cases $p = q$ and $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq c_2\beta/\varepsilon$ samples.*

- *There exists constants $c_1, c_2 > 0$ such that, for all $\varepsilon < 1$ and $h \geq c_1 \log(1/\varepsilon)$, any algorithm, given a description of an Ising model q with no edge potentials(i.e., $\beta = 0$) and has sample access to an Ising model p, and which distinguishes between the cases $p = q$ and $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq c_2 h/\varepsilon$ samples.*

The construction and analysis appears in Section 4.9.3.3.

This lower bound shows that the dependence on β parameters by our algorithms cannot be avoided in general, though it may be sidestepped in certain cases. Notably, we show that testing independence of a forest-structured Ising model under no external field can be done using $\tilde{O}\left(\frac{n}{\varepsilon}\right)$ samples (Theorem 23).

4.9.3 Proofs

4.9.3.1 Proof of Theorem 39

This proof will follow via an application of Le Cam's two-point method. More specifically, we will consider two classes of distributions $\mathcal{P}$ and $\mathcal{Q}$ such that:

1. $\mathcal{P}$ consists of a single distribution $p \triangleq \mathcal{U}_n$;
2. $\mathcal{Q}$ consists of a family of distributions such that for all distributions $q \in \mathcal{Q}$, $d_{\text{SKL}}(p, q) \geq \varepsilon$;
3. There exists some constant $c > 0$ such that any algorithm which distinguishes p from a uniformly random distribution $q \in \mathcal{Q}$ with probability $\geq 2/3$ requires $\geq c\sqrt{n}/\varepsilon$ samples.

The third point will be proven by showing that, with $k < c\sqrt{n}/\varepsilon$ samples, the following two processes have miniscule total variation distance, and thus no algorithm can distinguish them:

- The process $p^{\otimes k}$, which draws k samples from p;
- The process $\bar{q}^{\otimes k}$, which selects q from $\mathcal{Q}$ uniformly at random, and then draws k samples from q.

We will let $p_i^{\otimes k}$ be the process $p^{\otimes k}$ restricted to the ith coordinate of the random vectors sampled, and $\bar{q}_i^{\otimes k}$ is defined similarly.

We proceed with a description of our construction. Let $\delta = \sqrt{3\varepsilon/2n}$. As mentioned before, $\mathcal{P}$ consists of the single distribution $p \triangleq \mathcal{U}_n$, the Ising model on n nodes with 0 potentials on every node and edge. Let $\mathcal{M}$ be the set of all 2^n vectors in the set $\{\pm\delta\}^n$. For each $M = (M_1, \ldots, M_n) \in \mathcal{M}$, we define a corresponding $q_M \in \mathcal{Q}$ where the node potential M_i is placed on node i.

Proposition 11. *For each $q \in \mathcal{Q}$, $d_{\text{SKL}}(q, \mathcal{U}_n) \geq \varepsilon$.*

Proof. Recall that

$$d_{\text{SKL}}(q, \mathcal{U}_n) = \sum_{v \in V} \delta \tanh(\delta).$$

Note that $\tanh(\delta) \geq 2\delta/3$ for all $\delta \leq 1$, which can be shown using a Taylor expansion. Therefore

$$d_{\mathrm{SKL}}(q, \mathcal{U}_n) \geq n \cdot \delta \cdot 2\delta/3 = 2n\delta^2/3 = \varepsilon.$$

□

The goal is to upper bound $d_{\mathrm{TV}}(p^{\otimes k}, \bar{q}^{\otimes k})$. Our approach will start with manipulations similar to the following lemma from [AD15].

Lemma 36. *For any two distributions p and q,*

$$2d_{\mathrm{TV}}^2(p, q) \leq d_{\mathrm{KL}}(q, p) \leq \log \mathbf{E}_q \left[\frac{q}{p}\right].$$

The first inequality is Pinsker's, and the second is Jensen's.

Similarly:

$$2d_{\mathrm{TV}}^2(p^{\otimes k}, \bar{q}^{\otimes k}) \leq d_{\mathrm{KL}}(\bar{q}^{\otimes k}, p^{\otimes k}) = n d_{\mathrm{KL}}(\bar{q}_1^{\otimes k}, p_1^{\otimes k}) \leq n \log \mathbf{E}_{\bar{q}_1^{\otimes k}} \left[\frac{\bar{q}_1^{\otimes k}}{p_1^{\otimes k}}\right].$$

The first inequality is Pinsker's, and the last inequality is Jensen's. The equality in the middle is the chain rule for KL divergence – this is because $p_i^{\otimes k}$ and $\bar{q}_i^{\otimes k}$ are independent over coordinates.

We proceed to bound the right-hand side. To simplify notation, let $p_+ = e^{\delta}/(e^{\delta} + e^{-\delta})$ be the probability that a node with parameter δ takes the value 1. Note that a node with parameter $-\delta$ takes the value 1 with probability $1 - p_+$. We will perform a sum over all realizations k_1 for the number of times that node 1 is observed to be 1.

$$
\begin{aligned}
\mathbf{E}_{\bar{q}_1^{\otimes k}}\left[\frac{\bar{q}_1^{\otimes k}}{p_1^{\otimes k}}\right] &= \sum_{k_1=0}^{k} \frac{(\bar{q}_1^{\otimes k}(k_1))^2}{p_1^{\otimes k}(k_1)} \\
&= \sum_{k_1=0}^{k} \frac{\left(\frac{1}{2}\binom{k}{k_1}(p_+)^{k_1}(1-p_+)^{k-k_1} + \frac{1}{2}\binom{k}{k-k_1}(p_+)^{k_1}(1-p_+)^{k_1}\right)^2}{\binom{k}{k_1}(1/2)^k} \\
&= \frac{2^k}{4}\sum_{k_1=0}^{k}\binom{k}{k_1}\left((p_+)^{2k_1}(1-p_+)^{2(k-k_1)} + (p_+)^{2(k-k_1)}(1-p_+)^{2k_1} + 2(p_+(1-p_+))^k\right) \\
&= \frac{2^k}{2}(p_+(1-p_+))^k \sum_{k_1=0}^{k}\binom{k}{k_1} + 2\cdot\frac{2^k}{4}\sum_{k_1=0}^{k}\left(\binom{k}{k_1}(p_+^2)^{k_1}((1-p_+)^2)^{k-k_1}\right)
\end{aligned}
$$

where the second equality uses the fact that $\bar{q}_1^{\otimes k}$ chooses the Ising model with parameter on node 1 being δ and $-\delta$ each with probability 1/2. Using the identity $\sum_{k_1=0}^{k}\binom{k}{k_1}a^{k_1}b^{k-k_1} = (a+b)^k$ gives that

$$
\mathbf{E}_{\bar{q}_1^{\otimes k}}\left[\frac{\bar{q}_1^{\otimes k}}{p_1^{\otimes k}}\right] = \frac{4^k}{2}(p_+(1-p_+))^k + \frac{2^k}{2}\left(2p_+^2 + 1 - 2p_+\right)^k.
$$

Substituting in the value for p_+ and applying hyperbolic trigenometric identities, the above expression simplifies to

$$
\begin{aligned}
&\frac{1}{2}\left(\left(\operatorname{sech}^2(\delta)\right)^k + \left(1+\tanh^2(\delta)\right)^k\right) \\
\leq\ & 1 + \binom{k}{2}\delta^4 \\
=\ & 1 + \binom{k}{2}\frac{9\varepsilon^2}{4n^2}
\end{aligned}
$$

where the inequality follows by a Taylor expansion.

This gives us that

$$
2d_{\mathrm{TV}}^2(p^{\otimes k}, \bar{q}^{\otimes k}) \leq n\log\left(1 + \binom{k}{2}\frac{9\varepsilon^2}{4n^2}\right) \leq \frac{9k^2\varepsilon^2}{4n}.
$$

If $k < 0.9\cdot\sqrt{n}/\varepsilon$, then $d_{\mathrm{TV}}^2(p^{\otimes k}, \bar{q}^{\otimes k}) < 49/50$ and thus no algorithm can distinguish between the two with probability $\geq 99/100$. This completes the proof of Theorem 39.

4.9.3.2 Proof of Theorem 40

This lower bound similarly applies Le Cam's two-point method, as described in the previous section. We proceed with a description of our construction. Assume that n is even. As before, $\mathcal{P}$ consists of the single distribution $p \triangleq \mathcal{U}_n$, the Ising model on n nodes with 0 potentials on every node and edge. Let $\mathcal{M}$ denote the set of all $(n-1)!!$ perfect matchings on the clique on n nodes. Each $M \in \mathcal{M}$ defines a corresponding $q_M \in \mathcal{Q}$, where the potential $\delta = \sqrt{3\varepsilon/n}$ is placed on each edge present in the graph.

The following proposition follows similarly to Proposition 11.

Proposition 12. *For each $q \in \mathcal{Q}$, $d_{\mathrm{SKL}}(q, \mathcal{U}_n) \geq \varepsilon$.*

The goal is to upper bound $d_{\mathrm{TV}}(p^{\otimes k}, \bar{q}^{\otimes k})$. We apply Lemma 36 to $2d_{\mathrm{TV}}^2(p^{\otimes k}, \bar{q}^{\otimes k})$ and focus on the quantity inside the logarithm. Let $X^{(i)} \in \{\pm 1\}^n$ represent the realization of the ith sample and $X_u \in \{\pm 1\}^k$ represent the realization of the k samples on node u. Let $H(.,.)$ represent the Hamming distance between two vectors, and for sets S_1 and S_2, let $S = S_1 \uplus S_2$ be the very commonly used multiset addition operation (i.e., combine all the elements from S_1 and S_2, keeping duplicates). Let M_0 be the perfect matching with edges $(2i-1, 2i)$ for all $i \in [n/2]$.

$$\begin{aligned}\mathbf{E}_{\bar{q}^{\otimes k}}\left[\frac{\bar{q}^{\otimes k}}{p^{\otimes k}}\right] &= \sum_{X=(X^{(1)},\ldots,X^{(k)})} \frac{(\bar{q}^{\otimes k}(X))^2}{p^{\otimes k}(X)} \\ &= 2^{nk} \sum_{X=(X^{(1)},\ldots,X^{(k)})} (\bar{q}^{\otimes k}(X))^2\end{aligned}$$

We can expand the inner probability as follows. Given a randomly selected perfect matching, we can break the probability of a realization X into a product over the edges. By examining the PMF of the Ising model, if the two endpoints of a given edge agree, the probability is multiplied by a factor of $\left(\frac{e^{\delta}}{2(e^{\delta}+e^{-\delta})}\right)$, and if they disagree, a factor of $\left(\frac{e^{-\delta}}{2(e^{\delta}+e^{-\delta})}\right)$. Since (given a matching) the samples are independent, we take the product of this over all k samples. We average this quantity using a uniformly random choice of perfect matching. Writing these

ideas mathematically, the expression above is equal to

$$
\begin{aligned}
&2^{nk} \sum_{X=(X^{(1)},\ldots,X^{(k)})} \left(\frac{1}{(n-1)!!} \sum_{M \in \mathcal{M}} \prod_{(u,v)\in M} \prod_{i=1}^{k} \left(\frac{e^{\delta}}{2(e^{\delta}+e^{-\delta})} \right)^{\mathbb{1}(X_u^{(i)}=X_v^{(i)})} \left(\frac{e^{-\delta}}{2(e^{\delta}+e^{-\delta})} \right)^{\mathbb{1}(X_u^{(i)}\neq X_v^{(i)})} \right)^2 \\
&= 2^{nk} \sum_{X=(X^{(1)},\ldots,X^{(k)})} \left(\frac{1}{(n-1)!!} \sum_{M \in \mathcal{M}} \prod_{(u,v)\in M} \left(\frac{1}{2(e^{\delta}+e^{-\delta})} \right)^{k} e^{\delta(k-H(X_u,X_v))} e^{-\delta H(X_u,X_v)} \right)^2 \\
&= \left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}} \right)^{nk} \sum_{X=(X^{(1)},\ldots,X^{(k)})} \left(\frac{1}{(n-1)!!} \sum_{M \in \mathcal{M}} \prod_{(u,v)\in M} \exp(-2\delta H(X_u,X_v)) \right)^2 \\
&= \left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}} \right)^{nk} \frac{1}{(n-1)!!^2} \sum_{X=(X^{(1)},\ldots,X^{(k)})} \left(\sum_{M \in \mathcal{M}} \prod_{(u,v)\in M} \exp(-2\delta H(X_u,X_v)) \right)^2 \\
&= \left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}} \right)^{nk} \frac{1}{(n-1)!!^2} \sum_{X=(X^{(1)},\ldots,X^{(k)})} \sum_{M_1,M_2 \in \mathcal{M}} \prod_{(u,v)\in M_1 \uplus M_2} \exp(-2\delta H(X_u,X_v))
\end{aligned}
$$

At this point, we note that if we fix the matching M_1, summing over all perfect matchings M_2 gives the same value irrespective of the value of M_1. Therefore, we multiply by a factor of $(n-1)!!$ and fix the choice of M_1 to be M_0.

$$
\begin{aligned}
&\left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}} \right)^{nk} \frac{1}{(n-1)!!} \sum_{M \in \mathcal{M}} \sum_{X=(X^{(1)},\ldots,X^{(k)})} \prod_{(u,v)\in M_0 \uplus M} \exp(-2\delta H(X_u,X_v)) \\
&= \left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}} \right)^{nk} \frac{1}{(n-1)!!} \sum_{M \in \mathcal{M}} \left(\sum_{X^{(1)}} \prod_{(u,v)\in M_0 \uplus M} \exp\left(-2\delta H\left(X_u^{(1)}, X_v^{(1)}\right)\right) \right)^k
\end{aligned}
$$

We observe that multiset union of two perfect matchings will form a collection of even length cycles (if they contain the same edge, this forms a 2-cycle), and this can be rewritten as follows.

$$\left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}}\right)^{nk}\frac{1}{(n-1)!!}\sum_{M\in\mathcal{M}}\left(\sum_{X^{(1)}}\prod_{\substack{\text{cycles}C\\ \in M_0\uplus M}}\prod_{(u,v)\in C}\exp\left(-2\delta H\left(X_u^{(1)},X_v^{(1)}\right)\right)\right)^k$$

$$=\left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}}\right)^{nk}\frac{1}{(n-1)!!}\sum_{M\in\mathcal{M}}\left(\prod_{\substack{\text{cycles }C\\ \in M_0\uplus M}}\sum_{X_C^{(1)}}\prod_{(u,v)\in C}\exp\left(-2\delta H\left(X_u^{(1)},X_v^{(1)}\right)\right)\right)^k \tag{4.49}$$

We now simplify this using a counting argument over the possible realizations of $X^{(1)}$ when restricted to edges in cycle C. Start by noting that

$$\sum_{X_C^{(1)}}\prod_{(u,v)\in C}\left(e^{2\delta}\right)^{-2H\left(X_u^{(1)},X_v^{(1)}\right)}=2\sum_{i=0}^{n/2}\left(\binom{|C|-1}{2i-1}+\binom{|C|-1}{2i}\right)\left(e^{2\delta}\right)^{-2i}.$$

This follows by counting the number of possible ways to achieve a particular Hamming distance over the cycle. The $|C|-1$ (rather than $|C|$) and the grouping of consecutive binomial coefficients arises as we lose one "degree of freedom" due to examining a cycle, which fixes the Hamming distance to be even. Now, we apply Pascal's rule and can see

$$2\sum_{i=0}^{n/2}\left(\binom{|C|-1}{2i-1}+\binom{|C|-1}{2i}\right)\left(e^{2\delta}\right)^{-2i}=2\sum_{i=0}^{n/2}\binom{|C|}{2i}\left(e^{2\delta}\right)^{-2i}.$$

This is twice the sum over the even terms in the binomial expansion of $(1+e^{-2\delta})^{|C|}$. The odd terms may be eliminated by adding $(1-e^{-2\delta})^{|C|}$, and thus (4.49) is equal to the following.

$$\left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}}\right)^{nk}\frac{1}{(n-1)!!}\sum_{M\in\mathcal{M}}\left(\prod_{\substack{\text{cycles } C\\ \in M_0\uplus M}}(1+e^{-2\delta})^{|C|}+(1-e^{-2\delta})^{|C|}\right)^{k}$$

$$=\left(\frac{e^{\delta}}{e^{\delta}+e^{-\delta}}\right)^{nk}\frac{1}{(n-1)!!}\sum_{M\in\mathcal{M}}\left(\prod_{\substack{\text{cycles } C\\ \in M_0\uplus M}}\left(\frac{e^{\delta}+e^{-\delta}}{e^{\delta}}\right)^{|C|}\left(1+\left(\frac{e^{\delta}-e^{-\delta}}{e^{\delta}+e^{-\delta}}\right)^{|C|}\right)\right)^{k}$$

$$=\mathbf{E}\left[\left(\prod_{\substack{\text{cycles } C\\ \in M_0\uplus M}}\left(1+\tanh^{|C|}(\delta)\right)\right)^{k}\right] \tag{4.50}$$

where the expectation is from choosing a uniformly random perfect matching $M\in\mathcal{M}$. At this point, it remains only to bound Equation (4.50). Noting that for all $x>0$ and $t\geq 1$,

$$1+\tanh^{t}(\delta)\leq 1+\delta^{t}\leq\exp\left(\delta^{t}\right),$$

we can bound (4.50) as

$$\mathbf{E}\left[\left(\prod_{\substack{\text{cycles } C\\ \in M_0\uplus M}}\left(1+\tanh^{|C|}(\delta)\right)\right)^{k}\right]\leq\mathbf{E}\left[\left(\prod_{\substack{\text{cycles } C\\ \in M_0\uplus M}}\exp\left(\delta^{|C|}\right)\right)^{k}\right].$$

For our purposes, it turns out that the 2-cycles will be the dominating factor, and we use the following crude upper bound. Let ζ be a random variable representing the number of 2-cycles in $M_0\uplus M$, i.e., the number of edges shared by both perfect matchings.

$$\mathbf{E}\left[\left(\prod_{\substack{\text{cycles } C\\ \in M_0\uplus M}}\exp\left(\delta^{|C|}\right)\right)^{k}\right]=\mathbf{E}\left[\left(\prod_{\substack{\text{cycles } C\\ \in M_0\uplus M\\ |C|\geq 4}}\exp\left(\delta^{|C|}\right)\right)^{k}\exp\left(\delta^{2}\zeta k\right)\right]\leq\exp\left(\delta^{4}\cdot n/4\cdot k\right)\mathbf{E}\left[\exp\left(\delta^{2}\zeta k\right)\right],$$

where in the last inequality, we used the facts that $\delta^{|C|}$ is maximized for $|C|\geq 4$ when $|C|=4$, and that there are at most $n/4$ cycles of length at least 4.

We examine the distribution of ζ. Note that

$$\mathbf{E}[\zeta] = \frac{n}{2} \cdot \frac{1}{n-1} = \frac{n}{2(n-1)}.$$

More generally, for any positive integer $z \leq n/2$,

$$\mathbf{E}[\zeta - (z-1) | \zeta \geq z-1] = \frac{n-2z+2}{2} \cdot \frac{1}{n-2z+1} = \frac{n-2z+2}{2(n-2z+1)}.$$

By Markov's inequality,

$$\Pr[\zeta \geq z | \zeta \geq z-1] = \Pr[\zeta - (z-1) \geq 1 | \zeta \geq z-1] \leq \frac{n-2z+2}{2(n-2z+1)}.$$

Therefore,

$$\Pr[\zeta \geq z] = \prod_{i=1}^{z} \Pr[\zeta \geq i | \zeta \geq i-1] \leq \prod_{i=1}^{z} \frac{n-2i+2}{2(n-2i+1)}.$$

In particular, note that for all $z < n/2$,

$$\Pr[\zeta \geq z] \leq (2/3)^z.$$

We return to considering the expectation above:

$$\begin{aligned}
\mathbf{E}\left[\exp\left(\delta^2 \zeta k\right)\right] &= \sum_{z=0}^{n/2} \Pr[\zeta = z] \exp\left(\delta^2 z k\right) \\
&\leq \sum_{z=0}^{n/2} \Pr[\zeta \geq z] \exp\left(\delta^2 z k\right) \\
&\leq \frac{3}{2} \sum_{z=0}^{n/2} (2/3)^z \exp\left(\delta^2 z k\right) \\
&= \frac{3}{2} \sum_{z=0}^{n/2} \exp\left((\delta^2 k - \log(3/2)) z\right) \\
&\leq \frac{3}{2} \cdot \frac{1}{1 - \exp\left(\delta^2 k - \log(3/2)\right)},
\end{aligned}$$

where the last inequality requires that $\exp\left(\delta^2 k - \log(3/2)\right) < 1$. This is true as long as

$k < \log(3/2)/\delta^2 = \frac{\log(3/2)}{3} \cdot \frac{n}{\varepsilon}$.

Combining Lemma 36 with the above derivation, we have that

$$\begin{aligned} 2d_{\mathrm{TV}}^2(p^{\otimes k}, \bar{q}^{\otimes k}) &\leq \log\left(\exp(\delta^4 nk/4) \cdot \frac{3}{2(1-\exp\left(\delta^2 k - \log(3/2)\right))}\right) \\ &= \delta^4 nk/4 + \log\left(\frac{3}{2(1-\exp\left(\delta^2 k - \log(3/2)\right))}\right) \\ &= \frac{9\varepsilon^2}{4n}k + \log\left(\frac{3}{2(1-\exp\left(3k\varepsilon/n - \log(3/2)\right))}\right). \end{aligned}$$

If $k < \frac{1}{25} \cdot \frac{n}{\varepsilon}$, then $d_{\mathrm{TV}}(p^{\otimes k}, \bar{q}^{\otimes k}) < 49/50$ and thus no algorithm can distinguish between the two cases with probability $\geq 99/100$. This completes the proof of Theorem 40.

4.9.3.3 Proof of Theorem 41

We provide constructions for our lower bounds of Theorem 41 which show that a dependence on β is necessary in certain cases.

Lemma 37. *There exists a constant $c > 0$ such that, for all $\varepsilon < 1$ and $\beta \geq 0$, any algorithm, given sample access to an Ising model p, which distinguishes between the cases $p \in \mathcal{I}$ and $d_{\mathrm{SKL}}(p, \mathcal{I}) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq c\beta/\varepsilon$ samples.*

Proof. Consider the following two models, which share some parameter $\tau > 0$:

1. An Ising model p on two nodes u and v, where $\theta_u^p = \theta_v^p = \tau$ and $\theta_{uv} = 0$.

2. An Ising model q on two nodes u and v, where $\theta_u^q = \theta_v^q = \tau$ and $\theta_{uv} = \beta$.

We note that $\mathbf{E}[X_u^p X_v^p] = \frac{\exp(2\tau+\beta)+\exp(-2\tau+\beta)-\exp(-\beta)}{\exp(2\tau+\beta)+\exp(-2\tau+\beta)+\exp(-\beta)}$ and $\mathbf{E}[X_u^q X_v^q] = \tanh^2(\tau)$. By (4.2), these two models have $d_{\mathrm{SKL}}(p, q) = \beta\left(\mathbf{E}[X_u^p X_v^p] - \mathbf{E}[X_u^q X_v^q]\right)$. For any for any fixed β sufficiently large and $\varepsilon > 0$ sufficiently small, τ can be chosen to make $\mathbf{E}[X_u^p X_v^p] - \mathbf{E}[X_u^q X_v^q] = \frac{\varepsilon}{\beta}$. This is because at $\tau = 0$, this is equal to $\tanh(\beta)$ and for $\tau \to \infty$, this approaches 0, so by continuity, there must be a τ which causes the expression to equal this value. Therefore, the SKL distance between these two models is ε. On the other hand, it is not hard to see that $d_{\mathrm{TV}}(p, q) = \Theta\left(\mathbf{E}[X_u^p X_v^p] - \mathbf{E}[X_u^q X_v^q]\right) = \Theta(\varepsilon/\beta)$, and therefore, to distinguish these models, we require $\Omega(\beta/\varepsilon)$ samples. □

Lemma 38. *There exists constants $c_1, c_2 > 0$ such that, for all $\varepsilon < 1$ and $\beta \geq c_1 \log(1/\varepsilon)$, any algorithm, given a description of an Ising model q with no external field (i.e., $h = 0$) and has sample access to an Ising model p, and which distinguishes between the cases $p = q$ and $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq c_2 \beta/\varepsilon$ samples.*

Proof. This construction is very similar to that of Lemma 37. Consider the following two models, which share some parameter $\tau > 0$:

1. An Ising model p on two nodes u and v, where $\theta^p_{uv} = \beta$.

2. An Ising model q on two nodes u and v, where $\theta^p_{uv} = \beta - \tau$.

We note that $\mathbf{E}[X^p_u X^p_v] = \tanh(\beta)$ and $\mathbf{E}[X^q_u X^q_v] = \tanh(\beta - \tau)$. By (4.2), these two models have $d_{\mathrm{SKL}}(p, q) = \tau \left(\mathbf{E}[X^p_u X^p_v] - \mathbf{E}[X^q_u X^q_v]\right)$. Observe that at $\tau = \beta$, $d_{\mathrm{SKL}}(p, q) = \beta \tanh(\beta)$, and at $\tau = \beta/2$, $d_{\mathrm{SKL}}(p, q) = \frac{\beta}{2}(\tanh(\beta) - \tanh(\beta/2)) = \frac{\beta}{2}(\tanh(\beta/2)\operatorname{sech}(\beta)) \leq \beta \exp(-\beta) \leq \varepsilon$, where the last inequality is based on our condition that β is sufficiently large. By continuity, there exists some $\tau \in [\beta/2, \beta]$ such that $d_{\mathrm{SKL}}(p, q) = \varepsilon$. On the other hand, it is not hard to see that $d_{\mathrm{TV}}(p, q) = \Theta\left(\mathbf{E}[X^p_u X^p_v] - \mathbf{E}[X^q_u X^q_v]\right) = \Theta(\varepsilon/\beta)$, and therefore, to distinguish these models, we require $\Omega(\beta/\varepsilon)$ samples. □

The lower bound construction and analysis for the h lower bound follow almost identically, with the model q consisting of a single node with parameter h.

Lemma 39. *There exists constants $c_1, c_2 > 0$ such that, for all $\varepsilon < 1$ and $h \geq c_1 \log(1/\varepsilon)$, any algorithm, given a description of an Ising model q with no edge potentials(i.e., $\beta = 0$) and has sample access to an Ising model p, and which distinguishes between the cases $p = q$ and $d_{\mathrm{SKL}}(p, q) \geq \varepsilon$ with probability at least $99/100$ requires $k \geq c_2 h/\varepsilon$ samples.*

Together, Lemmas 37, 38, and 39 imply Theorem 41.

4.10 Weak Learning of Rademachers

In this section, we examine the concept of "weakly learning" Rademacher random variables. This problem we study is classical, but our regime of study and goals are slightly different.

Suppose we have k samples from a random variable, promised to either be $Rademacher(1/2+\lambda)$ or $Rademacher(1/2-\lambda)$, for some $0 < \lambda \leq 1/2$. How many samples do we need to tell which case we are in? If we wish to be correct with probability (say) $\geq 2/3$, it is folklore that $k = \Theta(1/\lambda^2)$ samples are both necessary and sufficient. In our weak learning setting, we focus on the regime where we are sample limited (say, when λ is very small), and we are unable to gain a constant benefit over randomly guessing. More precisely, we have a budget of k samples from some $Rademacher(p)$ random variable, and we want to guess whether $p > 1/2$ or $p < 1/2$. The "margin" $\lambda = |p - 1/2|$ may not be precisely known, but we still wish to obtain the maximum possible advantage over randomly guessing, which gives us probability of success equal to $1/2$. We show that with any $k \leq 1/4\lambda^2$ samples, we can obtain success probability $1/2 + \Omega(\lambda\sqrt{k})$. This smoothly interpolates within the "low sample" regime, up to the point where $k = \Theta(1/\lambda^2)$ and folklore results also guarantee a constant probability of success. We note that in this low sample regime, standard concentration bounds like Chebyshev and Chernoff give trivial guarantees, and our techniques require a more careful examination of the Binomial PMF.

We go on to examine the same problem under alternate centerings – where we are trying to determine whether $p > \mu$ or $p < \mu$, generalizing the previous case where $\mu = 1/2$. We provide a simple "recentering" based reduction to the previous case, showing that the same upper bound holds for all values of μ. We note that our reduction holds even when the centering μ is not explicitly known, and we only have limited sample access to $Rademacher(\mu)$.

We start by proving the following lemma, where we wish to determine the direction of bias with respect to a zero-mean Rademacher random variable.

Lemma 40. *Let $X_1, \ldots, X_k$ be i.i.d. random variables, distributed as $Rademacher(p)$ for any $p \in [0, 1]$. There exists an algorithm which takes $X_1, \ldots, X_k$ as input and outputs a value $b \in \{\pm 1\}$, with the following guarantees: there exists constants $c_1, c_2 > 0$ such that for any $p \neq \frac{1}{2}$,*

$$\Pr\left(b = \mathbf{sign}\left(\lambda\right)\right) \geq \begin{cases} \frac{1}{2} + c_1|\lambda|\sqrt{k} & \text{if } k \leq \frac{1}{4\lambda^2} \\ \frac{1}{2} + c_2 & \text{otherwise,} \end{cases}$$

where $\lambda = p - \frac{1}{2}$. If $p = \frac{1}{2}$, then $b \sim Rademacher\left(\frac{1}{2}\right)$.

Proof. The algorithm is as follows: let $S = \sum_{i=1}^{k} X_i$. If $S \neq 0$, then output $b = \mathbf{sign}(S)$, otherwise output $b \sim \mathit{Rademacher}\left(\frac{1}{2}\right)$.

The $p = 1/2$ case is trivial, as the sum S is symmetric about 0. We consider the case where $\lambda > 0$ (the negative case follows by symmetry) and when k is even (odd k can be handled similarly). As the case where $k > \frac{1}{4\lambda^2}$ is well known (see Lemma 23), we focus on the former case, where $\lambda \leq \frac{1}{2\sqrt{k}}$. By rescaling and shifting the variables, this is equivalent to lower bounding $\Pr\left(\mathit{Binomial}\left(k, \frac{1}{2} + \lambda\right) \geq \frac{k}{2}\right)$. By a symmetry argument, this is equal to

$$\frac{1}{2} + d_{\mathrm{TV}}\left(\mathit{Binomial}\left(k, \frac{1}{2} - \lambda\right), \mathit{Binomial}\left(k, \frac{1}{2} + \lambda\right)\right).$$

It remains to show this total variation distance is $\Omega(\lambda\sqrt{k})$.

$$\begin{aligned}
& d_{\mathrm{TV}}\left(\mathit{Binomial}\left(k, \frac{1}{2} - \lambda\right), \mathit{Binomial}\left(k, \frac{1}{2} + \lambda\right)\right) \\
\geq\ & d_{\mathrm{TV}}\left(\mathit{Binomial}\left(k, \frac{1}{2}\right), \mathit{Binomial}\left(k, \frac{1}{2} + \lambda\right)\right) \\
\geq\ & k \min_{\ell \in \{\lceil k/2 \rceil, \ldots, \lceil k/2 + k\lambda \rceil\}} \int_{1/2}^{1/2+\lambda} \Pr\left(\mathit{Binomial}\left(k-1, u\right) = l - 1\right) du && (4.51) \\
\geq\ & \lambda k \cdot \Pr\left(\mathit{Binomial}\left(k-1, 1/2 + \lambda\right) = k/2\right) \\
=\ & \lambda k \cdot \binom{k-1}{k/2} \left(\frac{1}{2} + \lambda\right)^{k/2} \left(\frac{1}{2} - \lambda\right)^{k/2-1} \\
\geq\ & \Omega(\lambda k) \cdot \sqrt{\frac{1}{2k}} \left(1 + \frac{1}{\sqrt{k}}\right)^{k/2} \left(1 - \frac{1}{\sqrt{k}}\right)^{k/2} && (4.52) \\
=\ & \Omega(\lambda\sqrt{k}) \cdot \left(1 - \frac{1}{k}\right)^{k/2} \\
\geq\ & \Omega(\lambda\sqrt{k}) \cdot \exp\left(-1/2\right) \left(1 - \frac{1}{k}\right)^{1/2} && (4.53) \\
=\ & \Omega(\lambda\sqrt{k}),
\end{aligned}$$

as desired.

(4.51) applies Proposition 2.3 of [AJ06]. (4.52) is by an application of Stirling's approximation and since $\lambda \leq \frac{1}{2\sqrt{k}}$. (4.53) is by the inequality $\left(1 - \frac{c}{k}\right)^k \geq \left(1 - \frac{c}{k}\right)^c \exp(-c)$. □

We now develop a corollary allowing us to instead consider comparisons with respect to

different centerings.

Corollary 12. *Let $X_1, \ldots, X_k$ be i.i.d. random variables, distributed as Rademacher(p) for any $p \in [0,1]$. There exists an algorithm which takes $X_1, \ldots, X_k$ and $q \in [0,1]$ as input and outputs a value $b \in \{\pm 1\}$, with the following guarantees: there exists constants $c_1, c_2 > 0$ such that for any $p \neq q$,*

$$\Pr\left(b = \mathbf{sign}\left(\lambda\right)\right) \geq \begin{cases} \frac{1}{2} + c_1|\lambda|\sqrt{k} & \text{if } k \leq \frac{1}{4\lambda^2} \\ \frac{1}{2} + c_2 & \text{otherwise,} \end{cases}$$

where $\lambda = \frac{p-q}{2}$. If $p = q$, then $b \sim$ Rademacher $\left(\frac{1}{2}\right)$.

This algorithm works even if only given k i.i.d. samples $Y_1, \ldots, Y_k \sim$ Rademacher(q), rather than the value of q.

Proof. Let $X \sim$ Rademacher(p) and $Y \sim$ Rademacher(q). Consider the random variable Z defined as follows. First, sample X and Y. If $X \neq Y$, output $\frac{1}{2}(X - Y)$. Otherwise, output a random variable sampled as Rademacher $\left(\frac{1}{2}\right)$. One can see that $Z \sim$ Rademacher $\left(\frac{1}{2} + \frac{p-q}{2}\right)$.

Our algorithm can generate k i.i.d. samples $Z_i \sim$ Rademacher $\left(\frac{1}{2} + \frac{p-q}{2}\right)$ in this method using X_i's and Y_i's, where Y_i's are either provided as input to the algorithm or generated according to Rademacher(q). At this point, we provide the Z_i's as input to the algorithm of Lemma 40. By examining the guarantees of Lemma 40, this implies the desired result. □

4.11 KL Learning of Ising Models: An Attempt

One approach to testing problems is by learning the distribution which we wish to test. If the distance of interest is the total variation distance, then a common approach to learning is a cover-based method. One first creates a set of hypothesis distributions H which $O(\varepsilon)$-covers the space. Then by drawing $k = \tilde{O}(\log|H|/\varepsilon^2)$ samples from p, we can output a distribution from H with the guarantee that it is at most $O(\varepsilon)$-far from p. The algorithm works by computing a score based on the samples for each of the distributions in the hypothesis class and then choosing the one with the maximum score.

However, it is not clear if this approach would work for testing in KL-divergence (an easier problem than testing in SKL-divergence) because KL-divergence does not satisfy the triangle inequality. In particular, if p and q are far, and we learn a distribution $\hat{p}$ which is close to p, we no longer have the guarantee that $\hat{p}$ and q are still far. Even if this issue were somehow resolved, the best known sample complexity for learning follows from the maximum likelihood algorithm. We state the guarantees provided by Theorem 17 of [FOS08].

Theorem 42 (Theorem 17 from [FOS08])**.** *Let $b, a, \varepsilon > 0$ such that $a < b$. Let $\mathcal{Q}$ be a set of hypothesis distributions for some distribution p over the space X such that at least one $q^* \in \mathcal{Q}$ is such that $d_{\mathrm{KL}}(p||q^*) \leq \varepsilon$. Suppose also that $a \leq q(x) \leq b$ for all $q \in \mathcal{Q}$ and for all x such that $p(x) > 0$. Then running the maximum likelihood algorithm on $\mathcal{Q}$ using a set S of i.i.d. samples from p, where $|S| = k$, outputs a $q^{ML} \in \mathcal{Q}$ such that $d_{\mathrm{KL}}(p||q^{ML}) \leq 4\varepsilon$ with probability $1 - \delta$ where*

$$\delta \leq (|\mathcal{Q}| + 1) \exp\left(\frac{-2k\varepsilon^2}{\log^2\left(\frac{b}{a}\right)}\right).$$

To succeed with probability at least $2/3$, we need that

$$k \geq \frac{\log\left(3(|\mathcal{Q}| + 1)\right) \log^2\left(\frac{b}{a}\right)}{2\varepsilon^2}$$

For the Ising model, a KL-cover $\mathcal{Q}$ would consist of creating a $\mathrm{poly}(n/\varepsilon)$ mesh for each parameter. Since there are $O(n^2)$ parameters, the cover will have a size of $\mathrm{poly}(n/\varepsilon)^{n^2}$. Letting β and h denote the maximum edge and node parameter (respectively), then the ratio b/a in the above theorem is such that

$$\frac{b}{a} \geq \exp\left(O(n^2\beta + nh)\right).$$

Therefore, the number of samples required by this approach would be

$$\begin{aligned} k &= O\left(\frac{n^2 \log\left(\frac{n}{\varepsilon}\right) \dot{(}n^2\beta + nh)^2}{\varepsilon^2}\right) \\ &= \tilde{O}\left(\frac{n^6\beta^2 + n^4h^2}{\varepsilon^2}\right) \end{aligned}$$

which is more expensive than our baseline, the localization algorithm of Theorem 22. Additionally, this algorithm is computationally inefficient, as it involves iterating over all hypotheses in the exponentially large set $\mathcal{Q}$. To summarize, there are a number of issues preventing a learning-based approach from giving an efficient tester.

Chapter 5

Private Distribution Testing and Property Estimation

5.1 Introduction

In several modern fields of research and application, we often wish to perform statistical inference on data which contains sensitive information about individuals. For example, in medical studies, where the data may contain individuals' health records and whether they carry some disease which bears a social stigma. Alternatively, one can consider a map application which suggests routes based on aggregate positions of individuals, which contains delicate information including users' residence data. It may thus be crucial to guarantee that operating on the samples needed to test a statistical hypothesis protects sensitive information about the samples. This does not preclude our overall goals of statistical analysis, as we are trying to infer properties of the population p, and not the samples which are drawn from said population.

That said, without careful experimental design, published statistical findings may be prone to leaking sensitive information about the sample. As a notable example, it was recently shown that one can determine the identity of some individuals who participated in genome-wide association studies [HSR+08]. This realization has motivated a surge of interest in developing data sharing techniques with an explicit focus on maintaining privacy of the data [JS13, USF13, YFSU14, SSB16].

Privacy-preserving computation has enjoyed significant study in a number of fields, including statistics and almost every branch of computer science, including cryptography, machine learning, algorithms, and database theory – see, e.g., [Dal77, AW89, AA01, DN03, Dwo08, DR14] and references therein. Perhaps the most celebrated notion of privacy, proposed by theoretical computer scientists, is *differential privacy* [DMNS06]. Informally, an algorithm is differentially private if its outputs on neighboring datasets (differing in a single element) are statistically close (for a more precise definition, see Section 5.2). Differential privacy has become the standard for theoretically-sound data privacy, leading to its adoption by several organizations, including Google, Apple, and the US Census Bureau [EPK14, Dif17, DLS+17].

In this chapter, our goal is to provide algorithms for various statistical tasks such as distribution testing and property estimation with the additional constraint of differential privacy. In particular, we wish for our algorithms to simultaneously provide:

- **Correctness**: With high probability, the algorithm should should be accurate;
- **Privacy**: The algorithm should be differentially private, for any dataset which it is provided.

Notice that the correctness constraint is the standard one which we have considered so far, but the privacy constraint is considered in addition. We emphasize that the privacy constraint is *worst-case*: no matter how the samples were generated (e.g., if our distributional assumptions were wrong, if the input is some very unlikely set of samples, or even if the dataset is entirely adversarially generated), we must guarantee privacy. The pertinent question is how much the requirement of privacy increases the number of samples that are needed to guarantee correctness. We introduce the study of several problems in this setting, including identity testing, and estimation of entropy, support coverage, support size, and distance to uniformity.

5.1.1 Results, Techniques, and Discussion

In this section, we overview our results and the methods used to obtain them. Since the techniques are somewhat different, we discuss our results on identity testing in Section 5.1.1.1,

and on the estimation of other properties in Section 5.1.1.2.

5.1.1.1 Identity Testing

As we have already established, in the absence of privacy constraints, the sample complexity of identity testing is $\Theta\left(\frac{\sqrt{n}}{\alpha^2}\right)$. Our main theoretical result is the following upper bound on the sample complexity of private identity testing:

Theorem 43. *There exists an ε-differentially private algorithm for the $(\alpha, \beta_{\mathrm{I}}, \beta_{\mathrm{II}})$-identity testing problem for q, distinguishing the cases:*

- $p = q$;
- $d_{\mathrm{TV}}(p, q) \geq \varepsilon$.

The algorithm uses $O\left(\left(\frac{n^{1/2}}{\alpha^2} + \frac{\sqrt{n \log n}}{\alpha^{1.5}\varepsilon}\right) \cdot \log(1/\beta)\right)$ samples, where $\beta = \min(\beta_{\mathrm{I}}, \beta_{\mathrm{II}})$.

Theorem 43 is proved in Section 5.3.3. Algorithm 13 achieves the desired bound. Notice that privacy comes for free when the privacy requirement ε is $\Omega(\sqrt{\alpha})$ – for example when $\varepsilon = 10\%$ and the required statistical accuracy is 3%.

The precise constants sitting in the $O(\cdot)$ notation of the sample complexity of Theorem 43 are given in the proof. We experimentally verify the sample efficiency of our tests by comparing them to recently proposed private statistical tests [GLRV16, KR17], discussed in more detail shortly. Fixing a differential privacy and type I, type II error constraints, we compare how many samples are required by our and their methods to distinguish between hypotheses that are $\alpha = 0.1$ apart in total variation distance. We find that different algorithms are more efficient depending on the regime and properties desired by the analyst. Our experiments and further discussion of the tradeoffs are presented in Section 5.5.1.

A standard approach to turn an algorithm differentially private is to use repetition. As already mentioned above, absent differential privacy constraints, statistical tests have been provided that use an optimal $m = O(\frac{\sqrt{n}}{\alpha^2} \cdot \log \frac{1}{\beta})$ number of samples. A trivial way to get $(\varepsilon, 0)$-differential privacy using such a non-private test is to create $O(1/\varepsilon)$ datasets, each comprising m samples from p, and run the non-private test on one of these datasets, chosen randomly. It is clear that changing the value of a single element in the combined dataset

may only affect the output of the test with probability at most ε. Thus the output is $(\varepsilon, 0)$-differentially private; see Section 5.3.1 for a proof. The issue with this approach is that the total number of samples that it draws is $m/\varepsilon = O(\frac{\sqrt{n}}{\varepsilon\alpha^2} \cdot \log \frac{1}{\beta})$, which is higher than our target. See Corollary 13.

A different approach towards private hypothesis testing is to look deeper into the non-private tests and try to "privatize" them. The most sample-efficient tests are variations of the classical χ^2-test. They compute the number of times, N_i, that element $i \in [n]$ appears in the sample and aggregate those counts using a statistic that equals, or is close to, the χ^2-divergence between the empirical distribution defined by these counts and the hypothesis distribution q. They accept q if the statistic is low and reject q if it is high, using some threshold.

A reasonable approach to privatize such a test is to add noise, e.g. Laplace$(1/\varepsilon)$ noise, to each count N_i, before running the test. It is well known that adding Laplace$(1/\varepsilon)$ noise to a set of counts makes them differentially private, see Theorem 48. However, it also increases the variance of the statistic. This has been noticed empirically in recent work of [GLRV16] for the χ^2-test. We show that the variance of the optimal χ^2-style test statistic significantly increases if we add Laplace noise to the counts, in Section 5.3.2.1, thus increasing the sample complexity from $O(\sqrt{n})$ to $\Omega(n^{3/4})$. So this route, too, seems problematic.

A last approach towards designing differentially private tests is to exploit the distance beween the null and the alternative hypotheses. A correct test should accept the null with probability close to 1, and reject an alternative that is α-far from the null with probability close to 1, but there are no requirements for correctness when the alternative is very close to the null. We could thus try to interpolate smoothly between datasets that we expect to see when sampling the null and datasets that we expect to see when sampling an alternative that is far from the null. Rather than outputting "accept" or "reject" by merely thresholding our statistic, we would like to tune the probability that we output "reject" based on the value of our statistic, and make it so that the "reject" probability is ε-Lipschitz as a function of the dataset. Moreover, the probability should be close to 0 on datasets that we expect to see under the null and close to 1 on datasets that we expect to see under an alternative that is α-far. As we show in Section 5.3.2.2, χ^2-style statistics have high sensitivity, requiring

$\omega(\sqrt{n})$ samples to be made appropriately Lipschitz.

While both the approach of adding noise to the counts, and that of turning the output of the test Lipschitz fail in isolation, our test actually goes through by intricately combining these two approaches. It has two steps:

1. A *filtering step,* whose goal is to "reject" when p is blatantly far from q. This step is performed by comparing the counts N_i with their expectations under q, after having added Laplace($1/\varepsilon$) noise to these counts. If the noisy counts deviate from their expectation, taking into account the extra variance introduced by the noise, then we can safely "reject." Moreover, because noise was added, this step is differentially private.

2. If the filtering step fails to reject, we perform a *statistical step.* This step just computes the χ^2-style statistic from [ADK15], *without adding noise to the counts.* The crucial observation is that if the filtering step does not reject, then the statistic is actually ε-Lipschitz with respect to the counts, and thus the value of the statistic is still differentially private. We use the value of the statistic to determine the bias of a coin that outputs "reject."

Details of our test are given in Section 5.3.3.

5.1.1.2 Property Estimation

Results. Our main results show that we can privately estimate many properties of interest at a very low cost. In particular, we focus on estimation of entropy, support size, support coverage, and distance to uniformity. For example, if one wishes to privately estimate entropy, this incurs an additional additive cost in the sample complexity which is very close to linear in $1/\alpha\varepsilon$. We draw attention to two features of this bound. First, this is independent of n. All the problems we consider have complexity $\Theta(n/\log n)$, so in the primary regime of study where $n \gg 1/\alpha\varepsilon$, this small additive cost is dwarfed by the inherent sample complexity of the non-private problem. Second, the bound is almost linear in $1/\alpha\varepsilon$. We note that performing even the most basic statistical task privately, estimating the bias of a coin, incurs this linear dependence. Surprisingly, we show that much more sophisticated inference tasks can be privatized at almost no cost. In particular, these properties imply that the

additive cost of privacy is $o(1)$ in the most studied regime where the support size is large. In general, this is not true – for many other problems, including distribution estimation and hypothesis testing, the additional cost of privacy depends significantly on the support size or dimension [DHS15, CDK17, ASZ17, ADR18]. We also provide lower bounds, showing that our upper bounds are almost tight.

More formally, our results are as follows:

Theorem 44. *The sample complexity of support coverage estimation is*

$$C(S_k,\alpha,\varepsilon)=\begin{cases} O\left(\frac{k\log(1/\alpha)}{\log k}+\frac{k\log(1/\alpha)}{\log(2+\varepsilon k)}\right), & \text{when } k\geq\frac{1}{\alpha\varepsilon} \\ O\left(\frac{1}{\alpha^2}+\frac{1}{\alpha\varepsilon}\right), & \text{when } \frac{1}{\alpha}\leq k\leq\frac{1}{\alpha\varepsilon} \\ O\left(k^2+\frac{k}{\varepsilon}\right). & \text{when } k\leq\frac{1}{\alpha} \end{cases}$$

Furthermore,

$$C(S_k,\alpha,\varepsilon)=\Omega\left(\frac{k\log(1/\alpha)}{\log k}+\frac{1}{\alpha\varepsilon}\right).$$

Theorem 45. *The sample complexity of support size estimation is*

$$C(S,\alpha,\varepsilon)=\begin{cases} O\left(\frac{n\log^2(1/\alpha)}{\log n}+\frac{n\log^2(1/\alpha)}{\log(2+\varepsilon n)}\right), & \text{when } n\geq\frac{1}{\alpha\varepsilon} \\ O\left(n\log(1/\alpha)+\frac{1}{\alpha\varepsilon}\right), & \text{when } \frac{1}{\alpha}\leq n\leq\frac{1}{\alpha\varepsilon} \\ O\left(n\log n+\frac{n}{\varepsilon}\right). & \text{when } n\leq\frac{1}{\alpha} \end{cases}$$

Furthermore,

$$C(S,\alpha,\varepsilon)=\begin{cases} \Omega\left(\frac{n\log^2(1/\alpha)}{\log n}+\frac{1}{\alpha\varepsilon}\right), & \text{when } n\geq\frac{1}{\alpha} \\ \Omega\left(n\log n+\frac{n}{\varepsilon}\right). & \text{when } n\leq\frac{1}{\alpha} \end{cases}$$

Theorem 46. *Let $\lambda>0$ be* any *small fixed constant. For instance, λ can be chosen to be any constant between* 0.01 *and* 1*. We have the following upper bounds on the sample complexity of estimating distance to uniformity:*

$$C(\|p-\mathcal{U}_n\|_1,\alpha,\varepsilon)=O\left(\frac{n}{\alpha^2}+\frac{1}{\alpha\varepsilon}\right)$$

and

$$C(\|p-\mathcal{U}_n\|_1,\alpha,\varepsilon)=O\left(\frac{n}{\lambda^2\alpha^2\log n}+\left(\frac{1}{\alpha\varepsilon}\right)^{1+\lambda}\right).$$

Furthermore,

$$C(\|p-\mathcal{U}_n\|_1,\alpha,\varepsilon)=\Omega\left(\frac{n}{\alpha^2\log n}+\frac{n^{1/2}}{\alpha\varepsilon^{1/2}}+\frac{n^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}}+\frac{1}{\alpha\varepsilon}\right).$$

Theorem 47. *Let $\lambda>0$ be* any *small fixed constant. For instance, λ can be chosen to be any constant between* 0.01 *and* 1*. We have the following upper bounds on the sample complexity of entropy estimation:*

$$C(H,\alpha,\varepsilon)=O\left(\frac{n}{\alpha}+\frac{\log^2(\min\{n,m\})}{\alpha^2}+\frac{1}{\alpha\varepsilon}\log\left(\frac{1}{\alpha\varepsilon}\right)\right)$$

and

$$C(H,\alpha,\varepsilon)=O\left(\frac{n}{\lambda^2\alpha\log n}+\frac{\log^2(\min\{n,m\})}{\alpha^2}+\left(\frac{1}{\alpha\varepsilon}\right)^{1+\lambda}\right).$$

Furthermore,

$$C(H,\alpha,\varepsilon)=\Omega\left(\frac{n}{\alpha\log n}+\frac{\log^2(\min\{n,m\})}{\alpha^2}+\frac{\log n}{\alpha\varepsilon}\right).$$

Discussion. At a high level, we wish to emphasize the following two points:

1. Our upper bounds show that the cost of privacy in these settings is often negligible compared to the sample complexity of the non-private statistical task, especially when we are dealing with distributions over a large support. Furthermore, our upper bounds are almost tight in all parameters.

2. The algorithmic complexity introduced by the requirement of privacy is minimal, consisting only of a single step which noises the output of an estimator. In other words, our methods are realizable in practice, and we demonstrate the effectiveness on several synthetic and real-data examples.

Before we continue, we emphasize that, in Theorems 44 and 45, we consider the "sublinear" regime to be of primary interest (when $k\geq\frac{1}{\alpha\varepsilon}$ or $n\geq\frac{1}{\alpha\varepsilon}$, respectively), both technically,

and in terms of parameter regimes which may be of greatest interest in practice. We include results for other regimes mostly for completeness.

First, we examine our results on support coverage and support size estimation in the sublinear regime, when $k \geq \frac{1}{\alpha\varepsilon}$ (focusing on support coverage for simplicity, but support size is similar). In this regime, if $\varepsilon = \Omega(k^\gamma/k)$ for any constant $\gamma > 0$, then up to constant factors, our upper bound is within a constant factor of the optimal sample complexity without privacy constratints. In other words, for most meaningful values of ε, privacy comes for free. In the non-sublinear regime for these problems, we provide upper and lower bounds which match in a number of cases. We note that in this regime, the cost of privacy may not be a lower order term – however, this regime only occurs when one requires very high accuracy, or unreasonably large privacy, which we consider to be of somewhat lesser interest.

Next, we turn our attention to estimating distance to uniformity and entropy. We note that the second upper bound in Theorems 46 and 47 has a parameter λ that indicates a tradeoff between the sample complexity incurred in the first and third term. This parameter determines the degree of a polynomial to be used. As the degree becomes smaller (corresponding to a large λ), accuracy of the polynomial estimator decreases, however, at the same time, low-degree polynomials have a small sensitivity, allowing us to privatize the outcome.

In terms of our theoretical results, one can think of $\lambda = 0.01$. With this parameter setting, it can be observed that our upper bounds are almost tight. For example, one can see that the upper and lower bounds match to either logarithmic factors (when looking at the first upper bound), or a very small polynomial factor in $1/\alpha\varepsilon$ (when looking at the second upper bound). For our experimental results on entropy estimation, we empirically determined an effective value for the parameter λ on a single synthetic instance. We then show that this choice of parameter generalizes, giving highly-accurate private estimation in other instances, on both synthetic and real-world data.

Approach. Our approach works by choosing statistics for these tasks which possess bounded sensitivity, which is well-known to imply privacy under the Laplace or Guassian[1] mechanism.

[1]This intentional misspelling is dedicated to the memory of Michael B. Cohen. While we weren't close friends, he invariably left a strong impression on everyone whose life he touched, and I was no exception. He is sorely missed.

We note that bounded sensitivity of statistics is not always something that can be taken for granted. Indeed, for many fundamental tasks, optimal algorithms for the non-private setting may be highly sensitive, thus necessitating crucial modifications to obtain differential privacy, as evidenced by our results on identity testing. Thus, careful choice and design of statistics must be a priority when performing inference with privacy considerations.

To this end, we leverage recent results of [ADOS17], which studies estimators for non-private versions of the problems we consider. The main technical work in their paper exploits bounded sensitivity to show sharp cutoff-style concentration bounds for certain estimators, which operate using the principle of best-polynomial approximation. They use these results to show that a single algorithm, the Profile Maximum Likelihood (PML), can estimate all these properties simultaneously. On the other hand, we consider the sensitivity of these estimators for purposes of privacy – the same property is utilized by both works for very different purposes, a connection which may be of independent interest.

We note that bounded sensitivity of a statistic may be exploited for purposes other than privacy. For instance, by McDiarmid's inequality, any such statistic also enjoys very sharp concentration of measure, implying that one can boost the success probability of the test at an additive cost which is logarithmic in the inverse of the failure probability. One may naturally conjecture that, if a statistical task is based on a primitive which concentrates in this sense, then it may also be privatized at a low cost. However, this is not true – estimating a discrete distribution in ℓ_1 distance is such a task, but the cost of privatization depends significantly on the support size [DHS15].

One can observe that, algorithmically, our method is quite simple: compute the non-private statistic, and add a relatively small amount of Laplace noise. The non-private statistics have recently been demonstrated to be practical [OSW16, WY18], and the additional cost of the Laplace mechanism is minimal. This is in contrast to several differentially private algorithms which invoke significant overhead in the quest for privacy. Our algorithms attain almost-optimal rates (which are optimal up to constant factors for most parameter regimes of interest), while simultaneously operating effectively in practice, as demonstrated in our experimental results.

Experimental Results. We demonstrate the efficacy of our method with experimen-

tal evaluations. As a baseline, we compare with the non-private algorithms of [OSW16] and [WY18]. Overall, we find that our algorithms' performance is nearly identical, showing that, in many cases, privacy comes (essentially) for free. We begin with an evaluation on synthetic data. Then, inspired by [VV13, OSW16], we analyze text corpus consisting of words from Hamlet, in order to estimate the number of unique words which occur. Finally, we investigate name frequencies in the US census data. This setting has been previously considered by [OSW16], but we emphasize that this is an application where private statistical analysis is critical. This is proven by efforts of the US Census Bureau to incorporate differential privacy into the 2020 US census [DLS+17].

5.1.2 Related Work

Recently, there has been significant interest in performing statistical tasks under differential privacy constraints. Several papers have addressed our original motivation of privately performing GWASs [JS13, USF13, YFSU14, SSB16]. A number of recent works [WZ10, Smi11, WLK15, GLRV16, KR17, Rog17] (and a work simultaneous with ours on identity testing, focused on independence testing [KSF17]) investigate differentially private hypothesis testing in the asymptotic regime. In particular, they fix a desired significance (type I error) and privacy requirement, and study the asymptotic distribution of the test statistics. [VS09] give some finite sample corrections required to compute valid p-values after noising. Our work on identity testing (published as [CDK17]) was the first to focus on private hypothesis testing in the minimax setting. Following the publication of [CDK17], further works have established tight bounds on identity and equivalence testing [ASZ17, ADR18]. Some have recently studied testing in the stricter *local* privacy model, in both the asymptotic setting [GR18] and the minimax setting [She18, ACFT18]. There has also been study on private distribution learning, in both the local and the global privacy model [Smi11, BNSV15, DHS15, KS16, WHW+16, KBR16, DJW17, KV18, ASZ18, KLSU18, YB18]. Here, we wish to estimate parameters of the distribution, rather than just a particular property of interest. Private supervised learning has also been a lively field of study [KLN+11, CH11, BBKN14, FX15, BNS15, BNS16a, BNS16b]. [RRST16] studies differential privacy for the purpose of generating valid p-values for adaptive hypothesis testing (the general di-

rection of privacy for adaptive data analysis has recently enjoyed a great deal of study, see, e.g., [DFH+15]). A number of other statistical problems have been studied with privacy requirements, including clustering [WWS15, BDL+17, NS18], principal component analysis [CSS13, KT13, HP14, DTTZ14, GGB18], ordinary least squares [She17], and much more. An interesting work which is unrelated to our goal is [GM18], which investigates the complexity of property testing whether an algorithm is differentially private. See [DR14, Vad17] for more general coverage about the theory of differential privacy.

5.1.3 Organization

In Section 5.2, we go over some additional preliminaries. In Sections 5.3 and 5.4, we present our results on private identity testing and property estimation, respectively. We conclude with experimental results in Section 5.5.

5.2 Preliminaries

In this chapter, we diverge slightly from our standard notation. In particular, we use ε and δ for parameters of differential privacy. We use α for the distance/accuracy parameter (rather than the usual ε), and β for the probability of failure (rather than the usual δ).

Privacy Preliminaries. We have the following basic definition of differential privacy.

Definition 13. *A randomized algorithm M with domain $\mathbb{N}^n$ is* (ε, δ)-differentially private *if for all $S \subseteq \mathrm{Range}(M)$ and for all pairs of inputs D, D' such that $\|D - D'\|_1 \leq 1$:*

$$\Pr\left[M(D) \in S\right] \leq e^{\varepsilon} \Pr\left[M(D') \in S\right] + \delta.$$

If $\delta = 0$, the guarantee is called pure *differential privacy, and we refer to it as ε-differential privacy.*

We also recall the definition of zero-concentrated differential privacy from [BS16] and its relationship to differential privacy.

Definition 14. *A randomized algorithm M with domain $\mathbb{N}^n$ is* ρ-zero-concentrated differentially private *(ρ-zCDP) if for all pairs of inputs D, D' such that $\|D - D'\|_1 \leq 1$ and all $\alpha \in (1, \infty)$:*

$$\mathrm{D}_\alpha(M(D)||M(D')) \leq \rho\alpha,$$

where D_α is the α-Rényi divergence between the distribution of $M(D)$ and $M(D')$.

Proposition 13 (Propositions 1.3 and 1.4 of [BS16])**.** *If a mechanism M_1 satisfies $(\varepsilon, 0)$-differential privacy, then M_1 satisfies $\frac{\varepsilon^2}{2}$-zCDP. If a mechanism M_2 satisfies ρ-zCDP, then M_2 satisfies $(\rho + 2\sqrt{\rho \log(1/\delta)}, \delta)$-differential privacy for any $\delta > 0$.*

Property Testing and Estimation Definitions Preliminaries.

Definition 15. *An algorithm for the* $(\alpha, \beta_{\mathrm{I}}, \beta_{\mathrm{II}})$-identity testing *problem with respect to a (known) distribution q takes m samples from an (unknown) distribution p and has the following guarantees:*

- *If $p = q$, then with probability at least $1 - \beta_{\mathrm{I}}$ it outputs "$p = q$;"*
- *If $d_{\mathrm{TV}}(p, q) \geq \alpha$, then with probability at least $1 - \beta_{\mathrm{II}}$ it outputs "$p \neq q$."*

In particular, β_{I} and β_{II} are the type I and type II errors of the test. Parameter α is the radius of distinguishing accuracy. Notice that, when p satisfies neither of cases above, the algorithm's output may be arbitrary.

Let $\Delta \triangleq \{(p(1), \ldots, p(n)) : p(i) \geq 0, \sum_{i=1}^n p(i) = 1, 1 \leq n \leq \infty\}$ be the set of discrete distributions over a countable support. Let Δ_n be the set of distributions in Δ with at most n non-zero probability values. A *property* $f(p)$ is a mapping from $\Delta \to \mathbb{R}$. We now describe the classical distribution property estimation problem, and then state the problem under differential privacy.

Definition 16. *Given α, β, f, and independent samples X_1^m from an unknown distribution p, design an estimator $\hat{f} : X_1^m \to \mathbb{R}$ such that with probability at least $1 - \beta$, $\left|\hat{f}(X_1^m) - f(p)\right| < \alpha$. The* sample complexity *of $\hat{f}$, $C_{\hat{f}}(f, \alpha, \beta) \triangleq \min\{n : \Pr\left[\left|\hat{f}(X_1^m) - f(p)\right| > \alpha\right] < \beta\}$ is the smallest number of samples to estimate f to accuracy α, and error β. We study the*

problem for $\beta = 1/3$, and by the median trick, we can boost the success probability to $1-\beta$ with an additional multiplicative $\log(1/\beta)$ *more samples. Therefore, focusing on $\beta = 1/3$, we define $C_{\hat{f}}(f,\alpha) \triangleq C_{\hat{f}}(f,\alpha,1/3)$. The sample complexity of estimating a property $f(p)$ is the minimum sample complexity over all estimators: $C(f,\alpha) = \min_{\hat{f}} C_{\hat{f}}(f,\alpha)$.*

Given $\alpha, \varepsilon, \beta$, f, and independent samples X_1^m from an unknown distribution p, design an ε-differentially private estimator $\hat{f} : X_1^m \to \mathbb{R}$ such that with probability at least $1-\beta$, $\left|\hat{f}(X_1^m) - f(p)\right| < \alpha$. Similar to the non-private setting, the sample complexity *of ε-differentially private estimation problem is $C(f,\alpha,\varepsilon) = \min_{\hat{f}:\hat{f} \text{ is } \varepsilon\text{-DP}} C_{\hat{f}}(f,\alpha,1/3)$, the smallest number of samples m for which there exists such an ε-DP $\pm\alpha$ estimator with error probability at most 1/3.*

We note that if an algorithm is to satisfy both the privacy and correctness conditions, the latter condition need only be satisfied *stochastically*: the algorithm only needs to be accurate with probability $1-\beta$, where the probability is over the randomness of the algorithm and the sampling process, and may only need to be accurate when the underlying distribution p satisfies some assumptions (e.g., in the identity testing case, when either $p = q$ or $d_{\mathrm{TV}}(p,q) \geq \alpha$). On the other hand, the former privacy property must be satisfied for *all realizations* of the samples from p (and in particular, for identity testing, when p does not fall into the two cases of interest).

Tools for Private Statistical Estimation. In their original paper [DMNS06] provides a scheme for differential privacy, known as the Laplace mechanism. This method adds Laplace noise to a non-private scheme in order to make it private. We first define the sensitivity of an estimator, and then state their result in our setting.

Definition 17. *The* sensitivity *of an estimator $\hat{f} : [n]^m \to \mathbb{R}$ is* $\Delta_{m,\hat{f}} \triangleq \max_{d_{hamming}(X_1^m, Y_1^m) \leq 1} \left|\hat{f}(X_1^m) - \hat{f}(Y_1^m)\right|$. *Let $D_{\hat{f}}(\alpha,\varepsilon) = \min\{m : \Delta_{m,\hat{f}} \leq \alpha\varepsilon\}$.*

Lemma 41.

$$C(f,\alpha,\varepsilon) = O\left(\min_{\hat{f}}\left\{C_{\hat{f}}(f,\alpha/2) + D_{\hat{f}}\left(\frac{\alpha}{4},\varepsilon\right)\right\}\right).$$

Proof. [DMNS06] showed that for a function with sensitivity $\Delta_{m,\hat{f}}$, adding Laplace noise $X \sim Lap(\Delta_{m,\hat{f}}/\varepsilon)$ makes the output ε-differentially private. By the definition of $D_{\hat{f}}(\frac{\alpha}{4},\varepsilon)$,

the Laplace noise we add has parameter at most $\frac{\alpha}{4}$. Recall that the probability density function of $Lap(b)$ is $\frac{1}{2b}e^{-\frac{|x|}{b}}$, hence we have $\Pr[\,|X| > \alpha/2\,] < \frac{1}{e^2}$. By the union bound, we get an additive error larger than $\alpha = \frac{\alpha}{2} + \frac{\alpha}{2}$ with probability at most $1/3 + \frac{1}{e^2} < 0.5$. Hence, with the median trick, we can boost the error probability to $1/3$, at the cost of a constant factor in the number of samples. □

This can be specialized to the specific case where we have a set of counts of n items.

Theorem 48 (Theorem 3.6 of [DR14])**.** *Given a set of counts $N_1, \ldots, N_n$, the noised counts $(N_1+Y_1, \ldots, N_n+Y_n)$ are $(\varepsilon, 0)$-differentially private when the Y_i's are i.i.d. random variables drawn from Laplace$(1/\varepsilon)$.*

To prove sample complexity lower bounds for differentially private estimators, we observe that the estimator can be used to test between two distributions with distinct property values, hence is a harder problem. For lower bounds on differentially private testing, [ASZ17] gives the following argument based on coupling:

Lemma 42. *Suppose there is a coupling between distributions p and q over $\mathcal{X}^m$, such that $\mathbf{E}\left[d_{\text{hamming}}\left(X_1^m, Y_1^m\right)\right] \leq D$. Then, any ε-differentially private algorithm that distinguishes between p and q with error probability at most $1/3$ must satisfy $D = \Omega\left(\frac{1}{\varepsilon}\right)$.*

5.3 `Priv'IT`: Private Identity Testing

In this section, we describe our results on private identity testing. We start in Section 5.3.1 by describing baseline approach, based on the paradigm of subsample and aggregate. This method is applicable to any decision problem. In Section 5.3.2, we describe roadblocks for some natural approaches to performing private identity testing. In Section 5.3.3, we explain how we bypass these roadblocks, and give a more efficient algorithm.

5.3.1 A Simple Upper Bound

In this section, we provide an $O\left(\frac{\sqrt{n}}{\alpha^2\varepsilon}\right)$ upper bound for the differentially private identity testing problem, based on the subsample and aggregate paradigm. More generally, we show

that if an algorithm requires a dataset of size m for a decision problem, then it can be made $(\varepsilon, 0)$-differentially private at a multiplicative cost of $1/\varepsilon$ in the sample size. This is a folklore result, but we include and prove it here for completeness.

Theorem 49. *Suppose there exists an algorithm for a decision problem P which succeeds with probability at least $1 - \beta$ and requires a dataset of size m. Then there exists an $(\varepsilon, 0)$-differentially private algorithm for P which succeeds with probability at least $\frac{4}{5}(1 - \beta) + 1/10$ and requires a dataset of size $O(m/\varepsilon)$.*

Proof. First, with probability $1/5$, we flip a coin and output yes or no with equal probability. This guarantees that we have probability at least $1/10$ of either outcome, which will allow us to satisfy the multiplicative guarantee of differential privacy.

We then draw $10/\varepsilon$ datasets of size m, and solve the decision problem (non-privately) for each of them. Finally, we select a random one of these computations and output its outcome.

The correctness follows, since we randomly choose the right answer with probability $1/10$, or with probability $4/5$, we solve the problem correctly with probability $1-\beta$. As for privacy, we note that, if we remove a single element of the dataset, we may only change the outcome of one of these computations. Since we pick a random computation, this is selected with probability $\varepsilon/10$, and thus the probability of any outcome is additively shifted by at most $\varepsilon/10$. Since we know the minimum probability of any output is $1/10$, this gives the desired multiplicative guarantee required for $(\varepsilon, 0)$-differential privacy. □

We obtain the following corollary by noting that the tester of [ADK15] (among others) requires $O(\sqrt{n}/\alpha^2)$ samples for identity testing.

Corollary 13. *There exists an $(\varepsilon, 0)$-differentially private testing algorithm for the $(\alpha, \beta_{\mathrm{I}}, \beta_{\mathrm{II}})$-identity testing problem for any distribution q which requires*

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

samples, where $\beta = \min(\beta_{\mathrm{I}}, \beta_{\mathrm{II}})$.

5.3.2 Roadblocks to Differentially Private Identity Testing

In this section, we describe roadblocks which prevent two natural approaches to differentially private testing from working.

In Section 5.3.2.1, we show that if one simply adds Laplace noise to the empirical counts of a dataset (i.e., runs the Laplace mechanism of Theorem 48) and then attempts to run an optimal identity tester, the variance of the statistic increases dramatically, and thus results in a much larger sample complexity, even for the case of uniformity testing. The intuition behind this phenomenon is as follows. When performing uniformity testing in the small sample regime (when the number of samples m is the square root of the domain size n), we will see a $(1 - o(1))n$ elements 0 times, $O(\sqrt{n})$ elements 1 time, and $O(1)$ elements 2 times. If we add $Laplace(10)$ noise to guarantee $(0.1, 0)$-differential privacy, this obliterates the signal provided by these collision statistics, and thus many more samples are required before the signal prevails.

In Section 5.3.2.2, we demonstrate that χ^2 statistics have high sensitivity, and thus are not naturally differentially private. In other words, if we consider a χ^2 statistic Z on two datasets D and D' which differ in one record, $|Z(D) - Z(D')|$ may be quite large. This implies that methods such as rescaling this statistic and interpreting it as a probability, or applying noise to the statistic, will not be differentially private until we have taken a large number of samples.

5.3.2.1 A Laplaced χ^2-statistic has large variance

Proposition 14. *Applying the Laplace mechanism to a dataset before applying the identity tester of [ADK15] results in a significant increase in the variance, even when considering the case of uniformity. More precisely, if we consider the statistic*

$$Z'(D) = \sum_{i \in [n]} \frac{(N_i + Y_i - m/n)^2 - (N_i + Y_i)}{m/n}$$

where N_i is the number of occurrences of symbol i in the dataset D (which is of size $Poisson(m)$) and $Y_i \sim Laplace(1/\varepsilon)$, then

- *If p is uniform, then $\mathbf{E}[Z'] = \frac{2n^2}{\varepsilon^2 m}$ and $\mathbf{Var}[Z'] \geq \frac{20n^3}{\varepsilon^4 m^2}$.*
- *If p is a particular distribution which is α-far in total variation distance from uniform, then $\mathbf{E}[Z'] = 4m\alpha^2 + \frac{2n^2}{\varepsilon^2 m}$.*

The variance of the statistic can be compared to that of the unnoised statistic, which is upper bounded by $m^2\alpha^4$. We can see that the noised statistic has larger variance until $m = \Omega(n^{3/4})$.

Proof. First, we compute the mean of Z'. Note that since $|D| \sim Poisson(m)$, the N_i's will be independently distributed as $Poisson(mp_i)$ (see, i.e., [ADK15] for additional discussion).

$$\begin{aligned}
\mathbf{E}[Z'] &= \mathbf{E}\Bigg[\sum_{i\in[n]} \frac{(N_i + Y_i - m/n)^2 - (N_i + Y_i)}{m/n}\Bigg] \\
&= \mathbf{E}\Bigg[\sum_{i\in[n]} \frac{(N_i - m/n)^2 - N_i}{m/n} \\
&\quad + \sum_{i\in[n]} \frac{Y_i^2 + 2Y_i(N_i - m/n) - Y_i}{m/n}\Bigg] \\
&= m \cdot \chi^2(p, q) + \sum_{i\in[n]} \frac{\frac{2}{\varepsilon^2}}{m/n} \\
&= m \cdot \chi^2(p, q) + \frac{2n^2}{\varepsilon^2 m}
\end{aligned}$$

In other words, the mean is a rescaling of the χ^2 distance between p and q, shifted by some constant amount. When $p = q$, the χ^2-distance between p and q is 0, and the expectation is just the second term. Focus on the case where n is even, and consider p such that $p_i = (1 + 2\alpha)/n$ if i is even, and $(1 - 2\alpha)/n$ otherwise. This is α-far from uniform in total variation distance. Furthermore, by direct calculation, $\chi^2(p, q) = 4\alpha^2$, and thus the expectation of Z' in this case is $4m\alpha^2 + \frac{2n^2}{\varepsilon^2 m}$.

Next, we examine the variance of Z'. Let $\lambda_i = mp_i$ and $\lambda'_i = mq_i = m/n$. By a similar computation as before, we have that

$$\begin{aligned}
\mathbf{Var}[Z'] = \sum_{i\in[n]} \frac{1}{\lambda_i'^2}\Bigg[&2\lambda_i^2 + 4\lambda_i(\lambda_i - \lambda'_i)^2 \\
&+ \frac{1}{\varepsilon^2}(8\lambda_i + 2(2\lambda_i - 2\lambda'_i - 1)^2) + \frac{20}{\varepsilon^4}\Bigg].
\end{aligned}$$

Since all four summands of this expression are non-negative, we have that

$$\mathbf{Var}[Z'] \geq \frac{20}{\varepsilon^4} \sum_{i \in [n]} \frac{1}{\lambda_i'^2} = \frac{20n^3}{\varepsilon^4 m^2}.$$

If we wish to use Chebyshev's inequality to separate these two cases, we require that $\mathbf{Var}[Z']$ is at most the square of the mean separation. In other words, we require that

$$\frac{20n^3}{\varepsilon^4 m^2} \leq m^2 \alpha^4,$$

or that

$$m = \Omega\left(\frac{n^{3/4}}{\varepsilon\alpha}\right).$$

□

5.3.2.2 A χ^2-statistic has high sensitivity

Consider the primary statistic which we use in Algorithm 13:

$$Z(D) = \frac{1}{m\alpha^2} \sum_{i \in [n]} \frac{(N_i - mq_i)^2 - N_i}{mq_i}.$$

As shown in Section 5.3.3, $\mathbf{E}[Z] = 0$ if $p = q$ and $\mathbf{E}[Z] \geq 1$ if $d_{\mathrm{TV}}(p, q) \geq \alpha$, and the variance of Z is such that these two cases can be separated with constant probability. A natural approach is to truncate this statistic to the range $[0, 1]$, interpret it as a probability and output the result of $Bernoulli(Z)$ – if $p = q$, the result is likely to be 0, and if $d_{\mathrm{TV}}(p, q) \geq \alpha$, the result is likely to be 1. One might hope that this statistic is naturally private. More specifically, we would like that the statistic Z has low sensitivity, and does not change much if we remove a single individual. Unfortunately, this is not the case. We consider datasets D, D', where D' is identical to D, but with one fewer occurrence of symbol i. It can be shown that the difference in Z is

$$|Z(D) - Z(D')| = \frac{2|N_i - mq_i - 1|}{m^2\alpha^2 q_i}$$

Letting q be the uniform distribution and requiring that this is at most ε (for the sake of privacy), we have a constraint which is roughly of the form

$$\frac{2N_i n}{m^2\alpha^2} \leq \varepsilon,$$

or that

$$m = \Omega\left(\frac{\sqrt{N_i}\sqrt{n}}{\varepsilon^{0.5}\alpha}\right).$$

In particular, if $N_i = n^c$ for any $c > 0$, this does not achieve the desired $O(\sqrt{n})$ sample complexity. One may observe that, if N_i is this large, looking at symbol i alone is sufficient to conclude p is not uniform, even if the count N_i had Laplace noise added. Indeed, our main algorithm of Section 5.3.3 works in part due to our formalization and quantification of this intuition.

5.3.3 `Priv'IT`: An Algorithm for Private Identity Testing

In this section, we prove our main testing upper bound:

Theorem 43. *There exists an ε-differentially private algorithm for the $(\alpha, \beta_{\mathrm{I}}, \beta_{\mathrm{II}})$-identity testing problem for q, distinguishing the cases:*

- $p = q$;
- $d_{\mathrm{TV}}(p, q) \geq \varepsilon$.

The algorithm uses $O\left(\left(\frac{n^{1/2}}{\alpha^2} + \frac{\sqrt{n\log n}}{\alpha^{1.5}\varepsilon}\right)\cdot\log(1/\beta)\right)$ samples, where $\beta = \min(\beta_{\mathrm{I}}, \beta_{\mathrm{II}})$.

The pseudocode for this algorithm is provided in Algorithm 13. We fix the constants $c_1 = 1/4$ and $c_2 = 3/40$. For a high-level overview of our algorithm's approach, we refer the reader to Section 5.1.1.1.

Proof of Theorem 43: We will prove the theorem for the case where $\beta = 1/3$, the general case follows at the cost of a multiplicative $\log(1/\beta)$ in the sample complexity from a standard amplification argument. To be more precise, we can consider splitting our dataset into $O(\log(1/\beta))$ sub-datasets and run the $\beta = 1/3$ test on each one independently. We return the majority result – since each test is correct with probability $\geq 2/3$, correctness of the

Algorithm 13 Priv'IT: A differentially private identity tester

Input: ε; an explicit distribution q; sample access to a distribution p
Define $\mathcal{A} \leftarrow \{i : q_i \geq c_1\alpha/n\}$, $\bar{\mathcal{A}} \leftarrow [n] \setminus \mathcal{A}$
Sample $Y_i \sim Laplace(2/c_2\varepsilon)$ for all $i \in \mathcal{A}$
if there exists $i \in \mathcal{A}$ such that $|Y_i| \geq \frac{2}{c_2\varepsilon} \log\left(\frac{1}{1-(1-c_2)^{1/|\mathcal{A}|}}\right)$ **then**
    return either "$p \neq q$" or "$p = q$" with equal probability
end if
Draw a multiset S of $Poisson(m)$ samples from p
Let N_i be the number of occurrences of the ith domain element in S
for $i \in \mathcal{A}$ **do**
    if $|N_i + Y_i - mq_i| \geq \frac{2}{c_2\varepsilon} \log\left(\frac{1}{1-(1-c_2)^{1/|\mathcal{A}|}}\right) + \max\left\{4\sqrt{mq_i \log n}, \log n\right\}$ **then**
        return "$p \neq q$"
    end if
end for
$Z \leftarrow \frac{2}{m\alpha^2} \sum_{i \in \mathcal{A}} \frac{(N_i - mq_i)^2 - N_i}{mq_i}$
Let T be the closest value to Z which is contained in the interval $[0, 1]$
Sample $b \sim Bernoulli(T)$
if $b = 1$ **then**
    return "$p \neq q$"
else
    return "$p = q$"
end if

overall test follows by Chernoff bound. It remains to argue privacy – note that a neighboring dataset will only result in a single sub-dataset being changed. Since we take the majority result, conditioning on the result of the other sub-tests, the result on this sub-dataset will either be irrelvant to or equal to the overall output. In the former case, any test is private, and in the latter case, we know that the individual test is ε-differentially private. Overall privacy follows by applying the law of total probability.

We require the following two claims, which give bounds on the random variables N_i and Y_i. Note that, due to the fact that we draw $Poisson(m)$ samples, each $N_i \sim Poisson(mp_i)$ independently.

Claim 8. $|Y_i| \leq \frac{2}{c_2\varepsilon} \log\left(\frac{1}{1-(1-c_2)^{1/|\mathcal{A}|}}\right)$ *simultaneously for all* $i \in \mathcal{A}$ *with probability exactly* $1 - c_2$.

Proof. The survival function of the folded Laplace distribution with parameter $2/c_2\varepsilon$ is $\exp(-c_2\varepsilon x/2)$, and the probability that a sample from it exceeding the value $\frac{2}{c_2\varepsilon} \log\left(\frac{1}{1-(1-c_2)^{1/|\mathcal{A}|}}\right)$

is equal to $1-(1-c_2)^{1/|\mathcal{A}|}$. The probability that probability that it does not exceed this value is $(1-c_2)^{1/|\mathcal{A}|}$, and since the Y_i's are independent, the probability that none exceeds this value is $1-c_2$, as desired. □

Claim 9. $|N_i - mp_i| \leq \max\left\{4\sqrt{mp_i \log n}, \log n\right\}$ *simultaneously for all* $i \in \mathcal{A}$ *with probability at least* $1 - \frac{2}{n^{0.84}} - \frac{1.1}{n}$.

Proof. We consider this in two cases. Let X be a $Poisson(\lambda)$ random variable. First, assume that $\lambda \geq e^{-3} \log n$. By Bennett's inequality, we have the following tail bound [Pol15, Can17]:

$$\Pr\left[|X-\lambda| \geq x\right] \leq 2\exp\left(-\frac{x^2}{2\lambda}\psi\left(\frac{x}{\lambda}\right)\right),$$

where

$$\psi(t) = \frac{(1+t)\log(1+t) - t}{t^2/2}.$$

Consider $x = 4\sqrt{\lambda \log n}$. At this point, we have

$$\psi(x/\lambda) = \psi(4\sqrt{\log n/\lambda}) \geq \psi(4e^{3/2}) \geq 0.23.$$

Thus,

$$\begin{aligned}\Pr\left[|X-\lambda| \geq 4\sqrt{\lambda \log n}\right] &\leq 2\exp\left(-0.23 \cdot 8 \log n\right) \\ &\leq 2n^{-1.84}.\end{aligned}$$

Now, we focus on the other case, where $\lambda \leq e^{-3} \log n$. Here, we appeal to Proposition 1 of [Kla00], which implies the following via Stirling's approximation:

$$\Pr\left[|X-\lambda| \geq k\lambda\right] \leq \frac{k}{k-1}\exp(-\lambda + k\lambda - k\lambda \log k).$$

We set $k\lambda = \log n$, giving the upper bound

$$\frac{k}{k-1} n^{1-\log k} \leq 1.1 \cdot n^{-2}.$$

We conclude by taking a union bound over $[n]$, with the argument for each $i \in [n]$

depending on whether $\lambda = mp_i$ is large or small. $\square$

We proceed with proving the two desiderata of this algorithm, correctness and privacy.

Correctness. We use the following two properties of the statistic $Z(D)$, which rely on the condition that $m = \Omega(\sqrt{n}/\alpha^2)$. The proofs of these properties are identical to the proofs of Lemma 2 and 3 in [ADK15], and are omitted.

Claim 10. *If $p = q$, then $\mathbf{E}[Z] = 0$. If $d_{\mathrm{TV}}(p, q) \geq \alpha$, then $\mathbf{E}[Z] \geq 1$.*

Claim 11. *If $p = q$, then $\mathbf{Var}[Z] \leq 1/1000$. If $d_{\mathrm{TV}}(p, q) \geq \alpha$, then $\mathbf{Var}[Z] \leq 1/1000 \cdot \mathbf{E}[Z]^2$.*

First, we note that, by Claim 8, the probability that we return in line 5 is exactly c_2. We now consider the case where $p = q$. We note that by Claim 9, the probability that we output "$p \neq q$" in line 10 is $o(1)$, and thus negligible. By Chebyshev's inequality, we get that $Z \leq 1/10$ with probability at least $9/10$, and we output "$p = q$" with probability at least $c_2/2 + (1 - c_2) \cdot (9/10 - c_2)^2 \geq 2/3$ (note that we subtract c_2 from $9/10$ since we are conditioning on an event with probability $1 - c_2$, and by union bound). Similarly, when $d_{\mathrm{TV}}(p, q) \geq \alpha$, Chebyshev's inequality gives that $Z \geq 9/10$ with probability at least $9/10$, and therefore we output "$p \neq q$" with probability at least $2/3$.

Privacy. We will prove $(0, c_2\varepsilon/2)$-differential privacy. By Claim 8, the probability that we return in line 5 is exactly c_2. Thus the minimum probability of any output of the algorithm is at least $c_2/2$, and therefore $(0, c_2\varepsilon/2)$-differential privacy implies $(\varepsilon, 0)$-differential privacy.

We first consider the possibility of rejecting in line 11. Consider two neighboring datasets D and D', which differ by 1 in the frequency of symbol i. Coupling the randomness of the Y_j's on these two datasets, the only case in which the output differs is when Y_i is such that the value of $|N_i + Y_i - mq_i|$ lies on opposite sides of the threshold for the two datasets. Since N_i differs by 1 in the two datasets, and the probability mass assigned by the PDF of Y_i to any interval of length 1 is at most $c_2\varepsilon/4$, the probability that the outputs differ is at most $c_2\varepsilon/4$. Therefore, this step is $(0, c_2\varepsilon/4)$-differentially private.

We next consider the value of Z for two neighboring datasets D and D', where D' has one fewer occurrence of symbol i. We only consider the case where we have not already

returned in line 11, as otherwise the value of Z is irrelevant for determining the output of the algorithm.

$$\begin{aligned}
&Z(D) - Z(D') \\
&= \frac{1}{m\alpha^2}\left[\frac{(N_i - mq_i)^2 - N_i}{mq_i} - \frac{(N_i - 1 - mq_i)^2 - (N_i - 1)}{mq_i}\right] \\
&= \frac{1}{m\alpha^2}\left[\frac{(N_i - mq_i)^2 - N_i}{mq_i} - \frac{(N_i - mq_i)^2 - 2(N_i - mq_i) + 1 - N_i + 1}{mq_i}\right] \\
&= \frac{2(N_i - mq_i - 1)}{m^2\alpha^2 q_i}.
\end{aligned}$$

Since we did not return in line 11,

$$\begin{aligned}
|N_i - mq_i| &\le \frac{4}{c_2\varepsilon}\log\left(\frac{1}{1-(1-c_2)^{1/n}}\right) + \max\left\{4\sqrt{mq_i\log n}, \log n\right\} \\
&\le \frac{4\log(n/c_2)}{c_2\varepsilon} + \max\left\{4\sqrt{mq_i\log n}, \log n\right\}
\end{aligned}$$

This implies that

$$\begin{aligned}
|Z(D) - Z(D')| &= \frac{2|N_i - mq_i - 1|}{m^2\alpha^2 q_i} \\
&\le \frac{2}{m^2\alpha^2 q_i}\left(\frac{6\log(n/c_2)}{c_2\varepsilon} + 4\sqrt{mq_i\log n}\right).
\end{aligned}$$

We will enforce that each of these terms are at most $c_2\varepsilon/8$.

$$\begin{aligned}
\frac{12\log(n/c_2)}{m^2\alpha^2 q_i c_2\varepsilon} \le \frac{c_2\varepsilon}{8} &\Rightarrow m \ge \sqrt{\frac{96}{c_2^2 c_1}}\frac{\sqrt{n\log(n/c_2)}}{\alpha^{1.5}\varepsilon} \\
\frac{8\sqrt{\log n}}{m^{1.5}\alpha^2\sqrt{q_i}} \le \frac{c_2\varepsilon}{8} &\Rightarrow m \ge \left(\frac{64}{c_2\sqrt{c_1}}\right)^{2/3}\frac{(n\log n)^{1/3}}{\alpha^{5/3}\varepsilon^{2/3}}
\end{aligned}$$

Since both terms are at most $c_2\varepsilon/8$, this step is $(0, c_2\varepsilon/4)$-differentially private. Combining with the previous step gives the desired $(0, c_2\varepsilon/2)$-differential privacy, and thus (as argued at the beginning of the privacy section of this proof) ε-pure differential privacy. $\square$

5.4 INSPECTRE: Private Property Estimation

In this section, we prove our results for support coverage in Section 5.4.1, support size in Section 5.4.2, distance to uniformity in Section 5.4.3, and entropy in Section 5.4.4. In each section, we first describe and analyze our algorithms for the relevant problem. We then go on to describe and analyze a lower bound construction, showing that our upper bounds are almost tight.

All our algorithms fall into the following simple framework:

1. Compute a non-private estimate of the property;
2. Privatize this estimate by adding Laplace noise, where the parameter is determined through analysis of the estimator and potentially computation of the estimator's sensitivity.

5.4.1 Support Coverage Estimation

In this section, we prove Theorem 44, about support coverage estimation:

Theorem 44. *The sample complexity of support coverage estimation is*

$$C(S_k, \alpha, \varepsilon) = \begin{cases} O\left(\frac{k\log(1/\alpha)}{\log k} + \frac{k\log(1/\alpha)}{\log(2+\varepsilon k)}\right), & \text{when } k \geq \frac{1}{\alpha\varepsilon} \\ O\left(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right), & \text{when } \frac{1}{\alpha} \leq k \leq \frac{1}{\alpha\varepsilon} \\ O\left(k^2 + \frac{k}{\varepsilon}\right). & \text{when } k \leq \frac{1}{\alpha} \end{cases}$$

Furthermore,

$$C(S_k, \alpha, \varepsilon) = \Omega\left(\frac{k\log(1/\alpha)}{\log k} + \frac{1}{\alpha\varepsilon}\right).$$

Our upper bound is analyzed in Section 5.4.1.1, while our lower bound is proved in Section 5.4.1.2.

5.4.1.1 Upper Bound for Support Coverage Estimation

We split the analysis into two regimes. First, we focus on the case where $k \leq \frac{1}{\alpha\varepsilon}$, and we prove the upper bound $O\left(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)$. Note that the problem is identical for any $\alpha < \frac{1}{k}$, since

this corresponds to estimating the support coverage exactly, and the above bound simplifies to $O\left(k^2 + \frac{k}{\varepsilon}\right)$. The algorithm in this case is simple: since $m = \Omega(k)$, we group the dataset into m/k batches of size k. Let Y_j be the number of unique symbols observed in batch j. Our estimator is

$$\hat{S}_k(X_1^m) = \frac{k}{m} \sum_{j=1}^{m/k} Y_j.$$

Observe that $\mathbf{E}[Y_j] = S_k(p)$, and that $\mathbf{Var}[Y_j] \leq k$. The latter can be seen by observing that Y_j is the sum of k negatively correlated indicator random variables, each one being the indicator of whether that sample in the batch is the first time the symbol is observed. This gives that $\hat{S}_k(X_1^m)$ is an unbiased estimator of $S_k(p)$, with variance $O(k^2/m)$. By Chebyshev's inequality, since we want an estimate which is accurate up to $\pm\alpha k$, this gives us that $C_{\hat{S}_k}(S_k(p), \alpha/2) = O\left(\frac{1}{\alpha^2}\right)$. Furthermore, we can see that the sensitivity of $\hat{S}_k(X_1^m)$ is at most $2k/m$. By Lemma 41, there is a private algorithm for support coverage estimation as long as

$$\Delta\left(\frac{\hat{S}_k(X_1^m)}{k}\right) \leq \alpha\varepsilon.$$

With the above bound on sensitivity, this is true with $m = O(1/\alpha\varepsilon)$, giving the desired upper bound.

Now, we turn our attention to the case where $k \geq \frac{1}{\alpha\varepsilon}$, and we prove the upper bound $O\left(\frac{k\log(1/\alpha)}{\log k} + \frac{k\log(1/\alpha)}{\log(2+\varepsilon k)}\right)$. Let φ_i be the number of symbols that appear i times in X_1^m. We will use the following non-private support coverage estimator from [OSW16]:

$$\hat{S}_k(X_1^m) = \sum_{i=1}^{m} \varphi_i \left(1 - (-t)^i \cdot \Pr[\, Z \geq i \,]\right),$$

where Z is a Poisson random variable with mean r (which is a parameter to be instantiated later), and $t = (k-m)/m$.

Our private estimator of support coverage is derived by adding Laplace noise to this non-private estimator with the appropriate noise parameter, and thus the performance of our private estimator, is analyzed by bounding the sensitivity and the bias of this non-private estimator according to Lemma 41.

The sensitivity and bias of this estimator is bounded in the following lemmas.

Lemma 43. *Suppose $k > 2m$, then the maximum coefficient of φ_i in $\hat{S}_k(p)$ is at most $1 + e^{r(t-1)}$.*

Proof. By the definition of Z, we know $\Pr[\, Z \geq i \,] = \sum_{j=i}^{\infty} e^{-r} \frac{r^j}{j!}$, hence we have:

$$\begin{aligned} |1 + (-t)^i \cdot \Pr[\, Z \geq i \,]\,| &\leq 1 + t^i \sum_{j=i}^{\infty} e^{-r} \frac{r^j}{j!} \\ &\leq 1 + e^{-r} \sum_{j=i}^{\infty} \frac{(rt)^j}{j!} \\ &\leq 1 + e^{-r} \sum_{j=0}^{\infty} \frac{(rt)^j}{j!} \\ &= 1 + e^{r(t-1)} \end{aligned}$$

□

The bias of the estimator is bounded in Lemma 4 of [ADOS17]:

Lemma 44. *Suppose $k > 2m$, then*

$$\left| \mathbf{E}\left[\hat{S}_k(X_1^m) \right] - S_k(p) \right| \leq 2 + 2e^{r(t-1)} + \min(k, S(p)) \cdot e^{-r}.$$

Using these results, letting $r = \log(1/\alpha)$, [OSW16] showed that there is a constant C, such that with $m = C \frac{k}{\log k} \log(1/\alpha)$ samples, with probability at least 0.9,

$$\left| \frac{\hat{S}_k(X_1^m)}{k} - \frac{S_k(p)}{k} \right| \leq \alpha.$$

Our upper bound in Theorem 44 is derived by the following analysis of the sensitivity of $\frac{\hat{S}_k(X_1^m)}{k}$.

If we change one sample in X_1^m, at most two of the φ_j's change. Hence by Lemma 43,

the sensitivity of the estimator satisfies

$$\Delta\left(\frac{\hat{S}_k(X_1^m)}{k}\right) \leq \frac{2}{k}\cdot\left(1+e^{r(t-1)}\right). \tag{5.1}$$

By Lemma 41, there is a private algorithm for support coverage estimation as long as

$$\Delta\left(\frac{\hat{S}_k(X_1^m)}{k}\right) \leq \alpha\varepsilon,$$

which by (5.1) holds if

$$2(1+\exp(r(t-1))) \leq \alpha\varepsilon k.$$

Let $r = \log(3/\alpha)$, note that $t-1 = \frac{k}{m} - 2$. Suppose $\alpha\varepsilon k > 2$, then, the condition above reduces to

$$\log\left(\frac{3}{\alpha}\right)\cdot\left(\frac{k}{m}-2\right) \leq \log\left(\frac{1}{2}\alpha\varepsilon k - 1\right).$$

This is equivalent to

$$\begin{aligned} m &\geq \frac{k\log(3/\alpha)}{\log(\frac{1}{2}\alpha\varepsilon k-1)+2\log(3/\alpha)} \\ &= \frac{k\log(3/\alpha)}{\log(\frac{3}{2}\varepsilon k - 3/\alpha)+\log(3/\alpha)} \end{aligned}$$

Suppose $\alpha\varepsilon k > 2$, then the condition above reduces to the requirement that

$$m = \Omega\left(\frac{k\log(1/\alpha)}{\log(2+\varepsilon k)}\right).$$

5.4.1.2 Lower Bound for Support Coverage Estimation

We now prove the lower bound described in Theorem 44. Note that the first term in the lower bound is the sample complexity of non-private support coverage estimation, shown in [OSW16]. Therefore, we turn our attention to prove the last term in the sample complexity.

Consider the following two distributions. u_1 is uniform over $[k(1+\alpha)]$. u_2 is distributed over $k+1$ elements $[k]\cup\{\triangle\}$ where $u_2[i] = \frac{1}{k(1+\alpha)}\forall i \in [k]$ and $u_2[\triangle] = \frac{\alpha}{1+\alpha}$. Moreover, $\triangle \notin [k(1+\alpha)]$. Then,

$$S_k(u_1) = k(1+\alpha) \cdot \left(1 - \left(1 - \frac{1}{k(1+\alpha)}\right)^k\right),$$

and

$$S_k(u_2) = k \cdot \left(1 - \left(1 - \frac{1}{k(1+\alpha)}\right)^k\right) + \left(1 - \left(1 - \frac{\alpha}{1+\alpha}\right)^k\right)$$

hence,

$$\begin{aligned}
&S_k(u_2) - S_k(u_1) \\
&= k\alpha \cdot \left(1 - \left(1 - \frac{1}{k(1+\alpha)}\right)^k\right) - \left(1 - \left(1 - \frac{\alpha}{1+\alpha}\right)^k\right) \\
&= \Omega(\alpha k)
\end{aligned}$$

Hence we know there support coverage differs by $\Omega(\alpha k)$. Moreover, their total variation distance is $\frac{\alpha}{1+\alpha}$. The following lemma is folklore, based on the coupling interpretation of total variation distance, and the fact that total variation distance is subadditive for product measures.

Lemma 45. *For any two distributions p, and q, there is a coupling between m i.i.d. samples from the two distributions with an expected Hamming distance of $d_{\mathrm{TV}}(p,q) \cdot m$.*

Using Lemma 45 and $d_{\mathrm{TV}}(u_1, u_2) = \frac{\alpha}{1+\alpha}$, we have

Lemma 46. *Suppose u_1 and u_2 are as defined before, there is a coupling between u_1^m and u_2^m with expected Hamming distance equal to $\frac{\alpha}{1+\alpha}m$.*

Moreover, given m samples, we must be able to privately distinguish between u_1 and u_2 given an α accurate estimator of support coverage with privacy considerations. Thus, according to Lemma 42 and 46, we have:

$$\frac{\alpha}{1+\alpha}m \geq \frac{1}{\varepsilon} \Rightarrow m = \Omega\left(\frac{1}{\varepsilon\alpha}\right).$$

5.4.2 Support Size Estimation

In this section, we prove our main theorem about support size estimation, Theorem 45:

Theorem 45. *The sample complexity of support size estimation is*

$$C(S,\alpha,\varepsilon) = \begin{cases} O\left(\frac{n\log^2(1/\alpha)}{\log n} + \frac{n\log^2(1/\alpha)}{\log(2+\varepsilon n)}\right), & \text{when } n \geq \frac{1}{\alpha\varepsilon} \\ O\left(n\log(1/\alpha) + \frac{1}{\alpha\varepsilon}\right), & \text{when } \frac{1}{\alpha} \leq n \leq \frac{1}{\alpha\varepsilon} \\ O\left(n\log n + \frac{n}{\varepsilon}\right). & \text{when } n \leq \frac{1}{\alpha} \end{cases}$$

Furthermore,

$$C(S,\alpha,\varepsilon) = \begin{cases} \Omega\left(\frac{n\log^2(1/\alpha)}{\log n} + \frac{1}{\alpha\varepsilon}\right), & \text{when } n \geq \frac{1}{\alpha} \\ \Omega\left(n\log n + \frac{n}{\varepsilon}\right). & \text{when } n \leq \frac{1}{\alpha} \end{cases}$$

Our upper bound is described and analyzed in Section 5.4.2.1, while our lower bound appears in Section 5.4.2.2.

5.4.2.1 Upper Bound for Support Size Estimation

We split the analysis into two regimes. First we consider the "sparse" case, where the amount of data is relatively small. In particular, $m < \frac{n\log\frac{3}{\alpha}}{2}$. In this case we show a bound of $O\left(\frac{n\log^2(1/\alpha)}{\log n} + \frac{n\log^2(1/\alpha)}{\log(2+\varepsilon n)}\right)$. This upper bound is less than $\frac{n\log\frac{3}{\alpha}}{2}$ only when $n = \Omega\left(\frac{1}{\alpha\varepsilon}\right)$, which is the condition for the sparse case.

Sparse case In [OSW16], it is shown that the support coverage estimator can be used to obtain optimal results for estimating the support size of a distribution. In this fashion, taking $k = n\log(3/\alpha)$, we may use an estimator of the support coverage $S_k(p)$ as an estimator of $S(p)$. In particular, their result is based on the following observation.

Lemma 47. *Suppose $k \geq n\log(3/\alpha)$, then for any $p \in \Delta_{\geq\frac{1}{n}}$,*

$$|S_k(p) - S(p)| \leq \frac{\alpha n}{3}.$$

Proof. From the definition of $S_k(p)$, we have $S_k(p) \leq S(p)$. For the other side,

$$\begin{aligned} S(p) - S_k(p) &= \sum_x (1 - p(x))^k \leq \sum_x e^{-kp(x)} \\ &\leq n \cdot e^{-\log(3/\alpha)} \qquad = \frac{n\alpha}{3}. \end{aligned}$$

□

Therefore, estimating $S_k(p)$ for $k = n\log(3/\alpha)$, up to $\pm\alpha n/3$. Therefore, the goal is to determine the smallest value of m to solve the support coverage problem for $k = n\log(3/\alpha)$.

Suppose $r = \log(3/\alpha)$, and $k = n\log(3/\alpha) = n \cdot r$ in the support coverage problem. Then, we have

$$t = \frac{k}{m} - 1 = \frac{n\log(3/\alpha)}{m} - 1. \tag{5.2}$$

Then, by Lemma 44 in the previous section, we have

$$\begin{aligned} &\left|\mathbf{E}\left[\hat{S}_k(X_1^m)\right] - S(p)\right| \\ &\leq \left|\mathbf{E}\left[\hat{S}_k(X_1^m)\right] - S_k(p)\right| + |S_k(p) - S(p)| \\ &\leq 2 + 2e^{r(t-1)} + \min\{k, n\} \cdot e^{-r} + \frac{n\alpha}{3} \\ &\leq 2 + 2e^{r(t-1)} + n \cdot e^{-\log(3/\alpha)} + \frac{n\alpha}{3} \\ &\leq 2 + 2e^{r(t-1)} + 2\frac{n\alpha}{3}. \end{aligned}$$

We will find conditions on m such that the middle term above is at most $n\alpha$. Toward this end, note that $2e^{r(t-1)} \leq \alpha n$ holds if and only if $r(t-1) \leq \log\left(\frac{\alpha n}{2}\right)$. Plugging in (5.2), this holds when

$$\log(3/\alpha) \cdot \left(\frac{n\log(3/\alpha)}{m} - 2\right) \leq \log\left(\frac{\alpha n}{2}\right),$$

which is equivalent to

$$m \geq \frac{n \log^2(3/\alpha)}{\log \frac{\alpha n}{2} + 2 \log \frac{3}{\alpha}} = O\left(\frac{n \log^2(1/\alpha)}{\log n}\right)$$

where we have assumed without loss of generality that $\alpha > \frac{1}{n}$.

The computations for sensitivity are very similar. From Lemma 41, we need to find the value of m such that

$$2 + 2e^{r(t-1)} \leq \alpha \varepsilon n,$$

where we assume that $m \leq \frac{1}{2} n \log(3/\alpha)$, else we just add noise to the true number of observed distinct elements. By computations similar to the previous case, this reduces to

$$m \geq \frac{n \log^2(3/\alpha)}{\log \frac{\alpha \varepsilon n}{2} + \log \frac{3}{\alpha}}.$$

Therefore, this gives us a sample complexity of

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

for the sensitivity result to hold.

Dense case Then let us consider the dense case when $n \leq \frac{1}{\alpha\varepsilon}$. The algorithm under this case will be the following. Let $W(X_1^m)$ denote the set of symbols which appear in X_1^m and let N_x denote the number of times x appears, then our non-private estimator is

$$\hat{S}(X_1^m) = \sum_{x \in W(X_1^m)} \min\left\{1, \frac{N_x}{\frac{m}{3n}}\right\}.$$

To analyze the performance of the algorithm, we consider two cases, the case when $n \leq \frac{1}{\alpha}$ and the case when $\frac{1}{\alpha} \leq n \leq \frac{1}{\alpha\varepsilon}$.

When $n \leq \frac{1}{\alpha}$, we have $n\alpha < 1$, which means we need to know the exact support size. Our algorithm gives correct answer when all the symbols appearing at least $\frac{m}{3n}$ times. For any symbol x with $p(x) \geq \frac{1}{n}$, according to the Chernoff bound, $\Pr\left[N_x < \frac{m}{3n}\right] \leq \exp(-\frac{\frac{2m^2}{9n^2}}{m \cdot \frac{1}{n}}) = \exp(-\frac{2m}{9n})$. Let $m \geq 18n \log n$, we have $\Pr\left[N_x < \frac{m}{3n}\right] \leq \frac{1}{n^4}$. Then according to the union

bound, the probability of all the symbols appearing at least $\frac{m}{3n}$ is greater than $1 - \frac{1}{n^3}$. When $n \geq 2$, this is larger than $2/3$, which means our algorithm gives correct answer with probability more than $\frac{2}{3}$.

Furthermore, we can see that the sensitivity of $\hat{S}(X_1^m)$ is at most $3n/m$. By Lemma 41, there is a private algorithm for support size estimation as long as

$$\Delta\left(\hat{S}(X_1^m)\right) \leq \varepsilon.$$

With the above bound on sensitivity, this is true with $m = O(n/\varepsilon)$, giving the desired upper bound.

Next we consider the case when $\frac{1}{\alpha} \leq n \leq \frac{1}{\alpha\varepsilon}$. For any symbol x with $p(x) \geq \frac{1}{n}$, according to the same argument, $\Pr\left[\, N_x < \frac{m}{3n} \,\right] \leq \exp(-\frac{\frac{2m^2}{9n^2}}{m \cdot \frac{1}{n}}) = \exp(-\frac{2m}{9n})$. When $m \geq 9n \log(1/\alpha)$, we have $\Pr\left[\, N_x < \frac{m}{3n} \,\right] \leq \alpha^2 \leq 0.5\alpha$ if we suppose $\alpha < 0.5$. Let $Y(X_1^m) \triangleq \sum_{x \in S(p)} \mathbf{1}\{N_x \geq \frac{m}{3n}\}$, which is the number of symbols appearing more than $\frac{m}{3n}$ times. We know that $\mathbf{E}\left[Y(X_1^m)\right] > S(p)(1 - 0.5\alpha)$ by linearity of expectations. Moreover, $\mathbf{Var}\left[Y(X_1^m)\right] < 0.5\alpha \cdot S(p)$ since it is the sum of $S(p)$ negatively related Bernoulli random variables with bias less than 0.5α. According to Chebyshev's inequality,

$$\Pr[\,(1 - 0.5\alpha)S(p) < Y(X_1^m) < S(p) + \alpha S(p)\,] \geq 1 - \frac{1}{4.5\alpha S(p)} \geq 1 - \frac{1}{4.5n\alpha} \geq \frac{2}{3},$$

where the last inequality comes from the fact $n\alpha \geq 1$. Therefore,

$$\Pr[\,(S(p) - \alpha n < Y(X_1^m) < S(p) + \alpha n\,] \geq \Pr[\,(1 - 0.5\alpha)S(p) < Y(X_1^m) < S(p) + \alpha S(p)\,] \geq \frac{2}{3}.$$

Furthermore, we can see that the sensitivity of $\hat{S}(X_1^m)$ is the same, which is at most $3n/m$. By Lemma 41, there is a private algorithm for support coverage estimation as long as

$$\Delta\left(\hat{S}(X_1^m)\right) \leq n\alpha\varepsilon.$$

With the above bound on sensitivity, this is true with $m = O(\frac{1}{\alpha\varepsilon})$, giving the desired upper bound.

5.4.2.2 Lower Bound for Support Size Estimation

In this section, we prove a lower bound for support size estimation, as described in Theorem 45. The techniques are similar to those for support coverage in Section 5.4.1.2.

First let us focus on the case when $n \geq \frac{1}{\alpha}$, The first term of the complexity is the lower bounds for the non-private setting, which follows by combining the lower bound of [OSW16] for support coverage, with the equivalence between estimation of support size and coverage as implied by Lemma 47. We focus on the final term in the sequel.

Consider the following two distributions: u_1 is a uniform distribution over $[n]$ and u_2 is a uniform distribution over $[(1-\alpha)n]$. Then the support size of these two distribution differs by αn, and $d_{\mathrm{TV}}(u_1, u_2) = \alpha$.

Hence by Lemma 45, we know the following:

Lemma 48. *Suppose $u_1 \sim \mathcal{U}_n$ and $u_2 \sim \mathcal{U}_{(1-\alpha)n}$, there is a coupling between u_1^m and u_2^m with expected Hamming distance equal to αm.*

Moreover, given m samples, we must be able to privately distinguish between u_1 and u_2 given an α accurate estimator of entropy with privacy considerations. Thus, according to Lemma 42 and Lemma 48, we have:

$$\alpha m \geq \frac{1}{\varepsilon} \Rightarrow m = \Omega\left(\frac{1}{\varepsilon\alpha}\right).$$

Then we move to the second case when $n \leq \frac{1}{\alpha}$. Because $n\alpha < 1$, we need to recover the support size exactly. The first term of the complexity is the lower bound for the non-private setting which can be proved using a coupon collector style argument, so here we focus on the second term.

We consider the following two distributions: u_1 is a uniform distribution over $[n]$ and u_2 is a uniform distribution over $[n-1]$. We must distinguish between these two distributions, for which $d_{\mathrm{TV}}(u_1, u_2) = \frac{1}{n}$. Hence, by Lemma 45, we have

$$\frac{m}{n} \geq \frac{1}{\varepsilon} \Rightarrow m = \Omega\left(\frac{n}{\varepsilon}\right).$$

5.4.3 Distance to Uniformity Estimation

In this section, we prove our main theorem about estimating distance to uniformity, Theorem 46.

Theorem 46. *Let $\lambda > 0$ be* any *small fixed constant. For instance, λ can be chosen to be any constant between* 0.01 *and* 1*. We have the following upper bounds on the sample complexity of estimating distance to uniformity:*

$$C(\|p - \mathcal{U}_n\|_1, \alpha, \varepsilon) = O\left(\frac{n}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)$$

and

$$C(\|p - \mathcal{U}_n\|_1, \alpha, \varepsilon) = O\left(\frac{n}{\lambda^2\alpha^2 \log n} + \left(\frac{1}{\alpha\varepsilon}\right)^{1+\lambda}\right).$$

Furthermore,

$$C(\|p - \mathcal{U}_n\|_1, \alpha, \varepsilon) = \Omega\left(\frac{n}{\alpha^2 \log n} + \frac{n^{1/2}}{\alpha\varepsilon^{1/2}} + \frac{n^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}} + \frac{1}{\alpha\varepsilon}\right).$$

We describe and analyze two upper bounds. The first is based on the empirical estimator, and is described and analyzed in Section 5.4.3.1. The second is based on the method of best-polynomial approximation, and appears in Section 5.4.3.2. Finally, our lower bound is in Section 5.4.3.3.

5.4.3.1 Upper Bound for Estimating Distance to Uniformity: The Empirical Estimator

Our first private distance to uniformity estimator is based on adding Laplace noise into the empirical estimator. The parameter of the Laplace noise is dependent on the sensitivity of the empirical estimator. By analyzing its sensitivity and bias, we prove the first upper bound in Theorem 46,

Let $\hat{p}_m$ be the empirical distribution, and let $\|\hat{p}_m - \mathcal{U}_n\|_1$ be the distance to uniformity

of the empirical distribution. The theorem is based on the following three facts:

$$\Delta(\|\hat{p}_m - \mathcal{U}_n\|_1) = O\left(\min\left\{\frac{1}{n}, \frac{1}{m}\right\}\right), \tag{5.3}$$

$$\mathbf{E}\left[\|\hat{p}_m - \mathcal{U}_n\|_1\right] - \|p - \mathcal{U}_n\|_1 = O\left(\sqrt{\frac{n}{m}}\right), \tag{5.4}$$

$$\mathbf{Var}\left[\|\hat{p}_m - \mathcal{U}_n\|_1\right] = O\left(\frac{1}{m}\right). \tag{5.5}$$

With these three facts in hand, the sample complexity of the empirical estimator can be bounded as follows. By Lemma 41, we need $\Delta(\|\hat{p}_m - \mathcal{U}_n\|_1) \leq \alpha\varepsilon$, which gives $m = O\left(\frac{1}{\alpha\varepsilon}\right)$. We also need $\mathbf{E}\left[\|\hat{p}_m - \mathcal{U}_n\|_1\right] - \|p - \mathcal{U}_n\|_1 = O(\alpha)$ and $\mathbf{Var}\left[\|\hat{p}_m - \mathcal{U}_n\|_1\right] = O(\alpha^2)$, which gives $m = O\left(\frac{n}{\alpha^2}\right)$.

Proof of (5.3)**.** The largest change in any N_x when we change one symbol is one. Moreover, at most two N_x change. Clearly we have $\Delta(\|\hat{p}_m - \mathcal{U}_n\|_1) \leq \frac{2}{m}$.

Then suppose $n \geq m$, we use ϕ_i to denote the number of symbols which appear i times.

$$\begin{aligned}
\|\hat{p}_m - \mathcal{U}_n\|_1 &= \frac{\phi_0}{n} + \sum_{i=1}^{m} \phi_i \cdot \left(\frac{i}{m} - \frac{1}{n}\right) \\
&= \sum_{i=1}^{m} \frac{\phi_i \cdot i}{m} - \sum_{i=1}^{m} \frac{\phi_i}{n} + \phi_0 \cdot \frac{1}{n} \\
&= \frac{2}{n}\phi_0
\end{aligned}$$

The last equality comes from $\sum_{i=1}^{m} \phi_i = n - \phi_0$ and $\sum_{i=1}^{m} \phi_i \cdot i = m$.

The largest change in ϕ_0 when we change one symbol is one. Therefore when $n \geq m$, $\Delta(\|\hat{p}_m - \mathcal{U}_n\|_1) \leq \frac{2}{n}$.

Proof of (5.4)**.**

$$\begin{aligned}
\mathbf{E}\left[\|\hat{p}_m - \mathcal{U}_n\|_1\right] &= \mathbf{E}\left[\sum_{i=1}^{n}\left|\hat{p}_i - \frac{1}{n}\right|\right] \\
&\leq \sum_{i=1}^{n}\mathbf{E}\left[|\hat{p}_i - p_i|\right] + \sum_{i=1}^{n}\left|p_i - \frac{1}{n}\right| \\
&\leq \sum_{i=1}^{n}\mathbf{E}\left[|\hat{p}_i - p_i|\right] + \|p - \mathcal{U}_n\|_1 \\
&\leq \sqrt{n}\cdot\sum_{i=1}^{n}\mathbf{E}\left[\|\hat{p}_i - p_i\|_2\right] + \|p - \mathcal{U}_n\|_1 \qquad (5.6) \\
&\leq \sqrt{\frac{n}{m}} + \|p - \mathcal{U}_n\|_1 \qquad (5.7)
\end{aligned}$$

Equation (5.6) comes from Cauchy-Schwarz inequality.

Proof of (5.5)**.** We apply the bounded differences inequality in the form stated in Corollary 3.2 of [BLM13].

Lemma 49. *Let $f : \Omega^m \to \mathbb{R}$ be a function. Suppose further that*

$$\max_{z_1,\ldots,z_m,z_i^{'}}\left|f(z_1,\ldots,z_m) - f(z_1,\ldots,z_{i-1},z_i^{'},\ldots,z_m)\right| \leq c_i.$$

Then for independent variables $Z_1,\ldots,Z_m$,

$$\mathbf{Var}\left(f(Z_1,\ldots,Z_m)\right) \leq \frac{1}{4}\sum_{i=1}^{m}c_i^2.$$

By Lemma 49 and Equation (5.3), we have

$$\mathbf{Var}\left[\|\hat{p}_m - \mathcal{U}_n\|_1\right] \leq m\cdot\frac{1}{m^2} \leq \frac{1}{m}$$

5.4.3.2 Upper Bound for Estimating Distance to Uniformity: Best-Polynomial Approximation

We prove an upper bound on the sample complexity if one adds Laplace noise to the best-polynomial estimator. This will give us the second upper bound in Theorem 46. We use the algorithm of [ADOS17]. This estimator has the order-optimal sample complexity, but smaller sensitivity in comparison to previous estimators.

Lemma 50 (Lemma 7 of [ADOS17])**.** *Let $\lambda > 0$ be a fixed small constant, which may be taken to be any value between* 0.01 *and* 1*. Then there is an estimator with sample complexity*

$$O\left(\frac{1}{\lambda^2} \cdot \frac{n}{\alpha^2 \log n}\right),$$

and has sensitivity m^{λ}/m.

We can now invoke Lemma 41 on the estimator in this lemma to obtain the second upper bound on private entropy estimation.

5.4.3.3 Lower Bound for Estimating Distance to Uniformity

We now prove a lower bound for estimating distance to uniformity. The first term in the lower bound of Theorem 46 comes from lower bounds for non-private estimation (see, i.e, [JHW16]).

Note that estimating distance to uniformity is a harder problem than uniformity testing, which tests whether p is either uniform distribution or α-far away from it. According to the private uniformity testing lower bound given by Theorem 13 in [ASZ17],

$$\Theta\left(\frac{n}{\alpha^2 \log n} + \max\left\{\frac{n^{1/2}}{\alpha\varepsilon^{1/2}}, \frac{n^{1/3}}{\alpha^{4/3}\varepsilon^{2/3}}, \frac{1}{\alpha\varepsilon}\right\}\right)$$

We get the lower bound part on the sample complexity in Theorem 46.

5.4.4 Entropy Estimation

In this section, we prove our main theorem about entropy estimation, Theorem 47:

Theorem 47. *Let $\lambda > 0$ be* any *small fixed constant. For instance, λ can be chosen to be any constant between* 0.01 *and* 1. *We have the following upper bounds on the sample complexity of entropy estimation:*

$$C(H, \alpha, \varepsilon) = O\left(\frac{n}{\alpha} + \frac{\log^2(\min\{n, m\})}{\alpha^2} + \frac{1}{\alpha\varepsilon}\log\left(\frac{1}{\alpha\varepsilon}\right)\right)$$

and

$$C(H, \alpha, \varepsilon) = O\left(\frac{n}{\lambda^2\alpha\log n} + \frac{\log^2(\min\{n, m\})}{\alpha^2} + \left(\frac{1}{\alpha\varepsilon}\right)^{1+\lambda}\right).$$

Furthermore,

$$C(H, \alpha, \varepsilon) = \Omega\left(\frac{n}{\alpha\log n} + \frac{\log^2(\min\{n, m\})}{\alpha^2} + \frac{\log n}{\alpha\varepsilon}\right).$$

We describe and analyze two upper bounds. The first is based on the empirical entropy estimator, and is described and analyzed in Section 5.4.4.1. The second is based on the method of best-polynomial approximation, and appears in Section 5.4.4.2. Finally, our lower bound is in Section 5.4.4.3.

5.4.4.1 Upper Bound for Entropy Estimation: The Empirical Estimator

Our first private entropy estimator is derived by adding Laplace noise into the empirical estimator. The parameter of the Laplace distribution is $\frac{\Delta(H(\hat{p}_m))}{\varepsilon}$, where $\Delta(H(\hat{p}_m))$ denotes the sensitivity of the empirical estimator. By analyzing its sensitivity and bias, we prove an upper bound on the sample complexity for private entropy estimation and get the first upper bound in Theorem 47.

Let $\hat{p}_m$ be the empirical distribution, and let $H(\hat{p}_m)$ be the entropy of the empirical distribution. The theorem is based on the following three facts:

$$\Delta(H(\hat{p}_m)) = O\left(\frac{\log m}{m}\right), \tag{5.8}$$

$$|H(p) - \mathbf{E}[H(\hat{p}_m)]| = O\left(\frac{n}{m}\right), \tag{5.9}$$

$$\mathbf{Var}[H(\hat{p}_m)] = O\left(\frac{\log^2(\min\{n, m\})}{m}\right). \tag{5.10}$$

With these three facts in hand, the sample complexity of the empirical estimator can be

bounded as follows. By Lemma 41, we need $\Delta(H(\hat{p}_m)) \leq \alpha\varepsilon$, which gives $m = O\left(\frac{1}{\alpha\varepsilon}\log(\frac{1}{\alpha\varepsilon})\right)$. We also need $|H(p) - \mathbf{E}[H(\hat{p}_m)]| = O(\alpha)$ and $\mathbf{Var}[H(\hat{p}_m)] = O(\alpha^2)$, which gives $m = O\left(\frac{n}{\alpha} + \frac{\log^2(\min\{n,m\})}{\alpha^2}\right)$.

Proof of (5.8)**.** The largest change in any N_x when we change one symbol is one. Moreover, at most two N_x change. Therefore,

$$\begin{aligned}
\Delta(H(\hat{p}_m)) &\leq 2 \cdot \max_{j=1\ldots m-1} \left| \frac{j+1}{m}\log\frac{m}{j+1} - \frac{j}{m}\log\frac{m}{j} \right| \\
&= 2 \cdot \max_{j=1\ldots m-1} \left| \frac{j}{m}\log\frac{j}{j+1} + \frac{1}{m}\log\frac{m}{j+1} \right| && (5.11) \\
&\leq 2 \cdot \max_{j=1\ldots m-1} \max\left\{ \left| \frac{j}{m}\log\frac{j}{j+1} \right|, \left| \frac{1}{m}\log\frac{m}{j+1} \right| \right\} && (5.12) \\
&\leq 2 \cdot \max\left\{ \frac{1}{m}, \frac{\log m}{m} \right\}, \\
&= 2 \cdot \frac{\log m}{m}. && (5.13)
\end{aligned}$$

Proof of (5.9)**.** By the concavity of entropy function, we know that

$$\mathbf{E}[H(\hat{p}_m)] \leq H(p).$$

Therefore,

$$\begin{aligned}
&\mathbf{E}\left[\left|H\left(p\right)-H\left(\hat{p}_m\right)\right|\right]=H\left(p\right)-\mathbf{E}\left[H\left(\hat{p}_m\right)\right]\\
&=\mathbf{E}\left[\sum_x\left(\hat{p}_n(x)\log\hat{p}_n(x)-p(x)\log p(x)\right)\right]\\
&=\mathbf{E}\left[\sum_x\hat{p}_n(x)\log\frac{\hat{p}_n(x)}{p(x)}\right]+\mathbf{E}\left[\sum_x(\hat{p}_n(x)-p(x))\log p(x)\right]\\
&=\mathbf{E}\left[d_{\mathrm{KL}}\left(\hat{p}_m,p\right)\right] &(5.14)\\
&\leq\mathbf{E}\left[d_{\chi^2}\left(\hat{p}_m,p\right)\right] &(5.15)\\
&=\mathbf{E}\left[\sum_x\frac{(\hat{p}_n(x)-p(x))^2}{p(x)}\right]\\
&\leq\sum_x\frac{(p(x)/m)}{p(x)} &(5.16)\\
&=\frac{n}{m}. &(5.17)
\end{aligned}$$

Proof of (5.10)**.** The variance bound of $\frac{\log^2 n}{m}$ is given precisely in Lemma 15 of [JVHW17]. To obtain the other half of the bound, we use Lemma 49 and Equation (5.8)

$$\mathbf{Var}\left[H(\hat{p}_m)\right]\leq m\cdot\left(\frac{4\log^2 m}{m^2}\right)=\frac{4\log^2 m}{m}.$$

5.4.4.2 Upper Bound for Entropy Estimation: Best-Polynomial Approximation

We prove an upper bound on the sample complexity for private entropy estimation if one adds Laplace noise into best-polynomial estimator.This will give us the second upper bound in Theorem 47.

In the non-private setting the optimal sample complexity of estimating $H(p)$ over Δ_n is given by Theorem 1 of [WY16]

$$\Theta\left(\frac{n}{\alpha\log n}+\frac{\log^2(\min\{n,m\})}{\alpha^2}\right).$$

However, this estimator can have a large sensitivity. [ADOS17] designed an estimator that has the same sample complexity but a smaller sensitivity. We restate Lemma 6 of [ADOS17]

here:

Lemma 51. *Let $\lambda > 0$ be a fixed small constant, which may be taken to be any value between* 0.01 *and* 1. *Then there is an entropy estimator with sample complexity*

$$\Theta\left(\frac{1}{\lambda^2}\cdot\frac{n}{\alpha\log n}+\frac{\log^2(\min\{n,m\})}{\alpha^2}\right),$$

and has sensitivity m^λ/m.

We can now invoke Lemma 41 on the estimator in this lemma to obtain the upper bound on private entropy estimation.

5.4.4.3 Lower Bound for Entropy Estimation

We now prove the lower bound for entropy estimation. Note that any lower bound on privately testing two distributions p, and q such that $H(p) - H(q) = \Theta(\alpha)$ is a lower bound on estimating entropy.

We analyze the following construction for Proposition 2 of [WY16]. The two distributions p, and q over $[n]$ are defined as:

$$p(1) = \frac{2}{3},\ p(i) = \frac{1-p(1)}{n-1}, \text{for } i = 2,\ldots,n, \tag{5.18}$$

$$q(1) = \frac{2-\eta}{3},\ q(i) = \frac{1-q(1)}{n-1}, \text{for } i = 2,\ldots,n. \tag{5.19}$$

Then, by the grouping property of entropy,

$$H(p) = h(2/3) + \frac{1}{3}\cdot\log(n-1),\ \text{and } H(q) = h((2-\eta)/3) + \frac{1+\eta}{3}\cdot\log(n-1),$$

which gives

$$H(p) - H(q) = \Omega(\eta\log n).$$

For $\eta = \alpha/\log n$, the entropy difference becomes $\Theta(\alpha)$.

The total variation distance between p and q is $\eta/3$. By Lemma 45, there is a coupling over X_1^m, and Y_1^m generated from p and q with expected Hamming distance at most $d_{\mathrm{TV}}(p,q)\cdot m$. This along with Lemma 42 gives a lower bound of $\Omega\left(\log n/\alpha\varepsilon\right)$ on the sample complexity.

5.5 Experiments

In this section, we experimentally evaluate our methods for identity testing, entropy estimation, and support coverage on synthetic and real data. For identity testing, privacy seems to be a non-negligible cost, while for the other problems, privacy is very cheap: private estimators achieve accuracy which is comparable or near-indistinguishable to non-private estimators in many settings. Our results on identity testing, entropy estimation, and support coverage appear in Sections 5.5.1, 5.5.2, and 5.5.3, respectively. We present supplementary experimental results without discussion in Section 5.5.4. Code of our implementations are available at `https://github.com/hoonose/privit` and `https://github.com/HuanyuZhang/INSPECTRE`.

5.5.1 Identity Testing

We performed an empirical evaluation of our algorithm, `Priv'IT`, on synthetic datasets. All experiments were performed on a laptop computer with a 2.6 GHz Intel Core i7-6700HQ CPU and 8 GB of RAM. Significant discussion is required to provide a full comparison with prior work in this area, since performance of the algorithms varies depending on the regime.

We compared our algorithm with two recent algorithms for differentially private hypothesis testing:

1. The Monte Carlo Goodness of fit test with Laplace noise from [GLRV16], `MCGOF`;

2. The projected Goodness of Fit test from [KR17], `zCDP-GOF`.

We note that we implemented a modified version of `Priv'IT`, which differs from Algorithm 13 in lines 14 to 21. In particular, we instead consider a statistic

$$Z = \sum_{i \in \mathcal{A}} \frac{(N_i - mq_i)^2 - N_i}{mq_i}.$$

We add Laplace noise to Z, with scale parameter $\Theta(\Delta/\varepsilon)$, where Δ is the sensitivity of Z, which guarantees $(\varepsilon/2, 0)$-differential privacy. Then, similar to the other algorithms, we choose a threshold for this noised statistic such that we have the desired type I error. This

algorithm can be analyzed to provide identical theoretical guarantees as Algorithm 13, but with the practical advantage that there are fewer parameters to tune.

To begin our experimental evaluation, we started with uniformity testing. Our experimental setup was as follows. The algorithms were provided q as the uniform distribution over $[n]$. The algorithms were also provided with samples from some distribution p. This (unknown) p was q for the case $p = q$, or a distribution which we call the "Paninski construction" for the case $d_{\mathrm{TV}}(p, q) \geq \alpha$. The Paninski construction is a distribution where half the elements of the support have mass $(1 + \alpha)/n$ and half have mass $(1 - \alpha)/n$. We use this name for the construction as [Pan08] showed that this example is one of the hardest to distinguish from uniform: one requires $\Omega(\sqrt{n}/\alpha^2)$ samples to (non-privately) distinguish a random permutation of this construction from the uniform distribution. We fixed parameters $\varepsilon = 0.1$ and $\alpha = 0.1$. In addition, recall that Proposition 13 implies that pure differential privacy (the privacy guaranteed by `Priv'IT`) is stronger than zCDP (the privacy guaranteed by `zCDP-GOF`). In particular, our guarantee of ε-pure differential privacy implies $\varepsilon^2/2$-zCDP. As a result, we ran `zCDP-GOF` with a privacy parameter of 0.005-zCDP, which is equivalent to the amount of zCDP our algorithm provides. Our experiments were conducted on a number of different support sizes n, ranging from 10 to 10600. For each n, we ran the testing algorithms with increasing sample sizes m in order to discover the minimum sample size when the type I and type II errors were both empirically below 1/3. To determine these empirical error rates, we ran all algorithms 1000 times for each n and m, and recorded the fraction of the time each algorithm was correct. As the other algorithms take a parameter β_{I} as a target type I error, we input 1/3 as this parameter.

The results of our first test are provided in Figure 5-1. The x-axis indicates the support size, and the y-axis indicates the minimum number of samples required. We plot three lines, which demonstrate the empirical number of samples required to obtain 1/3 type I and type II error for the different algorithms. We can see that in this case, `zCDP-GOF` is the most statistically efficient, followed by `MCGOF` and `Priv'IT`.

To explain this difference in statistical efficiency, we note that the theoretical guarantees of `Priv'IT` imply that it performs well even when data is sparsely sampled. More precisely, one of the benefits of our tester is that it can reduce the variance induced by elements whose

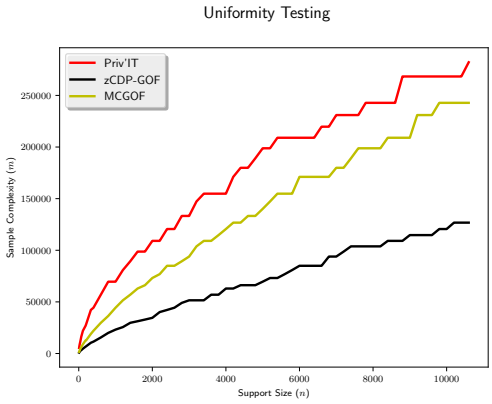


Figure 5-1: The sample complexities of `Priv'IT`, `MCGOF`, and `zCDP-GOF` for uniformity testing

expected number of occurrences is less than 1. Since none of these testers reach this regime (i.e., even `zCDP-GOF` at $n = 10000$ expects to see each element 10 times), we do not reap the benefits of `Priv'IT`. Ideally, we would run these algorithms on the uniform distribution at sufficiently large support sizes. However, since this is prohibitively expensive to do with thousands of repetitions (for any of these methods), we instead demonstrate the advantages of our tester on a different distribution.

Our second test is conducted with q being a 2-histogram[2], where all but a vanishing fraction of the probability mass is concentrated on a small, constant fraction of the support[3]. This serves as our proxy for a very large support, since now we will have elements which have a sub-constant expected number of occurrences. The algorithms are provided with samples from a distribution p, which is either q or a similar Paninski construction as before, where the total variation distance from q is placed on the support elements containing non-negligible mass. We ran the test on support sizes n ranging from 10 to 6800. All other parameters are the same as in the previous test.

The results of our second test are provided in Figure 5-2. In this case, we compare `Priv'IT` and `zCDP-GOF`, and note that our test is slightly better for all support sizes n, though the difference can be pronounced or diminished depending on the construction of the distribution q. We found that `MCGOF` was incredibly inefficient on this construction – even

[2]A k-histogram is a distribution where the domain can be partitioned into k intervals such that the distribution is uniform over each interval.

[3]In particular, in Figure 5-3, $n/200$ support elements contained $1 - 10/n$ probability mass, but similar trends hold with modifications of these parameters.

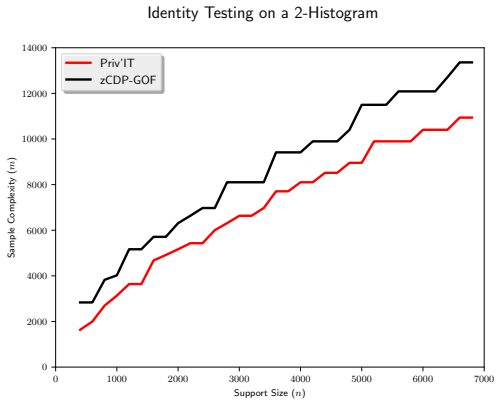


Figure 5-2: The sample complexities of `Priv'IT` and `zCDP-GOF` for identity testing on a 2-histogram

for $n = 400$ it required 130000 samples, which is a factor of 10 worse than `zCDP-GOF` on a support of size $n = 6800$. To explain this phenomenon, we can inspect the contribution of a single domain element i to their statistic:

$$\frac{(N_i + Y_i - mq_i)^2}{mq_i}.$$

In the case where $mq_i \ll 1$ and $p = q$, this is approximately equal to $\frac{Y_i^2}{mq_i}$. The standard deviation of this term will be of the order $\frac{1}{mq_i\varepsilon^2}$, which can be made arbitrarily large as $mq_i \to 0$. While `zCDP-GOF` may naively seem susceptible to this same pitfall, their projection method appears to elegantly avoid it.

As a final test, we note that `zCDP-GOF` guarantees zCDP, while `Priv'IT` guarantees (vanilla) differential privacy. In our previous tests, our guarantee was ε-differential privacy, while theirs was $\frac{\varepsilon^2}{2}$-zCDP: by Proposition 13, our guarantees imply theirs. In the third test, we revisit uniformity testing, but when *their guarantees imply ours*. More specifically, again with $\varepsilon = 0.1$, we ran `zCDP-GOF` with the guarantee of $\frac{\varepsilon^2}{2}$-zCDP and `Priv'IT` with the guarantee of $(\frac{\varepsilon^2}{2} + \varepsilon\sqrt{2\log(1/\delta)}, \delta)$ for various $\delta > 0$. We note that δ is often thought in theory to be "cryptographically small" (such as 2^{-100}), but we compare with a wide range of δ, both large and small: $\delta = 1/e^t$ for $t \in \{1, 2, 4, 8, 16\}$. This test was conducted on support sizes n ranging from 10 to 6000.

The results of our third test are provided in Figure 5-3. We found that, for all δ tested,

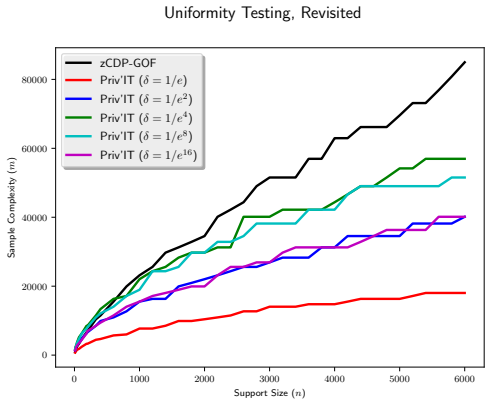


Figure 5-3: The sample complexities of `Priv'IT` and `zCDP-GOF` for uniformity testing, with approximate differential privacy

`Priv'IT` required fewer samples than `zCDP-GOF`. This is unsurprising for δ very large and small, since the differential privacy guarantees become very easy to satisfy, but we found it to be true for even "moderate" values of δ. This implies that if an analyst is satisfied with approximate differential privacy, she might be better off using `Priv'IT`, rather than an algorithm which guarantees zCDP.

While the main focus of our evaluation was statistical in nature, we will note that `Priv'IT` was more efficient in runtime than our implementation of `MCGOF`, and more efficient in memory usage than our implementation of `zCDP-GOF`. The former point was observed by noting that, in the same amount of time, `Priv'IT` was able to reach a trial corresponding to a support size of 20000, while `MCGOF` was only able to reach 10000. The latter point was observed by noting that `zCDP-GOF` ran out of memory at a support size of 11800. This is likely because `zCDP-GOF` requires matrix computations on a matrix of size $O(n^2)$. It is plausible that all of these implementations could be made more time and memory efficient, but we found our implementations to be sufficient for the sake of our comparison.

5.5.2 Entropy

We compare the performance of our entropy estimator with a number of alternatives, both private and non-private. Non-private algorithms considered include the plug-in estimator (`plug-in`), the Miller-Madow Estimator (`MM`) [Mil55], the sample optimal polynomial approximation estimator (`poly`) of [WY16]. We analyze the privatized versions of plug-in, and

poly in Sections 5.4.4.1 and 5.4.4.2, respectively. The implementation of the latter is based on code from the authors of [WY16][4]. We compare performance on different distributions including uniform, a distribution with two steps, Zipf(1/2), a distribution with Dirichlet-1 prior, and a distribution with Dirichlet-1/2 prior, and over varying support sizes.

While plug-in, and MM are parameter free, poly (and its private counterpart) have to choose the degree L of the polynomial to use, which manifests in the parameter λ in the statement of Theorem 47. [WY16] suggests the value of $L = 1.6 \log n$ in their experiments. However, since we add further noise, we choose a single L as follows: (i) Run privatized poly for different L values and distributions for $n = 2000$, $\varepsilon = 1$, (b) Choose the value of L that performs well across different distributions (See Figure 5-4). We choose $L = 1.2 \cdot \log n$ from this, and use it for all other experiments. To evaluate the sensitivity of poly, we computed the estimator's value at all possible input values, computed the sensitivity, (namely, $\Delta = \max_{d_{hamming}(X_1^m, Y_1^m) \leq 1} |\texttt{poly}(X_1^m) - \texttt{poly}(Y_1^m)|$), and added noise distributed as Lap $\left(0, \frac{\Delta}{\varepsilon}\right)$.

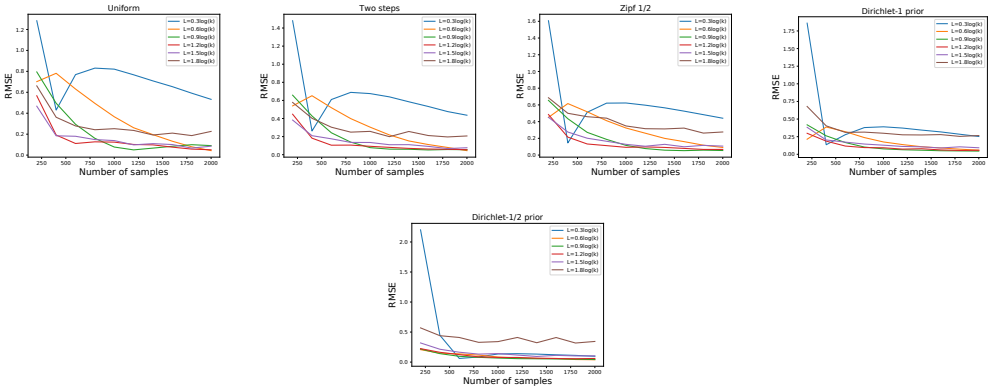


Figure 5-4: RMSE comparison between private Polynomial Approximation Estimators for entropy with various values for degree L, $n = 2000$, $\varepsilon = 1$. The degree L represents a bias-variance tradeoff: a larger degree decreases the bias but increases the sensitivity, necessitating the addition of Laplace noise with a larger variance.

The RMSE of various estimators for $n = 1000$, and $\varepsilon = 1$ for various distributions are illustrated in Figure 5-5. The RMSE is averaged over 100 iterations in the plots.

We observe that the performance of our private-poly is near-indistinguishable from the non-private poly, particularly as the number of samples increases. It also performs significantly better than all other alternatives, including the non-private Miller-Madow and the

[4]See https://github.com/Albuso0/entropy for their code for entropy estimation.

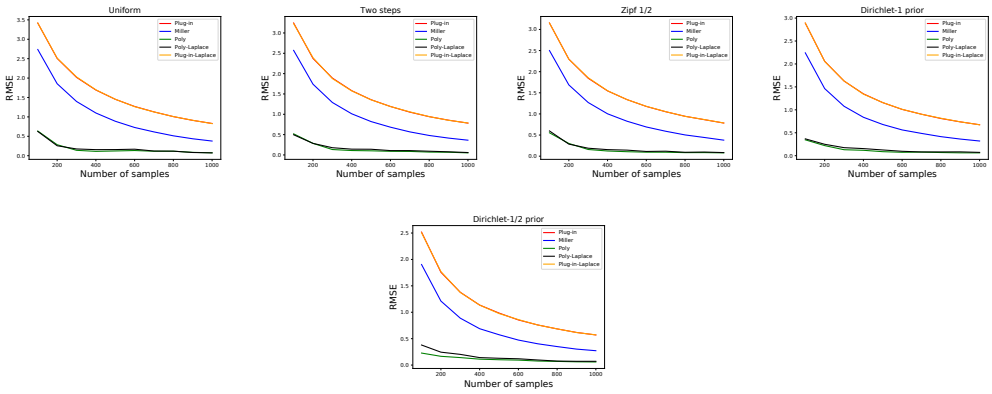


Figure 5-5: Comparison of various estimators for entropy, $n = 1000$, $\varepsilon = 1$.

plug-in estimator. The cost of privacy is minimal for several other settings of n and ε, for which results appear in Section 5.5.4.

5.5.3 Support Coverage

We investigate the cost of privacy for the problem of support coverage. We provide a comparison between the Smoothed Good-Toulmin estimator (SGT) of [OSW16] and our algorithm, which is a privatized version of their statistic (see Section 5.4.1.1). Our implementation is based on code provided by the authors of [OSW16]. As shown in our theoretical results, the sensitivity of SGT is at most $2(1 + e^r(t-1))$, necessitating the addition of Laplace noise with parameter $2(1 + e^{r(t-1)})/\varepsilon$. Note that while the theory suggests we select the parameter $r = \log(1/\alpha)$, α is unknown. We instead set $r = \frac{1}{2t}\log_e \frac{m(t+1)^2}{t-1}$, as previously done in [OSW16].

5.5.3.1 Evaluation on Synthetic Data

In our synthetic experiments, we consider different distributions over different support sizes n. We generate $m = n/2$ samples, and then estimate the support coverage at $k = m \cdot t$. For large t, estimation is harder. Some results of our evaluation on synthetic are displayed in Figure 5-6. We compare the performance of SGT, and privatized versions of SGT with parameters $\varepsilon = 1, 2$, and 10. For this instance, we fixed the domain size $n = 20000$. We ran the methods described above with $m = n/2$ samples, and estimated the support coverage at

$k = mt$, for t ranging from 1 to 10. The performance of the estimators is measured in terms of RMSE over 1000 iterations.

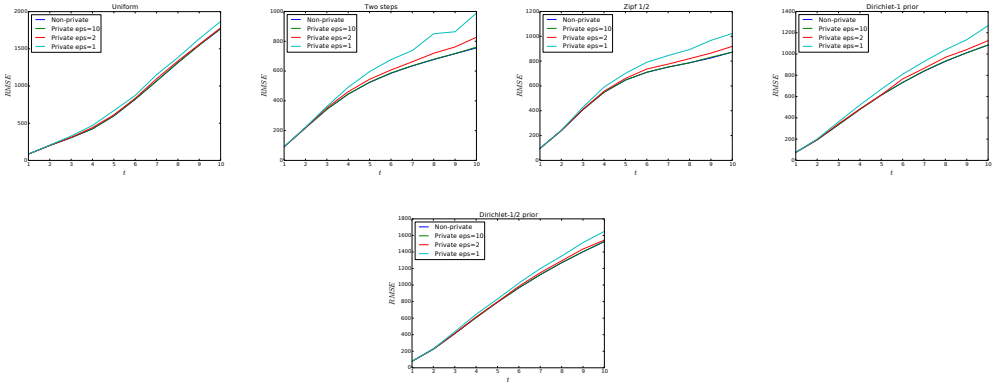


Figure 5-6: Comparison between our private support coverage estimator with non-private SGT when $n = 20000$

We observe that, in this setting, the cost of privacy is relatively small for reasonable values of ε. This is as predicted by our theoretical results, where unless ε is extremely small (less than $1/n$) the non-private sample complexity dominates the privacy requirement. However, we found that for smaller support sizes (as shown in Section 5.5.4.2), the cost of privacy can be significant. We provide an intuitive explanation for why no private estimator can perform well on such instances. To minimize the number of parameters, we instead argue about the related problem of support-size estimation. Suppose we are trying to distinguish between distributions which are uniform over supports of size 100 and 200. We note that, if we draw $m = 50$ samples, the "profile" of the samples (i.e., the histogram of the histogram) will be very similar for the two distributions. In particular, if one modifies only a few samples (say, five or six), one could convert one profile into the other. In other words, these two profiles are almost-neighboring datasets, but simultaneously correspond to very different support sizes. This pits the two goals of privacy and accuracy at odds with each other, thus resulting in a degradation in accuracy.

5.5.3.2 Evaluation on Census Data and Hamlet

We conclude with experiments for support coverage on two real-world datasets, the 2000 US Census data and the text of Shakespeare's play Hamlet, inspired by investigations in [OSW16]

and [VV17b]. Our investigation on US Census data is also inspired by the fact that this is a setting where privacy is of practical importance, evidenced by the proposed adoption of differential privacy in the 2020 US Census [DLS+17].

The Census dataset contains a list of last names that appear at least 100 times. Since the dataset is so oversampled, even a small fraction of the data is likely to contain almost all the names. As such, we make the task non-trivial by subsampling $k_{total} = 86080$ individuals from the data, obtaining 20412 distinct last names. We then sample m of the k_{total} individuals without replacement and attempt to estimate the total number of last names. Figure 5-7 displays the RMSE over 100 iterations of this process. We observe that even with an exceptionally stringent privacy budget of $\varepsilon = 0.5$, the performance is almost indistinguishable from the non-private SGT estimator.

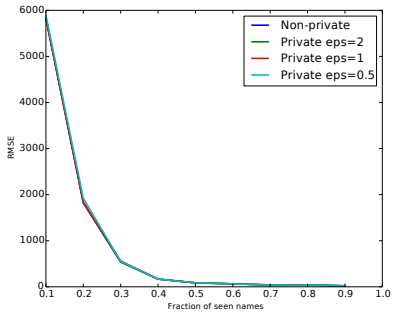


Figure 5-7: Comparison between our private support coverage estimator with the SGT on Census Data.

The Hamlet dataset has $k_{total} = 31,999$ words, of which 4804 are distinct. Since the distribution is not as oversampled as the Census data, we do not need to subsample the data. Besides this difference, the experimental setup is identical to that of the Census dataset. Once again, as we can see in Figure 5-8, we get near-indistinguishable performance between the non-private and private estimators, even for very small values of ε. Our experimental results demonstrate that privacy is realizable in practice, with particularly accurate performance on real-world datasets.

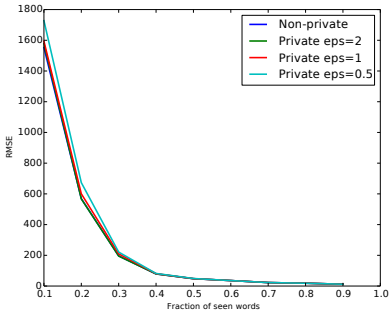


Figure 5-8: Comparison between our private support coverage estimator with the SGT on Hamlet.

5.5.4 Additional Experimental Results

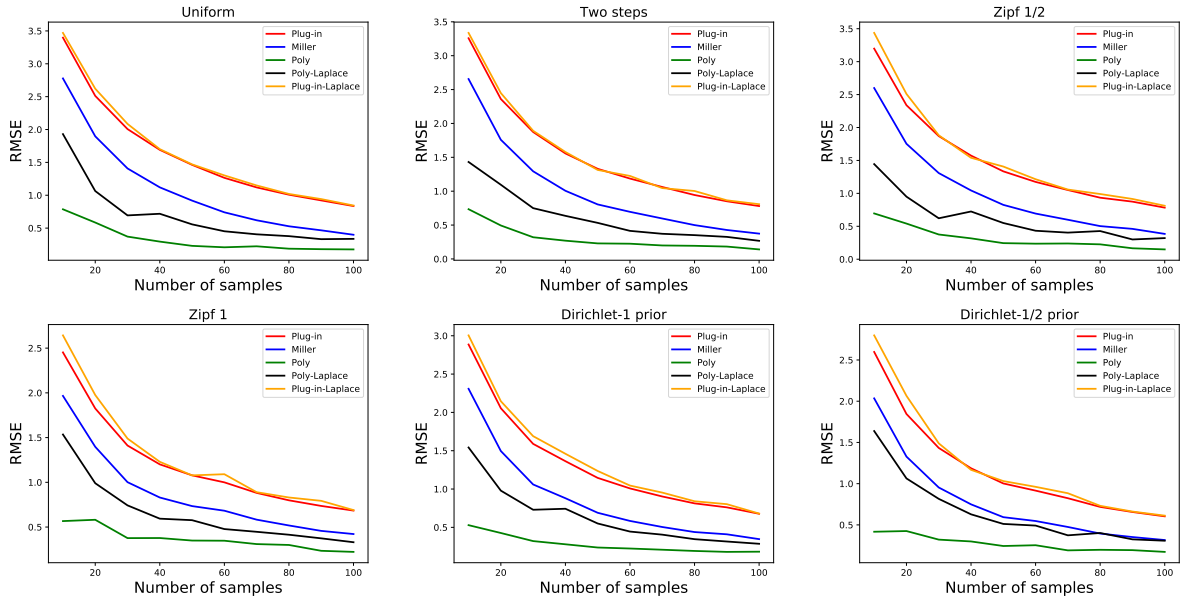
This section contains additional plots of our synthetic experimental results. Section 5.5.4.1 contains experiments on entropy estimation, while Section 5.5.4.2 contains experiments on estimation of support coverage.

5.5.4.1 Entropy Estimation

We present four more plots of our synthetic experimental results for entropy estimation. Figures 5-9 and 5-10 are on a smaller support of $n = 100$, with $\varepsilon = 1$ and 2, respectively. Figures 5-11 and 5-12 are on a support of $n = 1000$, with $\varepsilon = 0.5$ and 2.

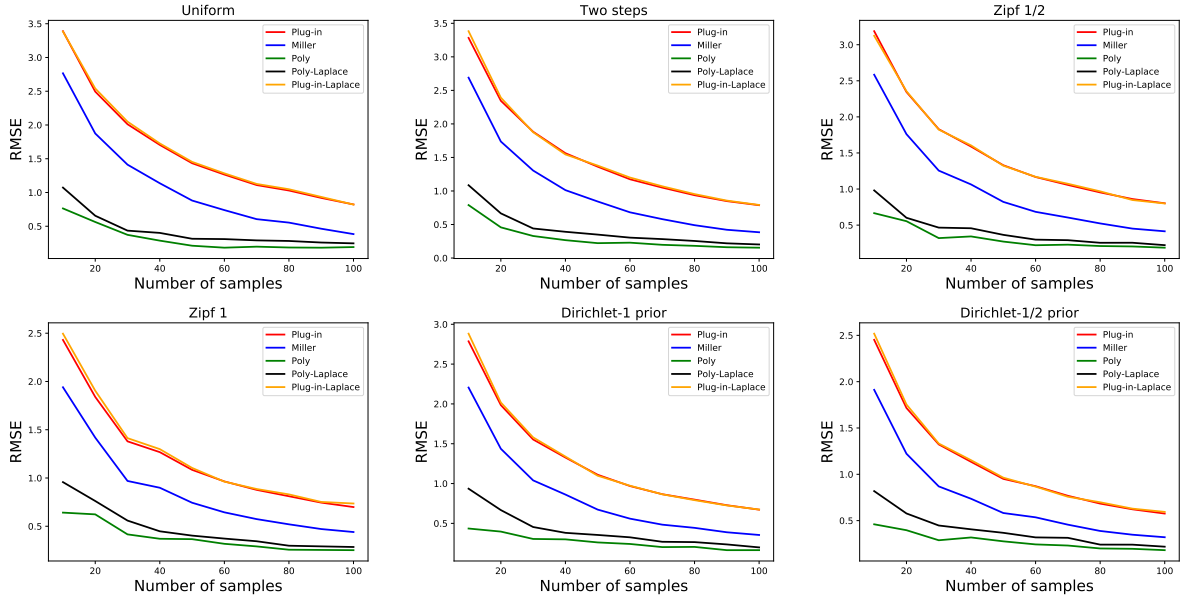
5.5.4.2 Support Coverage

We present three additional plots of our synthetic experimental results for support coverage estimation. In particular, Figures 5-13, 5-14, and 5-15 show support coverage for $n = 1000$, 5000, 100000.



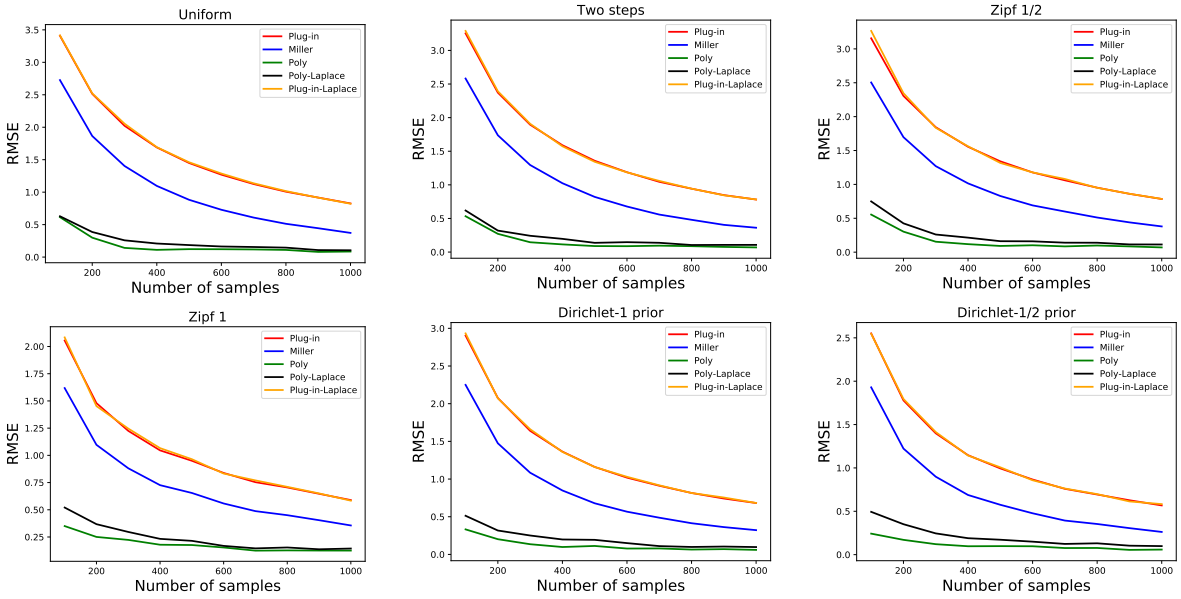
(a) (b) (c)

Figure 5-9: Comparison of various estimators for the entropy, $n = 100$, $\varepsilon = 1$.



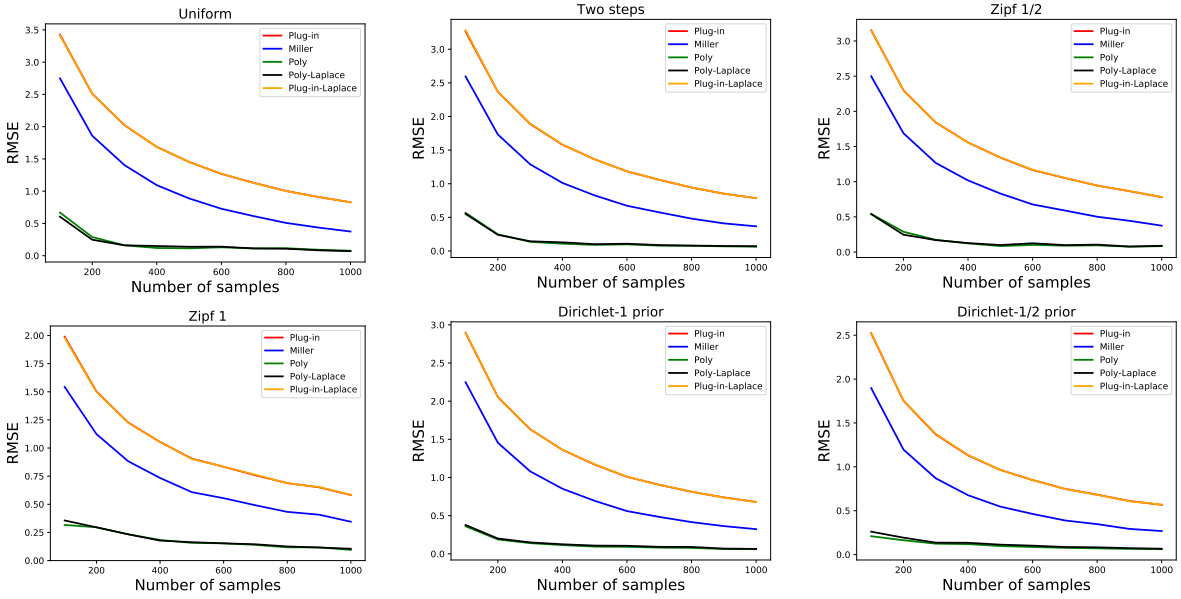
(a) (b) (c)

Figure 5-10: Comparison of various estimators for the entropy, $n = 100$, $\varepsilon = 2$.



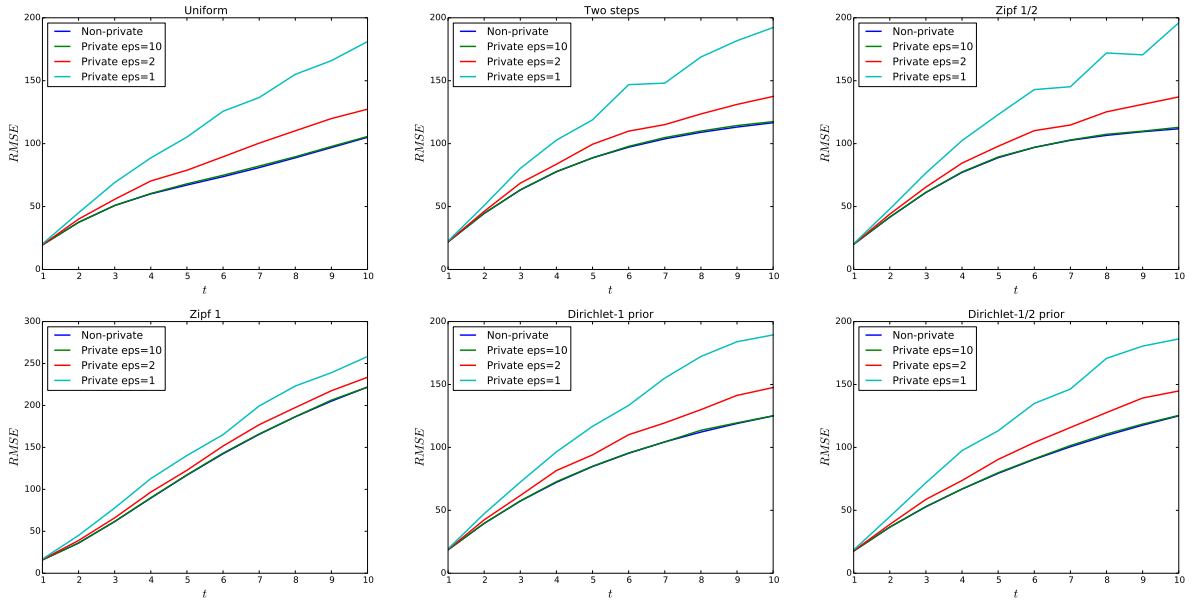
(a) (b) (c)

Figure 5-11: Comparison of various estimators for the entropy, $n = 1000$, $\varepsilon = 0.5$.



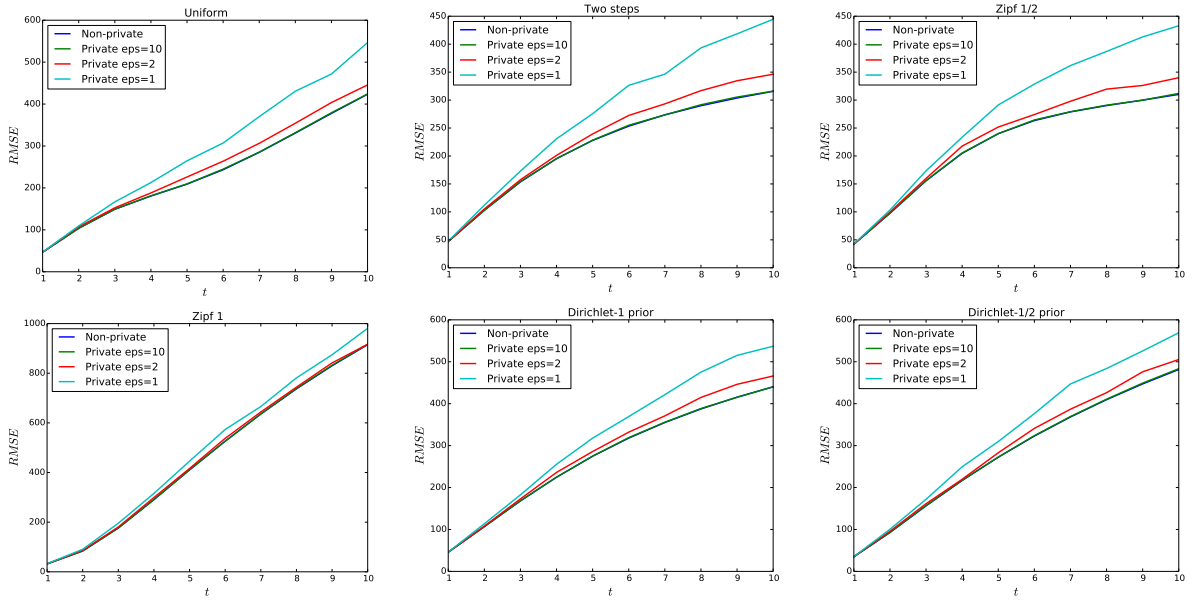
(a) (b) (c)

Figure 5-12: Comparison of various estimators for the entropy, $n = 1000$, $\varepsilon = 2$.



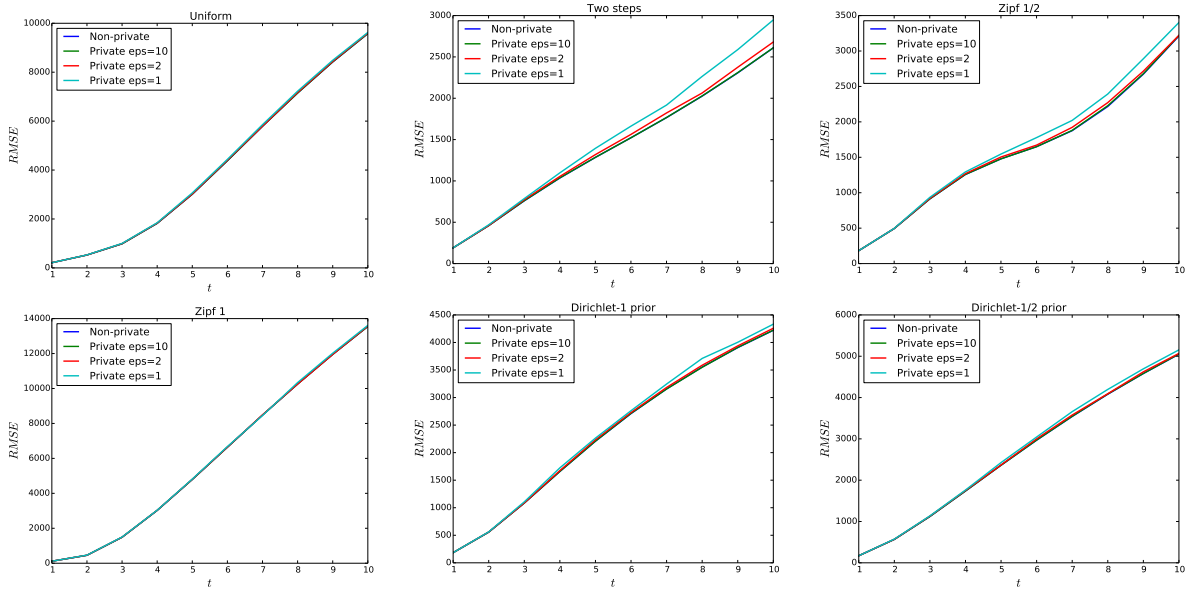
(a) (b) (c)

Figure 5-13: Comparison between the private estimator with the non-private SGT when $n = 1000$.



(a) (b) (c)

Figure 5-14: Comparison between the private estimator with the non-private SGT when $n = 5000$.



(a) (b) (c)

Figure 5-15: Comparison between the private estimator with the non-private SGT when $n = 100000$.

Chapter 6

Testing with Conditional Samples

6.1 Introduction

Up until this point, our objective has been to obtain algorithms which are *sublinear* in n, the size of the domain. However, as previously discussed in Chapter 4, in modern data analysis we may encounter settings in which the domain is exceptionally large, necessitating a complexity which is logarithmic in (or even independent of) the domain size. This goal seems at odds with polynomial lower bounds on the sample complexity of most natural testing questions (cf. Theorems 2 and 3). Some ways of avoiding these lower bounds involve instance-by-instance analysis (which is outside the scope of this thesis, see e.g. [ADJ$^+$11, ADJ$^+$12, AJOS13, VV17a, VV15, OS15, JHW16, BCG17, BW17a] for examples of this style of analysis) or assuming some sort of structure on the underlying density (a la Chapter 4). In this chapter, we pursue a different direction: we give ourself additional power when interacting with the distribution.

In recent years, the most popular method of augmenting the power of distribution testers in the theory community has been contained by the *conditional sampling* model. This model was recently introduced concurrently by Chakraborty, Fischer, Goldhirsh, and Matsliah [CFGM13, CFGM16] and Canonne, Ron, and Servedio [CRS14, CRS15]. The algorithm is able to *query* a distribution in the following way: it submits a query set S to an oracle, which returns a sample from the distribution conditioned on being from S. Additionally, we will distinguish between conditional sampling models where the algorithm's queries may be

adaptive (COND) or non-adaptive (NACOND) [1]. In comparison, we will use SAMP to refer to the standard sampling model.

The conditional model was introduced in part to capture the more dynamic nature of modern data collection. Classically, a statistician might refer to a pre-existing dataset, and then perform some statistical analysis upon it. Nowadays, the dataset and the analysis are often gathered and performed at the same time and by the same people, perhaps even with modifications to the data acquisition procedure based on the results of preliminary tests. In such an *interactive* statistical setting, design of efficient algorithms corresponds to a principled method for non-wasteful data collection in the design of experiments. This is also explored in the literature on *active learning*, discussed in Section 6.1.2.

One may wonder in which specific settings one will have access to a conditional sampling oracle. Some motivating examples are provided in [CFGM13], including testing lottery machines and asymmetric communication schemes. We highlight their example of political polling: when one attempts to generate poll numbers, this generally does not involve questioning wholly random individuals from the population. Instead, a pollster will *condition* their sample upon various demographics, including age, sex, and education. Another motivation for studying non-traditional oracles is to demonstrate their advantages to the database community. Indeed, if one can show that alternative models of data access can yield significantly faster algorithms, database researchers can work towards optimizing the cost of these non-traditional queries in their system [KXS+16].

Conditional sampling often dramatically reduces the complexity of distribution testing problems. For example, as previously discussed, given SAMP access to a distribution, the sample complexity of testing uniformity is $\Theta(\sqrt{n}/\varepsilon^2)$ [Pan08, VV17a, ADK15, DKN15b, DGPP16, DGPP18]. However, given COND access, the query complexity drops to $\tilde{\Theta}(1/\varepsilon^2)$ [FJO+15], completely removing the dependence on the support size. Similarly, significant qualitative savings can be realized for almost all natural distribution properties. To highlight an even more extreme example, consider the "estimation" version of the above problem: estimating the distance between a distribution and the uniform distribution. In SAMP, the complexity of this problem is known to be $\Theta\left(\frac{n}{\log n}\right)$ [VV10a, VV10b, VV11a,

[1] A formal definition of these concepts is given in Definition 18

VV11b, JHW16, HJW16, JVHW17]. But once again, given COND access, the complexity drops significantly, to $\tilde{O}\left(\frac{1}{\varepsilon^{20}}\right)$. In other words, one of the hardest property estimation tasks (with complexity which is "barely" sublinear) becomes one of the easiest (as the complexity is independent of n).

In this chapter, we present a number of results on distributional property testing and estimation and discuss their interplay with each other and the existing literature. More precisely, we will describe upper and lower bounds for uniformity, identity, equivalence testing and support-size estimation, in both the COND and NACOND models. Along the way, we will point out interesting qualitative relationships between the complexities of various problems, in an effort to identify precisely from where the power of the conditional sampling model is derived.

6.1.1 Results, Techniques, and Discussion

Our results are summarized pictorially in Table 6.1.

	SAMP	NACOND	COND
Uniformity	$\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [Pan08, VV17a]	$\tilde{O}\left(\frac{\log n}{\varepsilon^2}\right)$ [this work] $\Omega\left(\frac{\log n}{\varepsilon}\right)$ [this work]	$\tilde{\Theta}\left(\frac{1}{\varepsilon^2}\right)$ [CRS15]
Identity	$\Theta\left(\frac{\sqrt{n}}{\varepsilon^2}\right)$ [Pan08, VV17a]	$\tilde{O}\left(\frac{\log^2 n}{\varepsilon^2}\right)$ [this work] $\Omega\left(\frac{\log n}{\varepsilon}\right)$ [this work]	$\tilde{O}\left(\frac{1}{\varepsilon^2}\right)$ [FJO+15] $\Omega\left(\frac{1}{\varepsilon^2}\right)$[CRS15]
Equivalence	$\Theta\left(\max\left(\frac{n^{2/3}}{\varepsilon^{4/3}}, \frac{n^{1/2}}{\varepsilon^2}\right)\right)$ [CDVV14]	$\tilde{O}\left(\frac{\log^{12} n}{\varepsilon^2}\right)$ [this work] $\Omega\left(\frac{\log n}{\varepsilon}\right)$ [this work]	$\tilde{O}\left(\frac{\log\log n}{\varepsilon^5}\right)$ [FJO+15] $\Omega\left(\sqrt{\log\log n}\right)$ [this work]
Support-Size Estimation	$\Theta\left(\frac{n}{\log n}\right)$ [VV17b, OSW16]	$O\left(\text{poly}\left(\frac{\log n}{\varepsilon}\right)\right)$ [this work] $\Omega\left(\frac{\log n}{\varepsilon}\right)$ [this work]	$\tilde{O}\left(\frac{\log\log n}{\varepsilon^3}\right)$ [this work] $\Omega\left(\sqrt{\log\log n}\right)$ [CFGM13]

Table 6.1: Summary of results, and a comparison of various testing problems in different sampling oracle models. For the first three rows, problems get harder as one moves down and to the left in this table. The row on support-size estimation is incomparable with the other rows.

6.1.1.1 A Lower Bound for Adaptive Equivalence Testing

Our first result considers the sample complexity of testing equivalence with adaptive queries under the COND model. This resolves (in the negative) the question of whether constant-

query complexity was achievable, an open problem explicitly posed by Fischer [Fis14].

Theorem 50 (Adaptive Equivalence Testing Lower Bound)**.** *Any algorithm which, given* COND *access to unknown distributions p, q on $[n]$, distinguishes between the cases $p = q$ and $d_{\mathrm{TV}}(p, q) \geq 1/4$ with probability at least $2/3$ must make at least $\Omega\left(\sqrt{\log \log n}\right)$ queries.*

Combined with the $\tilde{O}(\log \log n)$ upper bound of Falahatgar et al. [FJO$^+$15], this almost (i.e., up to a quadratic factor) settles the sample complexity of this question. Furthermore, as the related task of identity testing *can* be performed with a constant number of queries in the COND model, this demonstrates an intriguing and intrinsic qualitative difference between the two problems. Our result can also be interpreted as showing a fundamental distinction from the usual sampling model, where both identity and equivalence testing have polynomial sample complexity.

In order to prove Theorem 50, we have to deal with one main conceptual issue: *adaptivity*. While the standard sampling model does not, by definition, allow any choice on what the next query to the oracle should be, this is no longer the case for COND algorithms. Quantifying the power that this grants an algorithm makes things much more difficult. To handle this point, we follow the approach of Chakraborty et al. [CFGM13] and focus on a restricted class of algorithms they introduce, called "core adaptive testers" (see Section 6.2 for a formal definition). They show that this class of testers is equivalent to general algorithms for the purpose of testing a broad class of properties, namely those which are invariant to any permutation of the domain. Using this characterization, it remains for us to show that none of these structurally much simpler core testers can distinguish whether they are given conditional access to (a) a pair of random identical distributions (p, p), or (b) two distributions (p, q) drawn according to a similar process, which are far apart.

At a high level, our lower bound works by designing instances where the property can be tested if and only if the support size is known to the algorithm. Our construction randomizes the support size by embedding the instance into a polynomially larger domain. Since the algorithm is only allowed a small number of queries, Yao's Minimax Principle allows us to argue that, with high probability, a deterministic algorithm is unable to "guess" the support size. This separates queries into several cases. First, in a sense we make precise, it is

somehow "predictable" whether or not a query will return an element previously observed. If this occurs, it is similarly predictable *which* element the query will return. On the other hand, if a fresh element is observed, the query set is either "too small" or "too large." In the former case, the query will entirely miss the support, and the sampling process is identical for both types of instance. In the latter case, the query will hit a large portion of the support, and the amount of information gleaned from a single sample is minimal.

At a lower level, this process itself is reminiscent of the "hard" instances underlying the lower bound of Canonne, Ron, and Servedio [CRS14] for testing identity (with a PAIRCOND oracle, which can only query on sets of size 2), with one pivotal twist. As in their work, both p and q are uniform within each of $\omega(1)$ "buckets" whose size grows exponentially and are grouped into "bucket-pairs." Then, q is obtained from p by internally redistributing the probability mass of each pair of buckets, so that the total mass of each pair is preserved but each particular bucket has mass going up or down by a constant factor (see Section 6.3.1 for details of the construction). However, we now add a final step, where in both p and q the resulting distribution's support is *scaled by a random factor*, effectively reducing it to a (randomly) negligible fraction of the domain. Intuitively, this last modification has the role of "blinding" the testing algorithm. We argue that unless its queries are on sets whose size somehow match (in a sense formalized in Section 6.3.2) this random size of the support, the sequences of samples it will obtain under p and q are almost identically distributed. The above discussion crucially hides many significant aspects and technical difficulties which we address in Section 6.3. Moreover, we observe that the lower bound we obtain seems to be optimal with regard to our proofs techniques (specifically, to the decision tree approach), and not an artifact of our lower bound instances. Namely, there appear to be conceptual barriers to strengthening our result, which would require new ideas.

6.1.1.2 An Upper Bound for Adaptive Support-Size Estimation

We provide the following theorem for adaptively estimating the support size of a distribution.

Theorem 51 (Adaptive Support-Size Estimation)**.** *Let $\tau > 0$ be any constant. There exists an adaptive algorithm which, given* COND *access to an unknown distribution p on $[n]$ (guaranteed to have probability mass at least τ/n on every element of its support) and accuracy*

parameter $\varepsilon \in (0,1)$*, makes* $\tilde{O}\left((\log\log n)/\varepsilon^3\right)$ *queries to the oracle*[2] *and outputs a value* $\tilde{\omega}$ *such that the following holds. With probability at least* $2/3$*,* $\tilde{\omega} \in [\frac{1}{1+\varepsilon} \cdot \omega, (1+\varepsilon) \cdot \omega]$*, where* $\omega = |\operatorname{supp}(p)|$.

Our algorithm is simple in spirit, and follows a guess-and-check strategy. In more detail, it first obtains a "reference point" *outside* the support, to check whether subsequent samples it may consider belong to the support. Then, it attempts to find a *rough upper bound* on the size of the support, of the form 2^{2^j} (so that only $\log\log n$ many options have to be considered); by using its reference point to check if a uniform random subset of this size contains, as it should, at least one point from the support. Once such an upper bound has been obtained using this double-exponential strategy, a refined bound is then obtained via a binary search on the new range of values for the exponent, $\{2^{j-1}, \ldots, 2^j\}$. Not surprisingly, our algorithm draws on similar ideas as in [RT16, Sto85], with some additional machinery to supplement the differences in the models. Interestingly, as a side-effect, this upper bound shows our analysis of Theorem 50 to be tight up to a quadratic improvement. Indeed, the lower bound construction we consider (see Section 6.3.1) can be easily "defeated" if an estimate of the support size is known, and therefore cannot yield better than a $\Omega(\log\log n)$ lower bound. Similarly, this further shows that the adaptive lower bound for support-size estimation of Chakraborty et al. [CFGM13] is also tight up to a quadratic improvement.

6.1.1.3 Anaconda: Non-Adaptive Upper Bounds

At this point, we have a developed understanding of the power of the COND oracle with respect to the aforementioned distribution testing problems. Perhaps surprisingly, the relative complexities of certain problems have qualitatively different relationships between SAMP and COND. To be precise, the sample complexities of identity testing and equivalence testing in SAMP are $\Theta(n^{1/2})$ ([Pan08, VV17a]) and $\Theta(n^{2/3})$ ([CDVV14]) respectively: there is a polynomial relationship between the two. However, their query complexities in COND are $\Theta(1)$ ([CRS15, FJO+15]) and $\log^{\Theta(1)}\log n$ ([FJO+15] and Theorem 50) respectively: there is a "chasm" between the two complexities, as we go from no dependence on the domain size to a doubly logarithmic one.

[2] We remark that the constant in the $\tilde{O}$ depends polynomially on $1/\tau$.

However, the picture is much less clear when it comes to the non-adaptive NACOND model. We know that the complexity of identity testing is $\operatorname{poly}\log n$ ([CFGM13] and Theorem 56 below), though the upper and lower bounds are quite far from each other. On the other hand, the complexity of equivalence testing is far less clear: the best lower bound is $\Omega(\log n)$ (Theorem 56 below), and the best upper bound is $O(n^{2/3})$ ([CDVV14]). Given the interesting qualitative behavior observed for the COND model, this begs the following question:

Question 7. *What is the relationship of the query complexities of identity and equivalence testing in the* NACOND *model?*

In particular, are they polynomially related, as in the SAMP model? Or is there a larger gap between the two, as in the COND model? Stated another way, do we require both conditional samples and adaptivity *simultaneously* in order to reap the benefits for testing equivalence?

We provide a qualitative resolution to this problem: we give a $\operatorname{poly}\log n$-query algorithm for equivalence testing.

Theorem 52 (Non-Adaptive Equivalence Testing)**.** *There exists an algorithm which, given* NACOND *access to unknown distributions p, q on $[n]$, makes $\tilde{O}\left(\frac{\log^{12} n}{\varepsilon^2}\right)$ queries to the oracle on each distribution and distinguishes between the cases $p = q$ and $d_{\mathrm{TV}}(p, q) \geq \varepsilon$ with probability at least $2/3$.*

For the special case of uniformity testing, we have a sharper analysis, allowing us to obtain a $\tilde{O}(\log n)$ query algorithm, which nearly matches the $\Omega(\log n)$ lower bound of Theorem 56:

Theorem 53 (Non-Adaptive Uniformity Testing)**.** *There exists an algorithm which, given* NACOND *access to an unknown distribution p on $[n]$, makes $\tilde{O}\left(\frac{\log n}{\varepsilon^2}\right)$ queries to the oracle on p and distinguishes between the cases $p = \mathcal{U}_n$ and $d_{\mathrm{TV}}(p, \mathcal{U}_n) \geq \varepsilon$ with probability at least $2/3$, where $\mathcal{U}_n$ is the uniform distribution on $[n]$.*

As a corollary of Theorem 53, we can obtain an improved upper bound for identity testing with an adaptation of the reduction from identity testing to uniformity testing of [CFGM16] (inspired by the bucketing techniques of [BFR+00, BFF+01]).

Theorem 54 (Non-Adaptive Identity Testing)**.** *There exists an algorithm which, given* NACOND *access to an unknown distribution p on $[n]$ and a description of a distribution q over $[n]$, makes $\tilde{O}\left(\frac{\log^2 n}{\varepsilon^2}\right)$ queries to the oracle on p and distinguishes between the cases $p = q$ and $d_{\mathrm{TV}}(p, q) \geq \varepsilon$ with probability at least $2/3$.*

We present a unified algorithm, Anaconda, for both equivalence and uniformity testing, the only difference is in the choice of parameters. Anaconda is quite simple to describe, requiring only four sentences below.[3] We consider this algorithmic simplicity to be an advantage of Anaconda, though we regret that its analysis is less simple.

Our bound for equivalence testing in the NACOND model is the first tailored to this setting. Specifically, the best upper bound was $O(n^{2/3})$ (for the harder problem of equivalence testing in the SAMP model [CDVV14]), and the best lower bound was $\Omega(\log n)$ (for the easier problem of uniformity testing in the NACOND model (Theorem 56)). These results left open the question of the true complexity of equivalence testing: is it polynomial in $\log n$, or polynomial in n? Our algorithm gives an exponential improvement in the query complexity by showing that the former is true: equivalence testing enjoys significant savings in the query complexity when we switch from the SAMP to the NACOND oracle model.

More generally, as mentioned before, our results expose a qualitatively interesting relationship between identity and equivalence testing in the NACOND model. In the standard sampling model (SAMP), the complexity of these problems is known to be polynomially related ($\Theta(n^{1/2})$ versus $\Theta(n^{2/3})$). However, in the conditional sampling model with adaptivity (COND), there is a "chasm" between these two complexities: one has a constant query complexity, while the other has a complexity which is doubly logarithmic in n ($\Theta(1)$ versus $\operatorname{poly}\log\log n$). Our results demonstrate that when we remove adaptivity from the conditional sampling model (NACOND), the relationship is qualitatively quite different. In this setting, the "chasm" closes, and the complexity of both problems is once again polynomially related: both are $\operatorname{poly}\log n$. Interestingly, this complexity is intermediate to the complexity of the same problems in the SAMP and COND models, by an exponential factor on either side. These relationships are all summarized in Table 6.1. We note that our results further address the aforementioned open problem of Fischer [Fis14], which inquires about the

[3] Perhaps if we tried harder, we could describe it in two sentences, plus the word "repeat."

complexity of equivalence testing with conditional samples.

In terms of specific sample complexities, we observe that our upper bound for uniformity testing is nearly tight: our $\tilde{O}\left(\frac{\log n}{\varepsilon^2}\right)$ upper bound is complemented by the $\Omega(\log n)$ lower bound of Theorem 56. It improves upon the algorithm of [CFGM13], which has query complexity $O\left(\frac{\log^{12.5} n}{\varepsilon^{17}}\right)$. Our algorithm for identity testing, with complexity $\tilde{O}\left(\frac{\log^2 n}{\varepsilon^2}\right)$, also significantly improves over theirs, which has a similar complexity as their algorithm for uniformity testing. We again mention that our bound for equivalence testing is exponentially better than the previous best algorithm for this problem (which is the $O(n^{2/3})$-query algorithm in the SAMP model of [CDVV14]).

Techniques and Proof Ideas At the core of our approach is reducing the problem from ℓ_1-testing to ℓ_∞-testing, the latter of which is much cheaper in terms of sample complexity. In particular, throughout this exposition, keep in mind that one can estimate a distribution up to ε in ℓ_∞-distance at a cost of $O(1/\varepsilon^2)$ samples (cf. Lemma 1). In order to give intuition on how such an approach could possibly work, we focus on two very simple instances of uniformity testing. In the first instance, p is a distribution with a single "spike": for some $i^* \in [n]$, $p(i^*) = \frac{1}{n} + \varepsilon$, and for $i \neq i^*$, $p(i) = \frac{1-\varepsilon}{n}$. This can be detected by simply choosing $S = [n]$ and querying it with NACOND $O(1/\varepsilon^2)$ times: the empirical distribution $\hat{p}(i^*)$ will have a similar spike, betraying that the distribution is non-uniform. In the second instance, p is the "Paninski construction" (used as the lower bound in [Pan08]): a random half of the domain elements have probability $\frac{1+\varepsilon}{n}$, while the other half have probability $\frac{1-\varepsilon}{n}$. This can be detected by choosing S to be two random symbols, and again querying this subset $O(1/\varepsilon^2)$ times. With constant probability, the two symbols will be from different sets. While the ℓ_∞ distance from uniformity on each symbol is only $\frac{\varepsilon}{n}$, in this *conditional* distribution, it is increased to ε, allowing easy detection.

These two examples illustrate the heart of our approach: our algorithm, ANACONDA, attempts to find a query set in which the discrepancy of a single item is large in comparison to the total probability mass of the set. One of our key lemmas (Lemma 61) shows that this is possible with probability $\geq \Omega\left(\frac{1}{\log n}\right)$. The flavor is somewhat reminiscent of Levin's Economical Work Investment Strategy [Gol14]. While the two instances above are

straightforward, a more careful analysis is required to avoid paying excess factors of $\log n$, particularly for uniformity and identity testing. That said, all the complexity is pushed to the analysis, and the algorithm itself is very simple to describe:

> First, the algorithm chooses a random power of two between 2 and n – roughly, this serves as a "guess" for (the inverse of) the size of the set which represents the discrepany between the distributions. Next, the algorithm chooses a random set $S \subseteq [n]$ of this size. Finally, it performs NACOND queries to S (on both distributions, for equivalence testing), in order to form an empirical distribution (which is accurate in ℓ_∞-distance) and check whether there is a discrepant symbol or not. This process is repeated several times, and if we fail to ever detect a discrepant symbol, we can conclude that the distributions are equal.

Since uniformity testing is relatively well-behaved, the key lemma mentioned above (Lemma 61) does most of the work. This is because in this setting, once we have a handle on the distribution of the discrepancy, it is easy to reason about how much of the mass from the uniform distribution is contained in a query set. We require a few additional concentration arguments on the total discrepancy and probability mass contained in the query set, as well as a separate analysis for the case where $|S|$ needs to be small and this concentration does not hold.

We then leverage our algorithm for uniformity testing to provide an algorithm for identity testing. This uses the reduction of [CFGM13][4], which partitions the domain so that the conditional distribution on each part is close to uniform, and tests for identity on each part. This requires a non-adaptive identity tester for distributions which are close to uniform (in ℓ_∞-distance) – we show our analysis for uniformity testing can be adapted to handle this case. Our application crucially modifies their reduction in order to minimize the sample complexity, as Anaconda can test against distributions which are further from uniform than theirs ($O(1/n)$, rather than $O(\varepsilon/n)$).

Finally, we turn to the most technically difficult problem of equivalence testing. This case turns out to be more challenging, as we must simultaneously reason about $p(i), p(S \setminus i), q(i)$,

[4]We note that the reduction of [Gol16], from identity testing to uniformity testing, is not known to apply in either the NACOND or COND models.

and $q(S \setminus i)$ – as mentioned prior, it is much easier to control the latter two quantities for uniformity testing. To establish our result, we must argue that ANACONDA identifies a set S where both differences $p(i) - q(i)$ and $p(S \setminus i) - q(S \setminus i)$ have opposite signs and are simultaneously relatively large compared to the magnitudes of $p(i), p(S \setminus i), q(i)$, and $q(S \setminus i)$ (Proposition 23). We consider the distribution of the discrepancy $p - q$, with a case analysis depending on the relationship between the "typical" magnitudes of the positive and negative discrepancies. If these magnitudes are close, then we can select a "smaller" set S (where "smaller" is defined based on these magnitudes) which has a reasonable probability of including a positively and negatively discrepant element of these magnitudes (Lemma 64). On the other hand, if these magnitudes are far, then with an appropriate choice of the size of the set S, there is a significant chance that our set will contain an element i with significant positive discrepancy $p(i) - q(i)$, while the total discrepancy in the set $p(S \setminus i) - q(S \setminus i)$ is very negative (Case 2 in Lemma 65). Despite all these technicalities, we emphasize that the algorithm itself is still quite simple; in particular, it is identical to the algorithm for uniformity testing (modulo some parameter modifications).

Besides the above testing results, we also sketch how to adapt our algorithm for adaptive support-size estimation (Theorem 51) to the non-adaptive setting. This exploits our $\tilde{O}(\log n)$-query algorithm for non-adaptive uniformity testing (Theorem 53) to obtain an $\tilde{O}(\log^2 n)$-query algorithm.

6.1.1.4 Non-Adaptive Lower Bounds

We conclude by complementing our NACOND upper bounds with NACOND lower bounds. Specifically, we establish a logarithmic lower bound on *non-adaptive* support-size estimation, for any (large enough) constant factor. This improves on the result of Chakraborty et al. [CFGM13], which gave a doubly logarithmic lower bound for constant factor support-size estimation.

Theorem 55 (Non-Adaptive Support-Size Estimation Lower Bound)**.** *Any algorithm which, given* NACOND *access to an unknown distribution p on $[n]$, estimates the size of the support up to a factor of $\gamma \geq \sqrt{2}$ must make at least $\Omega\left(\frac{\log n}{\log^2 \gamma}\right)$ queries.*

Moreover, the approach used to prove this theorem also implies an analogous lower bound on *non-adaptive* uniformity testing in the conditional model, answering a conjecture of Chakraborty et al. [CFGM13]:

Theorem 56 (Non-Adaptive Uniformity Testing Lower Bound)**.** *Any algorithm which, given* NACOND *access to an unknown distribution p on $[n]$, distinguishes between the cases $p = \mathcal{U}_n$ and $d_{\mathrm{TV}}(p, \mathcal{U}_n) \geq \varepsilon$ with probability at least $2/3$ must make at least $\Omega(\log n/\varepsilon)$ queries.*

These results complement $\operatorname{poly}\log(n)$-query upper bounds for uniformity, identity, and equivalence testing, and support-size estimation, as discussed in Section 6.1.1.3. This shows that all of these problems have query complexity $\log^{\Theta(1)} n$ in the NACOND model.

We proceed to outline our approach for proving Theorem 55. We define two families of distributions $\mathcal{P}$ and $\mathcal{Q}$, where an instance is either a draw (p, q) from $\mathcal{P} \times \mathcal{Q}$, or simply (p, p). Any distribution in $\mathcal{Q}$ has support size γ times that of its corresponding distribution in $\mathcal{P}$. Yet, we argue that no non-adaptive *deterministic* tester making too few queries can distinguish between these two cases, as the tuple of samples it will obtain from p or (the corresponding) q is almost identically distributed (where the randomness is over the choice of the instance itself). To show this last point, we analyze separately the case of "small" queries (conditioning on sets which turn out to be much smaller than the actual support size, and thus with high probability will not even intersect it) and the "large" ones (where the query set S is so big compared to the support T that a uniform sample from $S \cap T$ is essentially indistinguishable from a uniform sample from S). We conclude the proof by invoking Yao's Principle, carrying the lower bound back to the setting of non-adaptive *randomized* testers.

Interestingly, this argument essentially gives us Theorem 56 "for free." Indeed, the big-query-set case above is handled by proving that the distribution of samples returned on those queries is indistinguishable, both for $\mathcal{P}$ and $\mathcal{Q}$, from samples obtained from the *actual* uniform distribution. Considering again the small-query-set case separately, this allows us to argue that a random distribution from (say) $\mathcal{P}$ is indistinguishable from uniform.

6.1.1.5 Relation to the Ron-Tsur model

Recent work of Ron and Tsur [RT16] studies a model which is slightly different – and more favorable to the algorithm – than ours. In their setting, the algorithm still performs queries consisting of a subset of the domain, as in our case. However, the algorithm is also given the promise that the distribution is uniform on a subset of the domain, and whenever a query set contains 0 probability mass the oracle explicitly indicates this is the case. Their paper provides a number of results for support-size estimation in this model.

We point out two connections between our work and theirs. First, our $\Omega(\log n)$ lower bound for non-adaptive support-size estimation (Theorem 55) holds in the model of Ron and Tsur. Although lower bounds in the conditional sampling setting do not apply directly to their model, our construction and analysis do carry over, and provide a nearly tight answer to a question left unanswered in their paper. Also, our $\tilde{O}(\log \log n)$-query algorithm for adaptive support-size estimation (Theorem 51) can be seen as generalizing their result to the weaker conditional sampling model (most significantly, when we are not given the promise that the distribution be uniform).

6.1.2 Related Work

As mentioned before, the conditional sampling model was introduced in [CFGM13, CRS14] (the journal versions of these papers appear as [CRS15, CFGM16]). These initial works studied a number of distribution testing and property estimation problems in adaptive and non-adaptive settings, as well as under various types of conditional sampling oracles, including those which can only perform conditional samples on sets which are *simple*, for example, sets of size 2 or intervals. The general picture established by these works demonstrates that one can enjoy significantly reduced query complexity when one has conditional sampling access to a distribution.

Since the introduction of conditional sampling, a number of works have refined the complexity landscape of distribution testing problems in this model. [Can15a] provides bounds for testing monotonicity in the general conditional sampling model and when query sets must be intervals (as well as a number of other non-conditional distribution sampling mod-

els). [FJO+15] provides better algorithms for testing identity and equivalence, reducing the former complexity to $\tilde{O}(1/\varepsilon^2)$ (nearly matching the information theoretic limit), and the latter from $\operatorname{poly}\log n$ to $\operatorname{poly}\log\log n$. This equivalence testing upper bound is complemented by the $\Omega(\sqrt{\log\log n})$ lower bound of [ACK15b], showing that conditional samples grant a doubly-exponential improvement in the sample complexity for this problem, as well as demonstrating a "chasm" between the complexity of equivalence and identity testing with conditional samples. [ACK15b] also contains upper bounds for support-size estimation, as well as lower bounds for non-adaptive support-size estimation and uniformity testing. [KT18] delves deeper on non-adaptive distribution testing upper bounds, significantly reducing the complexity of testing uniformity, identity, and equivalence. An alternative method for proving COND lower bounds is presented in [BCG17], which involves reductions from testing to communication complexity protocols. [FLV17] studies the testing of composite hypotheses (a la Chapter 3) with adaptive and non-adaptive conditional samples. The quantum conditional sampling oracle is introduced in [SSJ17], demonstrating additional power over classical oracles for testing problems (both distributional and functional). Finally, [BC18] studies distribution testing on multivariate domains when query sets must be subcubes of the domain. The results contained in this thesis are from two of these works [ACK15b, KT18].

The conditional sampling model has also attracted attention outside the field of distribution testing. For instance, the weighted group testing model in [ACK15a] is inspired by the conditional sampling model. [GTZ17] studies the impact of a conditional oracle for more classical problems, such as k-means clustering and estimating the weight of a minimum spanning tree. [RT16] investigates the problem of estimating the size of a hidden set in a conditional-sampling-esque model. Finally, [GTZ18] gives verification-based algorithms for crowdsourcing tasks, in a method that is very reminiscent of conditional sampling.

There have been numerous other proposed oracle models which enable savings for distribution testing problems. Some examples include when the algorithm may query the PDF or CDF of the distribution [BDKR05, GMV06, RS09, CR14], or is given probability-revealing sample [OS18].

The conditional sampling model falls into a broader body of work on *interactive* learning, in which the algorithm has additional power when eliciting data. Perhaps the best known

line of work within this field is the *active learning* model for supervised learning. In this model, the algorithm is provided with unlabeled examples only (which are somehow "cheap" to obtain), and it may adaptively request labels for these points (which is considered to be a much more expensive operation). See [Set12, Han14] for surveys of this area, and [BBBY12] for a work on functional property testing in the active model.

6.1.3 Organization

In Section 6.2, we discuss preliminaries and the notation that we use throughout the chapter. In Section 6.3, we present a lower bound for testing equivalence in the COND model. In Section 6.4, we give an upper bound for estimating the support size of a distribution in the COND model. In Section 6.5, we describe ANACONDA, which results in a number of upper bounds in the NACOND model. Finally, in Section 6.6, we prove lower bounds for uniformity testing and support-size estimation in the NACOND model.

6.2 Preliminaries

For a set S, let $p(S) = \sum_{i \in S} p(i)$. Furthermore, let p_S be the conditional distribution of p restricted to S, i.e., $p_S(i) = p(i)/p(S)$.

We use the following definition of the conditional sampling model. Note that this uses the convention of [CFGM13] of sampling uniformly from query sets with 0 measure, rather than the convention of [CRS14] which immediately fails if given such a set, as the latter convention trivializes NACOND, reducing it to SAMP.

Definition 18. *A* conditional sampling *oracle for a distribution p is defined as follows: the oracle takes as input a* query set $S \subseteq [n]$*, and returns a symbol $i \in S$, where the probability that i is returned is equal to $p_S(i) = p(i)/p(S)$. If $p(S) = 0$, then a symbol $i \in S$ is returned uniformly at random.*

Given an adaptive conditional sampling *oracle (a* COND *oracle), the algorithm may query* adaptively*: before submitting each query set i, the algorithm is allowed to view the results of queries 1 through $i - 1$. In contrast, given a* non-adaptive conditional sampling *oracle*

(a NACOND *oracle), the algorithm must be non-adaptive: it must submit all query sets in advance of viewing any of their results.*

Adaptive Core Testers In order to deal with adaptivity in our lower bounds, we will use ideas introduced by Chakraborty et al. [CFGM13]. These ideas, for the case of *label-invariant* properties[5] allow one to narrow down the range of possible testers and focus on a restricted class of such algorithms called *adaptive core testers*. These core testers do not have access to the full information of the samples they draw, but instead only get to see the relations (inclusions, equalities) between the queries they make and the samples they get. Yet, Chakraborty et al. [CFGM13] show that any tester for a label-invariant property can be converted into a core tester with same query complexity; thus, it is enough to prove lower bounds against this – seemingly – weaker class of algorithms.

We here rephrase the definitions of a core tester and the view they have of the interaction with the oracle (the *configuration* of the samples), tailored to our setting.

Definition 19 (Atoms and partitions)**.** *Given a family $\mathcal{A} = (A_1, \ldots, A_m) \subseteq [n]^m$, the* atoms *generated by $\mathcal{A}$ are the (at most) 2^m distinct sets of the form $\bigcap_{r=1}^m C_r$, where $C_r \in \{A_r, [n] \setminus A_r\}$. The family of all such atoms, denoted $\mathrm{At}(\mathcal{A})$, is the* partition *generated by $\mathcal{A}$.*

This definition essentially captures "all sets (besides the A_i's) about which something can be learnt from querying the oracle on the sets of $\mathcal{A}$." Now, given such a sequence of queries $\mathcal{A} = (A_1, \ldots, A_m)$ and pairs of samples $\mathbf{s} = ((s_1^{(1)}, s_1^{(2)}), \ldots, (s_m^{(1)}, s_m^{(2)})) \in A_1^2 \times \cdots \times A_m^2$, we would like to summarize "all the label-invariant information available to an algorithm that obtains $((s_1^{(1)}, s_1^{(2)}), \ldots, (s_m^{(1)}, s_m^{(2)}))$ upon querying $A_1, \ldots, A_m$ for p and q." This calls for the following definition:

Definition 20 (m-configuration)**.** *Given $\mathcal{A} = (A_1, \ldots, A_m)$ and $\mathbf{s} = ((s_j^{(1)}, s_j^{(2)}))_{1 \le j \le m}$ as above, the* m-configuration *of $\mathbf{s}$ consists of the $6m^2$ bits indicating, for all $1 \le i, j \le m$, whether*

- *$s_i^{(\alpha)} = s_j^{(\beta)}$, for $\alpha, \beta \in \{1, 2\}$; and* *(relations between samples)*

[5] Recall that a property is label-invariant (or *symmetric*) if it is closed under relabeling of the elements of the support. More precisely, a property of distributions (resp. pairs of distributions) $\mathcal{C}$ is label-invariant if for any distribution $p \in \mathcal{C}$ (resp. $(p, q) \in \mathcal{C}$) and permutation σ of $[n]$, one has $p \circ \sigma \in \mathcal{C}$ (resp. $(p \circ \sigma, q \circ \sigma) \in \mathcal{C}$).

- $s_i^{(\alpha)} \in A_j$, *for* $\alpha \in \{1,2\}$. *(relations between samples and query sets)*

In other terms, it summarizes which is the unique atom $S_i \in \mathrm{At}(\mathcal{A})$ *that contains* $s_i^{(\alpha)}$*, and what collisions between samples have been observed.*

As aforementioned, the key idea is to argue that, without loss of generality, one can restrict one's attention to algorithms that only have access to m-configurations, and generate their queries in a specific (albeit adaptive) fashion:

Definition 21 (Core adaptive tester)**.** *A* core adaptive distribution tester *for pairs of distributions is an algorithm* $\mathcal{T}$ *that acts as follows.*

- *In the* i*-th phase, based only on its own internal randomness and the configuration of the previous queries* $A_1, \ldots, A_{i-1}$ *and samples obtained* $(s_1^{(1)}, s_1^{(2)}), \ldots, (s_{i-1}^{(1)}, s_{i-1}^{(2)})$ *– whose labels it does not actually know,* $\mathcal{T}$ *provides:*
  - *a number* ζ_i^A *for each* $A \in \mathrm{At}(A_1, \ldots, A_{i-1})$*, between* 0 *and* $|A \setminus \{s_j^{(1)}, s_j^{(2)}\}_{1 \le j \le i-1}|$ *(How many* fresh, not-already-seen *elements of each particular atom* A *should be included in the next query.)*
  - *sets* $K_i^{(1)}, K_i^{(2)} \subseteq \{1, \ldots, i-1\}$ *(Which of the samples* $s_1^{(k)}, \ldots, s_{i-1}^{(k)}$ *will be included in the next query. The labels of these samples are unknown, but are indexed by the index of the query which returned them.)*
- *based on these specifications, the next query* A_i *is drawn (but not revealed to* $\mathcal{T}$*) by*
  - *drawing uniformly at random a set* Λ_i *in*

$$\left\{ \Lambda \subseteq [n] \setminus \{s_j^{(1)}, s_j^{(2)}\}_{1 \le j \le i-1} \ : \ \forall A \in \mathrm{At}(A_1, \ldots, A_{i-1}),\ |\Lambda \cap A| = \zeta_i^A \right\} . \quad (6.1)$$

    That is, among all sets, containing only "fresh elements," whose intersection with each atom contains as many elements as $\mathcal{T}$ *requires.*
  - *adding the selected previous samples to this set:*

$$\Gamma_i \triangleq \left\{ s_j^{(1)} \ : \ j \in K_i^{(1)} \right\} \cup \left\{ s_j^{(2)} \ : \ j \in K_i^{(2)} \right\} ; \qquad A_i \triangleq \Lambda_i \cup \Gamma_i . \quad (6.2)$$

This results in a set A_i, not fully known to $\mathcal{T}$ besides the samples it already got and decided to query again; in which the labels *of the fresh elements are unknown, but the* proportions *of elements belonging to each atom are known.*

- *samples $s_i^{(1)} \sim (p)_{A_i}$ and $s_i^{(2)} \sim (q)_{A_i}$ are drawn (but not disclosed to $\mathcal{T}$). This defines the i-configuration of $A_1, \ldots, A_i$ and $(s_1^{(1)}, s_1^{(2)}), \ldots, (s_i^{(1)}, s_i^{(2)})$, which is revealed to $\mathcal{T}$. Put differently, the algorithm only learns (i) to which of the A_ℓ's the new sample belongs, and (ii) if it is one of the previous samples, in which stage(s) and for which of p, q it has already seen it.*

After $m = m(\varepsilon, n)$ such stages, $\mathcal{T}$ outputs either yes *or* no*, based only on the configuration of $A_1, \ldots, A_m$ and $(s_1^{(1)}, s_1^{(2)}), \ldots, (s_m^{(1)}, s_m^{(2)})$ (which is all the information it ever had access to).*

Note that in particular, $\mathcal{T}$ does not know the labels of samples it got, nor the actual queries it makes: it knows all about their sizes and sizes of their intersections, but not the actual "identity" of the elements they contain.

On the use of Yao's Principle in our lower bounds We recall Yao's Principle (e.g., see Chapter 2.2 of [MR95]), a technique which is ubiquitous in the analysis of randomized algorithms. Consider a set S of instances of some problem: what this principle states is that the worst-case expected cost of a randomized algorithm on instances in S is lower-bounded by the expected cost of the best deterministic algorithm on an instance drawn randomly from S.

As an example, we apply it in a standard way in Section 6.6: instead of considering a randomized algorithm working on a fixed instance, we instead analyze a *deterministic* algorithm working on a *random* instance. (We note that, importantly, the randomness in the samples returned by the COND oracle is "external" to this argument, and these samples behave identically in an application of Yao's Principle.)

On the other hand, our application in Section 6.3 is slightly different, due to our use of adaptive core testers. Once again, we focus on deterministic algorithms working on random instances, and the randomness in the samples is external and therefore unaffected by Yao's

Principle. However, we stress that the randomness in the choice of the set Λ_i is also external to the argument, and therefore unaffected – similar to the randomness in the samples, the algorithm has no control here. Another way of thinking about this randomness is via another step in the distribution over instances: after an instance (which is a pair of distributions) is randomly chosen, we permute the labels on the elements of the distribution's domain uniformly at random. We note that since the property in question is label-invariant, this does not affect its value. We can then use the adaptive core testing model as stated above for ease of analysis, observing that this can be considered an application of the principle of deferred decisions (as in Chapter 3.5 of [MR95]).

6.3 A Lower Bound for Adaptive Equivalence Testing

We prove Theorem 50, a lower bound on the sample complexity of testing equivalence between unknown distributions:

Theorem 50 (Adaptive Equivalence Testing Lower Bound)**.** *Any algorithm which, given* COND *access to unknown distributions* p, q *on* $[n]$*, distinguishes between the cases* $p = q$ *and* $d_{\mathrm{TV}}(p, q) \geq 1/4$ *with probability at least* $2/3$ *must make at least* $\Omega\left(\sqrt{\log \log n}\right)$ *queries.*

We construct two priors $\mathcal{Y}$ and $\mathcal{N}$ over *pairs* of distributions (p, q) over $[n]$. $\mathcal{Y}$ is a distribution over pairs of distributions of the form (p, p), namely the case when the distributions are identical. Similarly, $\mathcal{N}$ is a distribution over (p, q) with $d_{\mathrm{TV}}(p, q) \geq \frac{1}{4}$. We then show that no algorithm making $O\left(\sqrt{\log \log n}\right)$ queries to $\mathsf{COND}_p, \mathsf{COND}_q$ can distinguish between a draw from $\mathcal{Y}$ and $\mathcal{N}$ with constant probability (over the choice of (p, q), the randomness in the samples it obtains, and its internal randomness).

We describe the construction of $\mathcal{Y}$ and $\mathcal{N}$ in Section 6.3.1, and provide a detailed analysis in Section 6.3.2.

6.3.1 Construction

We now summarize how a pair of distribution is constructed under $\mathcal{Y}$ and $\mathcal{N}$. (Each specific step will be described in more detail in the subsequent paragraphs.)

1. **Effective Support**

   (a) Pick k_b from the set $\{0, 1, \ldots, \frac{1}{2}\log n\}$ at random.

   (b) Let $b = 2^{k_b}$ and $k \triangleq b \cdot n^{1/4}$.

2. **Buckets**

   (a) Choose ρ and r such that $\sum_{i=1}^{2r} \rho^i = n^{1/4}$.

   (b) Divide $\{1, \ldots, k\}$ into intervals $B_1, \ldots, B_{2r}$ with $|B_i| = b \cdot \rho^i$.

3. **Distributions**

   (a) For each $i \in [2r]$, assign probability mass $\frac{1}{2r}$ uniformly over B_i to generate distribution p.

   (b) For each $i \in [r]$ independently, pick π_i to be a Bernoulli with $\Pr(\pi_i = 0) = \frac{1}{2}$; if $\pi_i = 0$ then assign probability mass $\frac{1}{4r}$ and $\frac{3}{4r}$ over B_{2i-1} and B_{2i} respectively, else $\frac{3}{4r}$ and $\frac{1}{4r}$ respectively. This generates a distribution q.

4. **Support relabeling**

   (a) Pick a permutation $\sigma \in S_n$ of the *total* support n.

   (b) Relabel the symbols of D_1 and D_2 according to σ.

5. **Output:** Generate (p, p) for $\mathcal{Y}$, and (p, q) otherwise.

We now describe the various steps of the construction in greater detail.

Effective support. Both p and q, albeit distributions on $[n]$, will have (common) *sparse* support. The support size is taken to be $k \triangleq b \cdot n^{1/4}$. Note that, from the above definition, k is chosen uniformly at random from products of $n^{1/4}$ with powers of 2, resulting in values in $[n^{1/4}, n^{3/4}]$.

In this step b will act as a random scaling factor. The objective of this random scaling is to induce uncertainty in the algorithm's knowledge of the true support size of the distributions, and to prevent it from leveraging this information to test equivalence. In fact one can verify

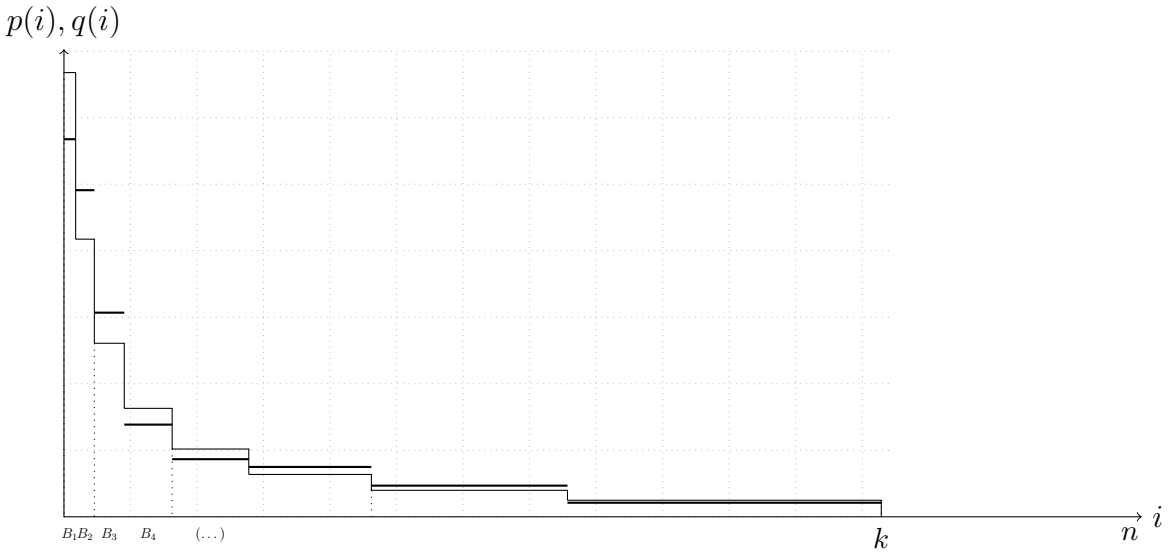


Figure 6-1: A no-instance (p, q) (before permutation).

that the class of distributions induced for a single value of b, namely all distributions have the same value of k, then one can distinguish the $\mathcal{Y}$ and $\mathcal{N}$ cases with only $O(1)$ conditional queries. The test would (roughly) go as follows. Since $|B_i|$ is known, one can choose a random subset S of the domain which (with high probability) has no intersection with B_i for $i \leq 2r-2$, a $O(1)$ size intersection with B_{2r-1}, and a $O(\rho)$ size intersection with B_{2r}. Perform $O(1)$ conditional queries over the set S, for both distributions. Given these queries, we can then identify which elements of S belong to B_{2r-1} or B_{2r} – namely, those which occur at most once belong to B_{2r}, and those which occur at least twice belong to B_{2r-1}. In a $\mathcal{Y}$ instance, then in both distributions, a $1/2$ fraction of queries will belong to B_{2r-1}, whereas in a $\mathcal{N}$ instance, one distribution will have either a $1/4$ or $3/4$ fraction of queries in B_{2r-1}, allowing us to distinguish the two cases.

Buckets. Our construction is inspired by the lower bound of Canonne, Ron, and Servedio [CRS14, Theorem 8] for the more restrictive PAIRCOND access model. We partition the support into $2r$ consecutive intervals (henceforth referred to as *buckets*) $B_1, \dots, B_{2r}$, where the size of the i-th bucket is $b\rho^i$. We note that r and ρ will be chosen such that $\sum_{i=1}^{2r} b\rho^i = bn^{1/4}$, i.e., the buckets fill the effective support.

Distributions. We output a pair of distributions (p, q). Each distribution that we construct is uniform within any particular bucket B_i. In particular, the first distribution assigns the same mass $1/2r$ to each bucket. Therefore, points within B_i have the same probability mass $1/2rb\rho^i$. For the $\mathcal{Y}$ case, the second distribution is identical to the first. For the $\mathcal{N}$ case, we pair buckets in r consecutive *bucket-pairs* $\Pi_1, \ldots, \Pi_r$, with $\Pi_i = B_{2i-1} \cup B_{2i}$. For the second distribution q, we consider the same buckets as p, but repartition the mass $1/r$ *within* each Π_i. More precisely, in each pair, one of the buckets gets now total probability mass $1/4r$ while the other gets $3/4r$ (so that the probability of every point is either decreased by a factor $1/2$ or increased by $3/2$). The choice of which goes up and which goes down is done uniformly and independently at random for each bucket-pair determined by the random choices of π_i's.

Random relabeling. The final step of the construction randomly relabels the symbols, namely is a random injective map from $[k]$ to $[n]$. This is done to ensure that no information about the individual symbol labels can be used by the algorithm for testing. For example, without this the algorithm can consider a few symbols from the first bucket and distinguish the $\mathcal{Y}$ and $\mathcal{N}$ cases. As mentioned in Section 6.2, for ease of analysis, the randomness in the choice of the permutation is, in some sense, deferred to the randomness in the choice of Λ_i during the algorithm's execution.

Summary. A no-instance (p, q) is thus defined by the following parameters: the support size k, the vector $(\pi_1, \ldots, \pi_r) \in \{0, 1\}^r$ (which only impacts q), and the final permutation σ of the domain. A yes-instance (p, p) follows an identical process, however, $\vec{\pi}$ has no influence on the final outcome. See Figure 6-1 for an illustration of such a (p, q) when σ is the identity permutation and thus the distribution is supported over the first k natural numbers.

Values for ρ and r. By setting $r = \log n/(8 \log \rho) + O(1)$, we have as desired $\sum_{i=1}^{2r} |B_i| = k$ and there is a factor $(1 + o(1))n^{1/4}$ between the height of the first bucket B_1 and the one of the last, B_{2r}. It remains to choose the parameter ρ itself; we shall take it to be $2^{\sqrt{\log n}}$, resulting in $r = \frac{1}{8}\sqrt{\log n} + O(1)$. (Note that for the sake of the exposition, we ignore technical details such as the rounding of parameters, e.g. bucket sizes; these can be easily taken care

of at the price of cumbersome case analyses, and do not bring much to the argument.)

6.3.2 Analysis

We now prove our main lower bound, by analyzing the behavior of core adaptive testers (as per Definition 21) on the families $\mathcal{Y}$ and $\mathcal{N}$ from the previous section. In Section 6.3.2.1, we argue that, with high probability, the sizes of the queries performed by the algorithm satisfy some specific properties. Conditioned upon this event, in Section 6.3.2.2, we show that the algorithm will get similar information from each query, whether it is running on a yes-instance or a no-instance.

Before moving to the heart of the argument, we state the following fact, straightforward from the construction of our no-instances:

Fact 2. *For any (p, q) drawn from $\mathcal{N}$, one has $d_{\mathrm{TV}}(p, q) = 1/4$.*

Moreover, as allowing more queries can only increase the probability of success, we hereafter focus on a core adaptive tester that performs exactly $q = \frac{1}{10}\sqrt{\log \log n}$ (adaptive) queries; and will show that it can only distinguish between yes and no-instances with probability $o(1)$.

6.3.2.1 Banning "bad queries"

As mentioned in Section 6.3.1, the draw of a yes or no-instance involves a random scaling of the size of the support of the distributions, meant to "blind" the testing algorithm. Recall that a testing algorithm is specified by a decision tree, which at step i, specifies how many unseen elements from each atom to include in the query ($\{\zeta_i^A\}$) and which previously seen elements to include in the query (sets $K_i^{(1)}, K_i^{(2)}$, as defined in Section 6.2), where the algorithm's choice depends on the observed configuration at that time. Note that, using Yao's Principle (as discussed in Section 6.2), these choices are deterministic for a given configuration – in particular, we can think of all $\{\zeta_i^A\}$ and $K_i^{(1)}, K_i^{(2)}$ in the decision tree as being fixed. In this section, we show that all ζ_i^A values satisfy with high probability some particular conditions with respect to the choice of distribution, where the randomness is over the choice of the support size.

First, we recall an observation from [CFGM13], though we modify it slightly to apply to configurations on pairs of distributions and we apply a slightly tighter analysis. This essentially limits the number of states an algorithm could be in by a function of how many queries it makes.

Proposition 15. *The number of nodes in a decision tree corresponding to a m-sample algorithm is at most 2^{6m^2+1}.*

Proof. As mentioned in Definition 20, an i-configuration can be described using $6i^2$ bits, resulting in at most 2^{6i^2} i-configurations. Since each i-configuration leads us to some node on the i-th level of the decision tree, the total number of nodes can be upper bounded by summing over the number of i-configurations for i ranging from 0 to m, giving us the desired bound. □

For the sake of the argument, we will introduce a few notions applying to the *sizes* of query sets: namely, the notions of a number being *small*, *large*, or *stable*, and of a vector being *incomparable*. Roughly speaking, a number is small if a uniformly random set of this size does not, in expectation, hit the largest bucket B_{2r} – in other words, the set is likely to be disjoint from the support. On the other hand, it is large if we expect such a set to intersect many bucket-pairs (i.e., a significant fraction of the support).

The definition of stable numbers is slightly more quantitative: a number β is stable if a random set of size β, for each bucket B_i, is either disjoint from B_i or has an intersection with B_i of size close to the expected value. In the latter case, we say the set *concentrates* over B_i. Finally, a vector of values (β_j) is incomparable if the union of random sets $S_1, \ldots, S_m$ of sizes $\beta_1, \ldots, \beta_m$ contains (with high probability) an amount of mass $p\left(\bigcup_j S_j\right)$ which is either much smaller or much larger than the probability $p(s)$ of any single element s.
We formalize these concepts in the definitions below. To motivate them, it will be useful to bear in mind that, from the construction described in Section 6.3.1, the expected intersection of a uniform random set of size β with a bucket B_i is of size $\beta b\rho^i/n$; while the expected probability mass from B_i it contains (under either p or q) is $\beta/2rn$.

Definition 22. *Let χ be an integer, and let $\varphi = \Theta(\chi^{5/2})$. A number β is said to be* small *if $\beta < \frac{n}{b\rho^{2r}}$; it is* large *(with relation to some integer χ) if $\beta \geq \frac{n}{b\rho^{2r-2\varphi}}$.*

Note that the latter condition equivalently means that, in expectation, a set of large size will intersect at least $\varphi + 1$ bucket-pairs (as it hits an expected $2\varphi + 1$ buckets, since $\beta|B_{2r-2\varphi}|/n \geq 1$). From the above definitions we get that, with high probability, a random set of any fixed size will in expectation either hit many or no buckets:

Proposition 16. *A number is either small or large with probability* $1 - O\left(\frac{\varphi \log \rho}{\log n}\right)$.

Proof. A number β is neither large nor small if $\frac{\rho^{2\varphi} n}{\beta \rho^{2r}} \leq b \leq \frac{n}{\beta \rho^{2r}}$. The ratio of the endpoints of the interval is $\rho^{2\varphi}$. Since $b = 2^{k_b}$, this implies that at most $\log \rho^{2\varphi} = 2\varphi \log \rho$ values of k_b could result in a fixed number falling in this range. As there are $\Theta(\log n)$ values for k_b, the proposition follows. □

The next definition characterizes the sizes of query sets for which the expected intersection with any bucket is either close to 0 (less than $1/\alpha$, for some threshold α), or very big (more than α). (It will be helpful to keep in mind that we will eventually use this definition with $\alpha = \text{poly}(m)$.)

Definition 23. *A number β is said to be* α-stable *(for $\alpha \geq 1$) if, for each $j \in [2r]$, $\beta \notin \left[\frac{n}{\alpha b \rho^j}, \frac{\alpha n}{b \rho^j}\right]$. A vector of numbers is said to be α-stable if all numbers it contains are α-stable.*

Proposition 17. *A number is α-stable with probability* $1 - O\left(\frac{r \log \alpha}{\log n}\right)$.

Proof. Fix some $j \in [2r]$. A number β does not satisfy the definition of α-stability for this j if $\frac{n}{\alpha \beta \rho^j} \leq b \leq \frac{n\alpha}{\beta \rho^j}$. Since $b = 2^{k_b}$, this implies that at most $\log 2\alpha$ values of k_b could result in a fixed number falling in this range. Noting that there are $\Theta(\log n)$ values for k_b and taking a union bound over all $2r$ values for j, the proposition follows. □

The following definition characterizes the sizes of query sets which have a probability mass far from the probability mass of any individual element. (For the sake of building intuition, the reader may replace ν in the following by the parameter b of the distribution.)

Definition 24. *A vector of numbers $(\beta_1, \ldots, \beta_\ell)$ is said to be* (α, τ)-incomparable with respect to ν *(for $\tau \geq 1$) if the two following conditions hold.*

- *$(\beta_1, \ldots, \beta_\ell)$ is α-stable.*

- *Let Δ_j be the minimum $\Delta \in \{0, \ldots, 2r\}$ such that $\frac{\beta_j \nu \rho^{2r-\Delta}}{n} \leq \frac{1}{\alpha}$, or $2r$ if no such Δ exists. For all $i \in [2r]$, $\frac{1}{2rn} \sum_{j=1}^{\ell} \beta_j \Delta_j \notin \left[\frac{1}{\tau 2r\nu\rho^i}, \frac{\tau}{2r\nu\rho^i}\right]$.*

Recall from the definition of α-stability of a number that a random set of this size either has essentially no intersection with a bucket or "concentrates over it" (i.e., with high probability, the probability mass contained in the intersection with this bucket is very close to the expected value). The above definition roughly captures the following. For any j, Δ_j is the number of buckets that will concentrate over a random set of size β_j. The last condition asks that the total probability mass from p (or q) enclosed in the union of m random sets of size $\beta_1, \ldots, \beta_\ell$ be a multiplicative factor of τ from the individual probability weight $1/2rb\rho^i$ of a single element from any of the $2r$ buckets.

Proposition 18. *Given that a vector of numbers of length ℓ is α-stable, it is (α, m^2)-incomparable with respect to b with probability at least $1 - O\left(\frac{r \log m}{\log n}\right)$.*

Proof. Fix any vector $(\beta_1, \ldots, \beta_\ell)$. By the definition above, for each value b such that $(\beta_1, \ldots, \beta_\ell)$ is α-stable, we have

$$\beta_j \cdot \frac{\alpha\rho^{2r}}{n} \leq \frac{\rho^{\Delta_j}}{b} < \beta_j \cdot \frac{\alpha\rho^{2r+1}}{n}, \quad j \in [\ell]$$

or, equivalently,

$$\frac{\log \frac{\alpha\beta_j}{n}}{\log \rho} + 2r + \frac{\log b}{\log \rho} \leq \Delta_j < \frac{\log \frac{\alpha\beta_j}{n}}{\log \rho} + 2r + \frac{\log b}{\log \rho} + 1, \quad j \in [\ell].$$

Writing $\lambda_j \triangleq \frac{\log \frac{\alpha\beta_j}{n}}{\log \rho} + 2r$ for $j \in [\ell]$, we obtain that

$$\sum_{j=1}^{\ell} \beta_j \Delta_j b = b \sum_{j=1}^{\ell} \beta_j (\lambda_j + O(1)) + \frac{b \log b}{\log \rho} \sum_{j=1}^{\ell} \beta_j. \tag{6.3}$$

- If it is the case that $\log \rho \cdot \sum_{j=1}^{\ell} \beta_j (\lambda_j + O(1)) \ll \log b \cdot \sum_{j=1}^{\ell} \beta_j$. Then, for any fixed $i \in [2r]$, to meet the second item of the definition of incomparability we need $\sum_{j=1}^{\ell} \beta_j \Delta_j b \notin [n/(200m\rho^i), 200mn/\rho^i]$. This is essentially, with the assumption (6.3)

above, requiring that

$$b \log b \notin \left[\frac{n \log \rho}{2m^2 \rho^i \sum_{j=1}^{\ell} \beta_j}, \frac{2m^2 n \log \rho}{\rho^i \sum_{j=1}^{\ell} \beta_j} \right].$$

Recalling that $b \log b = k_b 2^{k_b}$, this means that $O(\log m / \log\log m)$ values of k_b are to be ruled out. (Observe that this is the number of possible "bad values" for b without the condition from the case distinction above; since we add an extra constraint on b, there are at most this many values to avoid.)

- Conversely, if $\log \rho \cdot \sum_{j=1}^{\ell} \beta_j(\lambda_j + O(1)) \gg \log b \cdot \sum_{j=1}^{\ell} \beta_j$ the requirement becomes

$$b \notin \left[\frac{n \log \rho}{2m^2 \rho^i \sum_{j=1}^{\ell} \beta_j(\lambda_j + O(1))}, \frac{2m^2 n \log \rho}{\rho^i \sum_{j=1}^{\ell} \beta_j(\lambda_j + O(1))} \right].$$

ruling out this time $O(\log m)$ values for k_b.

- Finally, the two terms are comparable only if $\log b = \Theta\Big(\log \rho \cdot \sum_{j=1}^{\ell} \beta_j(\lambda_j + O(1)) \cdot \Big(\sum_{j=1}^{\ell} \beta_j \Big)^{-1} \Big)$; given that $\log b = k_b$, this rules out this time $O(1)$ values for k_b.

A union bound over the $2r$ possible values of i, and the fact that k_b can take $\Theta(\log n)$ values, complete the proof. □

We put these together to obtain the following lemma.

Lemma 52. *With probability at least* $1 - O\left(\frac{2^{6m^2+m}(r \log \alpha + \varphi \log \rho) + 2^{6m^2}(r \log m)}{\log n} \right)$*, the following holds for the decision tree corresponding to a* m*-query algorithm:*

- *the size of each atom is* α*-stable and either large or small;*
- *the size of each atom, after excluding elements we have previously observed,*[6] *is* α*-stable and either large or small;*
- *for each* i*, the vector* $(\zeta_i^A)_{A \in \mathrm{At}(A_1,\dots,A_i)}$ *is* (α, m^2)*-incomparable (with respect to* b*).*

[6]More precisely, we mean to say that for each $i \le m$, for every atom A defined by the partition of $(A_1, \dots, A_i)$, the values k_i^A and $|A \setminus \{s_1^{(1)}, s_1^{(2)}, \dots, s_{i-1}^{(1)}, s_{i-1}^{(2)}\}| - k_i^A$ are α-stable and either large or small;

Proof. From Proposition 15, there are at most 2^{6m^2+1} tree nodes, each of which contains one vector $(\zeta_i^A)_A$, and at most 2^m atom sizes. The first point follows from Propositions 16 and 17 and applying the union bound over all $2^{6m^2+1} \cdot 2 \cdot 2^m$ sizes, where we note the additional factor of 2 comes from either including or excluding the old elements. The latter point follows from Proposition 18 and applying the union bound over all 2^{6m^2+1} nodes in the tree (each containing a single ζ_i^A vector). □

6.3.2.2 Key lemma: bounding the variation distance between decision trees

In this section, we prove a key lemma on the variation distance between the distribution on leaves of any decision tree, when given access to either an instance from $\mathcal{Y}$ or $\mathcal{N}$. This lemma will in turn directly yield Theorem 50. Hereafter, we set the parameters α (the threshold for stability), φ (the parameter for smallness and largeness) and γ (an accuracy parameter for how well things concentrate over their expected value) as follows.[7] $\alpha \triangleq m^7$, $\varphi \triangleq m^{5/2}$ and $\gamma \triangleq 1/\varphi = m^{-5/2}$. (Recall further that $m = \frac{1}{10}\sqrt{\log\log n}$.)

Lemma 53. *Conditioned on the events of Lemma 52, consider the distribution over leaves of any decision tree corresponding to a m-query adaptive algorithm when the algorithm is given a yes-instance, and when it is given a no-instance. These two distributions have total variation distance $o(1)$.*

Proof. This proof is by induction on $1 \leq i \leq m$. We will have three inductive hypotheses, $\boldsymbol{E_1}(t), \boldsymbol{E_2}(t)$, and $\boldsymbol{E_3}(t)$. Assuming all three hold for all $t < i$, we prove $\boldsymbol{E_1}(i)$. Additionally assuming $\boldsymbol{E_1}(i)$, we prove $\boldsymbol{E_2}(i)$ and $\boldsymbol{E_3}(i)$.

Roughly, the first inductive hypothesis states that the query sets behave similarly to as if we had picked a random set of that size. It also implies that whether or not we get an element we have seen before is "obvious" based on past observances and the size of the query we perform. The second states that we never observe two distinct elements from the same bucket-pair. The third states that the next sample is distributed similarly in either a yes-instance or a no-instance. Note that this distribution includes both features which

[7] This choice of parameters is not completely arbitrary: combined with the setting of m, r and ρ, they ensure a total bound $o(1)$ on variation distance and probability of "bad events" as well as a (relative) simplicity and symmetry in the relevant quantities.

our algorithm can observe (i.e., the atom which the sample belongs to and if it collides with a previously seen sample), as well as those which it can not (i.e., which bucket-pair the observed sample belongs to). It is necessary to show the latter, since the bucket-pair a sample belongs to may determine the outcome of future queries.

More precisely, the three inductive hypotheses are as follows:

- $\boldsymbol{E_1}(i)$: In either a yes-instance or a no-instance, the following occurs: For an atom S in the partition generated by $A_1, \ldots, A_i$, let $S' = S \setminus \{s_1^{(1)}, s_1^{(2)}, \ldots, s_{i-1}^{(1)}, s_{i-1}^{(2)}\}$. For every such S', let $\ell^{S'}$ be the largest index $\ell \in \{0, \ldots, 2r\}$ such that $\frac{|S'|b\rho^\ell}{n} \leq \frac{1}{\alpha}$, or 0 if no such ℓ exists. We claim that $\ell^{S'} \in \{0, \ldots, 2r - \varphi - 2\} \cup \{2r\}$, and say S' is small if $\ell^{S'} = 2r$ and large otherwise. Additionally:

  - for $j \leq \ell^{S'}$, $|S' \cap B_j| = 0$;
  - for $j > \ell^{S'}$, $|S' \cap B_j|$ lies in $[1 - i\gamma, 1 + i\gamma] \frac{|S'|b\rho^j}{n}$.

  Furthermore, let p_1 and p_2 be the probability mass contained in Λ_i and Γ_i, respectively. Then $\frac{p_1}{p_1+p_2} \leq O\left(\frac{1}{m^2}\right)$ or $\frac{p_2}{p_1+p_2} \leq O\left(\frac{1}{m^2}\right)$ (that is, either almost all the probability mass comes from elements which we have not yet observed, or almost all of it comes from previously seen ones).

- $\boldsymbol{E_2}(i)$: No two elements from the set $\{s_1^{(1)}, s_1^{(2)}, \ldots, s_i^{(1)}, s_i^{(2)}\}$ belong to the same bucket-pair.

- $\boldsymbol{E_3}(i)$: Let T_i^{yes} be the random variable representing the atoms and bucket-pairs[8] containing $(s_i^{(1)}, s_i^{(2)})$, as well as which of the previous samples they intersect with, when the i-th query is performed on a yes-instance, and define T_i^{no} similarly for no-instances. Then $d_{\mathrm{TV}}(T_i^{\mathsf{yes}}, T_i^{\mathsf{no}}) \leq O\left(1/m^2 + 1/\rho + \gamma + 1/\varphi\right) = o(1)$.

We will show that $\boldsymbol{E_1}(i)$ holds with probability $1 - O\left(2^i \exp(-2\gamma^2\alpha/3)\right)$ and $\boldsymbol{E_2}(i)$ holds with probability $1 - O\left(i/\varphi\right)$. Let T^{yes} be the random variable representing the m-configuration and the bucket-pairs containing each of the observed samples in a yes-instance, and define T^{no} similarly for a no-instance. We note that this random variable determines which leaf of

[8] If a sample $s_i^{(k)}$ does not belong to any bucket (if the corresponding i-th query did not intersect the support), it is marked in T_i^{yes} with a "dummy label" to indicate so.

the decision tree we reach, which $\boldsymbol{E_3}(m)$ bounds. We can then take a union bound over all $i \in [m]$ to upper bound the probability that $\boldsymbol{E_1}(i)$ and $\boldsymbol{E_2}(i)$ do not hold, and use $\boldsymbol{E_3}(i)$ and the coupling interpretation of total variation distance to upper bound the probability that T_i^{yes} and T_i^{no} ever differ. Any of these "failure" events happens with probability at most $O(2^m \exp(-\frac{2\gamma^2\alpha}{3}) + \frac{m^2}{\varphi} + \frac{1}{m} + \frac{m}{\rho} + m\gamma + \frac{m}{\varphi}) = o(1)$ (from our choice of α, γ, φ). This upper bounds the total variation distance between T^{yes} and T^{no}, giving the desired result.

We proceed with the inductive proofs of $\boldsymbol{E_1}(i)$, $\boldsymbol{E_2}(i)$, and $\boldsymbol{E_3}(i)$, noting that the base cases hold trivially for all three of these statements. Throughout this proof, recall that Λ_i is the set of unseen support elements which we query, and Γ_i is the set of previously seen support elements which we query.

Lemma 54. *Assume that $\boldsymbol{E_1}(t), \boldsymbol{E_2}(t), \boldsymbol{E_3}(t)$ hold for all $1 \le t \le i-1$. Then $\boldsymbol{E_1}(i)$ holds with probability at least $1 - O\left(2^i \exp\left(-\frac{2\gamma^2\alpha}{3}\right)\right) = 1 - 2^{i-\Omega(m^2)}$.*

Proof. We start with the first part of $\boldsymbol{E_1}(i)$ (the statement prior to "Furthermore"). Let S (and the corresponding S') be any atom as in $\boldsymbol{E_1}(i)$. First, we note that $\ell^{S'} \in \{0, \dots, 2r - \varphi - 2\} \cup \{2r\}$ since we are conditioning on Lemma 52: $|S'|$ is α-stable and either large or small, which enforces this condition.

Next, suppose S' is contained in some other atom T generated by $A_1, \dots, A_{i-1}$, and let $T' = T \setminus \{s_1^{(1)}, s_1^{(2)}, \dots, s_{i-1}^{(1)}, s_{i-1}^{(2)}\}$. Since $|S'| \le |T'|$, this implies that $\ell^{T'} \le \ell^{S'}$. We argue about $|T' \cap B_j|$ for three regimes of j:

- The first case is $j \le \ell^{T'}$. By the inductive hypothesis, $|T' \cap B_j| = 0$, so $|S' \cap B_j| = 0$ with probability 1.

- The next case is $\ell^{T'} < j \le \ell^{S'}$. Recall from the definition of a core adaptive tester that S' will be chosen uniformly at random from all subsets of T' of appropriate size. By the inductive hypothesis,

$$\frac{|T' \cap B_j|}{|T'|} \in [1 - (i-1)\gamma, 1 + (i-1)\gamma] \frac{b\rho^j}{n},$$

and therefore

$$\mathbf{E}\left[|S' \cap B_j|\right] \in [1-(i-1)\gamma, 1+(i-1)\gamma]\,\frac{|S'|b\rho^j}{n}, \;\; \text{implying } \mathbf{E}\left[|S' \cap B_j|\right] \leq \frac{2}{\alpha\rho^{\ell^{S'}-j}};$$

where the inequality is by the definition of $\ell^{S'}$ and using the fact that $(i-1)\gamma \leq 1$. Using a Chernoff bound for hypergeometric random variables (Lemma 3), and writing $\mu \triangleq \mathbf{E}\left[|S' \cap B_j|\right]$ for conciseness,

$$\begin{aligned}
\Pr[\,|S' \cap B_j| \geq 1\,] &= \Pr\left[\,|S' \cap B_j| \geq \left(1+\frac{1-\mu}{\mu}\right)\mu\,\right] \\
&\leq \exp\left(-\frac{(1-\mu)^2}{3\mu}\right) \\
&\leq \exp\left(-\frac{1}{12}\alpha\rho^{\ell^{S'}-j}\right),
\end{aligned}$$

where the second inequality holds because $\mu \leq 2/\alpha\rho^{\ell^{S'}-j}$ and $(1-\mu)^2 \geq 1/2$ for n sufficiently large.

- The final case is $j > \ell^{S'}$. As in the previous one,

$$\mathbf{E}\left[|S' \cap B_j|\right] \in [1-(i-1)\gamma, 1+(i-1)\gamma]\,\frac{|S'|b\rho^j}{n}, \;\; \text{implying } \mathbf{E}\left[|S' \cap B_j|\right] \geq \frac{\alpha\rho^{j-\ell^{S'}-1}}{2};$$

where the inequality is by the definition of $\ell^{S'}$, α-stability, and using the fact that $(i-1)\gamma \leq 1/2$. Using again a Chernoff bound for hypergeometric random variables (Lemma 3),

$$\begin{aligned}
\Pr\left[\,|S' \cap B_j| - \mathbf{E}\left[|S' \cap B_j|\right] \geq \gamma\frac{|S'|b\rho^j}{n}\,\right] &\leq \Pr[\,|S' \cap B_j| - \mathbf{E}\left[|S' \cap B_j|\right] \geq \gamma 2\mathbf{E}\left[|S' \cap B_j|\right]\,] \\
&\leq 2\exp\left(-\frac{(2\gamma)^2\mathbf{E}\left[|S' \cap B_j|\right]}{3}\right) \\
&\leq 2\exp\left(-\frac{2}{3}\gamma^2\alpha\rho^{j-\ell^{S'}-1}\right),
\end{aligned}$$

where the first inequality comes from $2(1-(i-1)\gamma) \geq 1$, the second is from Chernoff bound, and the third follows from $\mathbf{E}\left[|S' \cap B_j|\right] \geq \alpha\rho^{j-\ell^{S'}-1}/2$.

Since we wish to prove the statement for all buckets B_j simultaneously, we take a union bound over all j. Recall that we want to bound the probability that S' satisfies two conditions, for $j \le \ell^{S'}$, $|S' \cap B_j| = 0$; and for $j > \ell^{S'}$, $|S' \cap B_j|$ lies in $[1 - i\gamma, 1 + i\gamma] \frac{|S'|b\rho^j}{n}$. Using bounds from all three of these regimes of j, and a union bound, the probability that S' does not satisfy the conditions of $\boldsymbol{E_1}(i)$ is at most

$$\sum_{j \le \ell^{T'}} 0 + \sum_{\ell^{T'} < j \le \ell^{S'}} \exp\left(-\frac{1}{12}\alpha\rho^{\ell^{S'}-j}\right) + \sum_{j > \ell^{S'}} 2\exp\left(-\frac{2}{3}\gamma^2\alpha\rho^{j-\ell^{S'}-1}\right).$$

This probability is maximized when $\ell^{S'} = \ell^{T'} = 0$, in which case it is

$$\sum_{j=1}^{2r} 2\exp\left(-\frac{2}{3}\gamma^2\alpha\rho^{j-1}\right) \le \sum_{j=1}^{\infty} 2\exp\left(-\frac{2}{3}\gamma^2\alpha\rho^{j-1}\right) \le 3\exp\left(-\frac{2}{3}\gamma^2\alpha\right).$$

Taking a union bound over at most 2^i sets gives us the desired probability bound.

Finally, we prove the remainder of $\boldsymbol{E_1}(i)$ (the statement following "Furthermore"); this will follow from the definition of incomparability (Definition 24).

- First, we focus on Γ_i. Suppose that Γ_i contains at least one element with positive probability mass (if not, the statement trivially holds). Let p_2' be the probability mass of the heaviest element in Γ_i. Since our inductive hypothesis implies that Γ_i has no elements in the same bucket pair, the maximum possible value for p_2 is

$$\begin{aligned} p_2 &\le p_2' + \frac{3p_2'}{\rho} + \frac{3p_2'}{\rho^3} + \cdots \le p_2' + \frac{3p_2'}{\rho}\sum_{k=0}^{\infty}\frac{1}{\rho^{2k}} = \left(1 + \frac{3}{\rho}\frac{\rho^2}{\rho^2 - 1}\right)p_2' \\ &\le (1 + o(1))p_2' \end{aligned}$$

  Therefore, $p_2 \in [p_2', (1 + o(1))p_2']$. Supposing this heaviest element belongs to bucket j, we can say that $p_2 \in [\frac{1}{2}, (1 + o(1))\frac{3}{2}]\frac{1}{2rb\rho^j}$.

- Next, we focus on Λ_i. Consider some atom A, from which we selected k_A elements which have not been previously observed: call the set of these elements A'. In the first part of this proof, we showed that for each bucket B_k, either $|A' \cap B_k| = 0$ or $|A' \cap B_k| \in [1 - i\gamma, 1 + i\gamma]\,|A'|b\rho^k/n$. In the latter case, noting that $i\gamma \le \frac{1}{2}$ and that

the probability of an individual element in B_k is within $[1,3]\frac{1}{4rb\rho^k}$, the probability mass contained by $|A' \cap B_k|$ belongs to $[1,9]\frac{|A'|}{8rn}$. Recalling the definition of Δ_A as stated in Definition 24, as shown earlier in this proof, this non-empty intersection happens for exactly Δ_A buckets. Therefore, the total probability mass in Λ_i is in the interval $\left[\frac{1}{4}, \frac{9}{4}\right] \frac{1}{2rn} \sum_{A \in \mathrm{At}(A_1,\ldots,A_i)} \zeta_i^A \Delta_A$.

Recall that we are conditioning on Lemma 52 which states that the vector $(\zeta_i^A)_{A \in \mathrm{At}(A_1,\ldots,A_i)}$ is (α, m^2)-incomparable with respect to b. Applying this definition to the bounds just obtained on the probability masses in Λ_i and Γ_i gives Lemma 54. □

Lemma 55. *Assume that $\boldsymbol{E_1}(t), \boldsymbol{E_2}(t), \boldsymbol{E_3}(t)$ hold for all $1 \le t \le i-1$, and additionally $\boldsymbol{E_1}(i)$ holds. Then $\boldsymbol{E_2}(i)$ holds with probability at least $1 - O\left(\frac{i}{\varphi}\right)$.*

Proof. We focus on $s_i^{(1)}$. If $s_i^{(1)} \in \Gamma_i$, the conclusion is trivial, so suppose $s_i^{(1)} \in \Lambda_i$. From $\boldsymbol{E_1}(i)$, no small atom intersects any of the buckets, so let us condition on the fact that $s_i^{(1)}$ belongs to some large atom S. Since we want $s_i^{(1)}$ to fall in a distinct bucket-pair from $2(i-1)+1$ other samples, there are at most $2i-1$ bucket-pairs which $s_i^{(1)}$ should not land in. Using $\boldsymbol{E_1}(i)$, the maximum probability mass contained in the intersection of these bucket-pairs and S is $(1+i\gamma)(2i-1)|S|/rn$. Similarly, using the definition of a large atom, the minimum probability mass contained in S is $(1-i\gamma)\varphi|S|/rn$. Taking the ratio of these two terms gives an upper bound on the probability of breaking this invariant, conditioned on landing in S, as $O(i/\varphi)$, where we note that $\frac{1+i\gamma}{1-i\gamma} = O(1)$. Since the choice of which large atom was arbitrary, we can remove the conditioning. Taking a union bound for $s_i^{(1)}$ and $s_i^{(2)}$ gives the result. □

Lemma 56. *Assume that $\boldsymbol{E_1}(t), \boldsymbol{E_2}(t), \boldsymbol{E_3}(t)$ hold for all $1 \le t \le i-1$, and additionally $\boldsymbol{E_1}(i)$ holds. Then $\boldsymbol{E_3}(i)$ holds.*

Proof. We fix some setting of the intersection history, i.e. the configuration and the bucket-pairs the past elements belong to, and show that the results of the next query will behave similarly, whether the instance is a yes-instance or a no-instance. We note that, since we are assuming the inductive hypotheses hold, certain settings which violate these hypotheses are

not allowed. We also note that $s_i^{(1)}$ is distributed identically in both instances, so we focus on $s_i \triangleq s_i^{(2)}$ for the remainder of this proof.

First, we condition that, based on the setting of the past history, s_i will either come from Λ_i or Γ_i – this event happens with probability $1 - O(1/m^2)$.

Proposition 19. *In either a* yes*-instance or a* no*-instance, s_i will either come from Λ_i with probability $1 - O\left(\frac{1}{m^2}\right)$, or Γ_i with probability $1 - O\left(\frac{1}{m^2}\right)$, where the choice of which one is deterministic based on the fixed configuration and choice for the bucket-pairs of previously seen elements.*

Proof. This is simply a rephrasing of the portion of $\boldsymbol{E_1}(i)$ following "Furthermore." □

We now try to bound the total variation distance between T_i^{yes} and T_i^{no} conditioning on this event. In the case when it does not hold, we trivially bound the total variation distance by 1, incurring a cost of $O\left(1/m^2\right)$ to the total variation distance between the unconditioned variables. Since our target was for this quantity was $O\left(1/m^2 + 1/\rho + \gamma + 1/\varphi\right)$, it remains to show, in the conditioned space, that the total variation distance in either case is at most $O\left(1/\rho + \gamma + 1/\varphi\right) = O(1/m^{5/2})$. We break this into two cases, the first being when s comes from Γ_i. In this case, we incur a cost in total variation distance which is $O\left(1/\rho\right)$:

Proposition 20. *In either a* yes*-instance or a* no*-instance, condition that s_i comes from Γ_i. Then one of the following holds:*

- $|\Gamma_i \cap B_j| = 0$ *for all* $j \in [2r]$*, in which case s_i is distributed uniformly at random from the elements of Γ_i;*
- *or* $|\Gamma_i \cap B_j| \neq 0$ *for some* $j \in [2r]$*, in which case s_i will be equal to some $s \in \Gamma_i$ with probability $1 - O\left(1/\rho\right)$, where the choice of s is deterministic based on the fixed configuration and choice for the bucket-pairs of previously seen elements.*

Proof. The former case follows from the definition of the sampling model. For the latter case, let p be the probability mass of the heaviest element in Γ_i. Since our inductive hypothesis implies that Γ_i has no elements in the same bucket-pair, the maximum possible value for the

rest of the elements is

$$\frac{3p}{\rho} + \frac{3p}{\rho^3} + \frac{3p}{\rho^5} + \cdots \leq \frac{3p}{\rho} \sum_{k=0}^{\infty} \frac{1}{\rho^{2k}} = \frac{3p}{\rho} \frac{\rho^2}{\rho^2 - 1} = O\left(\frac{p}{\rho}\right).$$

Since the ratio of this value and p is $O(1/\rho)$, with probability $1 - O(1/\rho)$ the sample returned is the heaviest element in Γ_i. □

Finally, we examine the case when s comes from Λ_i:

Proposition 21. *Condition that s_i comes from Λ_i. Then either:*

- $|\Lambda_i \cap B_j| = 0$ *for all* $j \in [2r]$*, in which case* $d_{\mathrm{TV}}(T_i^{\mathsf{yes}}, T_i^{\mathsf{no}}) = 0$*;*
- *or* $|\Lambda_i \cap B_j| \neq 0$ *for some* $j \in [2r]$*, in which case* $d_{\mathrm{TV}}(T_i^{\mathsf{yes}}, T_i^{\mathsf{no}}) \leq O\left(\gamma + \frac{1}{\varphi}\right) = O\left(\frac{1}{m^{5/2}}\right)$

Proof. The former case follows from the definition of the sampling model – since Λ_i does not intersect any of the buckets, the sample will be labeled as such. Furthermore, the sample returned will be drawn uniformly at random from Λ_i, and the probability of each atom will be proportional to the cardinality of its intersection with Λ_i, in both the yes and the no-instances.

We next turn to the latter case. Let $\mathcal{X}$ be the event that, if the intersection of Λ_i and some atom A has a non-empty intersection with an odd number of buckets, then s_i does not come from the unpaired bucket. Note that $\boldsymbol{E_1}(i)$ and the definition of a large atom imply that an unpaired bucket can only occur if the atom intersects at least φ bucket-pairs: conditioned on the sample coming from a particular atom, the probability that it comes from the unpaired bucket is $O(1/\varphi)$. Since the choice of A was arbitrary, we may remove the conditioning, and note that $\Pr(\mathcal{X}) = 1 - O(1/\varphi)$.

Since

$$\begin{aligned} d_{\mathrm{TV}}(T_i^{\mathsf{yes}}, T_i^{\mathsf{no}}) &\leq d_{\mathrm{TV}}(T_i^{\mathsf{yes}}, T_i^{\mathsf{no}} \mid \mathcal{X}) \Pr(\mathcal{X}) + d_{\mathrm{TV}}(T_i^{\mathsf{yes}}, T_i^{\mathsf{no}} \mid \bar{\mathcal{X}}) \Pr(\bar{\mathcal{X}}) \\ &\leq d_{\mathrm{TV}}(T_i^{\mathsf{yes}}, T_i^{\mathsf{no}} \mid \mathcal{X}) + O(1/\varphi), \end{aligned} \tag{6.4}$$

it remains to show that $d_{\mathrm{TV}}(T_i^{\mathsf{yes}}, T_i^{\mathsf{no}} \mid \mathcal{X}) \leq O(\gamma)$.

First, we focus on the distribution over atoms, conditioned on $\mathcal{X}$. Let N^A be the number of bucket-pairs with which A intersects both buckets, i.e., conditioned on $\mathcal{X}$, the sample could come from $2N^A$ buckets, and let $N \triangleq \sum_{A \in \mathrm{At}(A_1,\ldots,A_i)} N^A$. By $\boldsymbol{E_1}(i)$, the maximum amount of probability mass that can be assigned to atom A is $\frac{(1+\gamma)|S|N^A/rn}{(1-\gamma)|S|N/rn}$, and the minimum is $\frac{(1-\gamma)|S|N^A/rn}{(1+\gamma)|S|N/rn}$, so the total variation distance in the distribution incurred by this atom is at most $O(\gamma N^A/N)$. Summing over all atoms, we get the desired result of $O(\gamma)$.

Finally, we bound the distance on the distribution over bucket-pairs, again conditioned on $\mathcal{X}$. By $\boldsymbol{E_1}(i)$ only large atoms will contain non-zero probability mass, so condition on the sample coming from some large atom A. Let N^A be the number of bucket-pairs with which A intersects both buckets, i.e., conditioned on $\mathcal{X}$, the sample could come from $2N^A$ buckets. Using $\boldsymbol{E_1}(i)$, the maximum amount of probability mass that can be assigned to any intersecting bucket-pair is $(1+\gamma)\frac{|A|}{rn}((1-\gamma)\frac{|A|}{rn}N^A)^{-1}$, and the minimum is $(1-\gamma)\frac{|A|}{rn}((1+\gamma)\frac{|A|}{rn}N^A)^{-1}$, so the total variation distance in the distribution incurred by this bucket-pair is at most $O(\gamma/N^A)$. Summing this difference over all N^A bucket-pairs, we get $\frac{2\gamma}{1-\gamma^2} = O(\gamma)$. Since the choice of large atom A was arbitrary, we can remove the conditioning on the choice of atom. The statement follows by applying the union bound on the distribution over bucket-pairs and the distribution over atoms. This concludes the proof of Proposition 21. □

We note that in both cases, the cost in total variation distance which is incurred is $O(\frac{1}{\rho} + \gamma + \frac{1}{\varphi})$, which implies $\boldsymbol{E_3}(i)$ – proving Lemma 56 . □

This concludes the proof of Lemma 53. □

With Lemma 53 in hand, the proof of the main theorem is straightforward:
Proof of Theorem 50: Conditioned on Lemma 52, Lemma 53 implies that the distribution over the leaves in a yes-instance vs. a no-instance is $o(1)$. Since an algorithm's choice to accept or reject depends deterministically on which leaf is reached, this bounds the difference between the conditional probability of reaching a leaf which accepts. Since Lemma 52 occurs with probability $1 - o(1)$, the difference between the unconditional probabilities is also $o(1)$.
□

6.4 An Upper Bound for Adaptive Support-Size Estimation

We prove our upper bound for constant-factor support-size estimation, Theorem 51:

Theorem 51 (Adaptive Support-Size Estimation)**.** *Let $\tau > 0$ be any constant. There exists an adaptive algorithm which, given* COND *access to an unknown distribution p on $[n]$ (guaranteed to have probability mass at least τ/n on every element of its support) and accuracy parameter $\varepsilon \in (0,1)$, makes $\tilde{O}\left((\log\log n)/\varepsilon^3\right)$ queries to the oracle*[9] *and outputs a value $\tilde{\omega}$ such that the following holds. With probability at least $2/3$, $\tilde{\omega} \in [\frac{1}{1+\varepsilon} \cdot \omega, (1+\varepsilon)\cdot\omega]$, where $\omega = |\operatorname{supp}(p)|$.*

Before describing and analyzing our algorithm, we shall need the following results, that we will use as subroutines: the first one will help us detecting when the support is already dense. The second, assuming the support is sparse enough, will enable us to find an element with zero probability mass, which can afterwards be used as a "reference" to verify whether any given element is inside or outside the support. Finally, the last one will use such a reference point to check whether a candidate support size σ is smaller or significantly bigger than the actual support size.

Lemma 57. *Given $\tau > 0$ and* COND *access to a distribution p such that each support element has probability at least τ/n, as well as parameters $\varepsilon \in (0,1/2), \delta \in (0,1)$, there exists an algorithm* TestSmallSupport *(Algorithm 15) that makes $\tilde{O}\left(1/(\tau\varepsilon^2) + 1/\tau^2\right)\cdot\log(1/\delta)$ queries to the oracle, and satisfies the following. (i) If $\operatorname{supp}(p) \geq (1-\varepsilon/2)n$, then it outputs* yes *with probability at least $1-\delta$; (ii) if $\operatorname{supp}(p) \leq (1-\varepsilon)n$, then it outputs* no *with probability at least $1-\delta$.*

Lemma 58. *Given* COND *access to a distribution p, an upper bound $t < n$ on $\operatorname{supp}(p)$, as well as parameter $\delta \in (0,1)$, there exists an algorithm* GetNonSupport *(Algorithm 16) that makes $\tilde{O}\left(\log^2\frac{1}{\delta}\log^{-2}\frac{n}{t}\right)$ queries to the oracle, and returns an element $r \in [n]$ such that $r \notin \operatorname{supp}(p)$ with probability at least $1-\delta$.*

[9] We remark that the constant in the $\tilde{O}$ depends polynomially on $1/\tau$.

Lemma 59. *Given* COND *access to a distribution p, inputs $\sigma \geq 2$ and $r \notin \operatorname{supp}(p)$, as well as parameters $\varepsilon \in (0, 1/2), \delta \in (0, 1)$, there exists an algorithm* IsAtMostSupportSize *(Algorithm 17) that makes $\tilde{O}\left(1/\varepsilon^2\right) \log(1/\delta)$ queries to the oracle, and satisfies the following. The algorithm returns either* yes *or* no*, and (i) if $\sigma \leq \operatorname{supp}(p)$, then it outputs* yes *with probability at least $1 - \delta$; (ii) if $\sigma > (1 + \varepsilon) \operatorname{supp}(p)$, then it outputs* no *with probability at least $1 - \delta$.*

We defer the proofs of these three lemmata, and for the time being, turn to the proof of the theorem.

Proof. The algorithm is given in Algorithm 14, and at a high-level works as follows: if first checks whether the support size is big (an $1 - O(\varepsilon)$ fraction of the domain), in which case it can already stop and return a good estimate. If this is not the case, however, then the support is sparse enough to efficiently find an element r *outside* the support, by taking a few uniform points, comparing and ordering them by probability mass (and keeping the lightest). This element r can then be used as a reference point in a (doubly exponential) search for a good estimate: for each guess $\tilde{\omega}$, a random subset S of size roughly $\tilde{\omega}$ is taken, a point x is drawn from p_S, and x is compared to r to check if $p(x) > 0$. If so, then S intersects the support, meaning that $\tilde{\omega}$ is an upper bound on ω; repeating until this is no longer the case results in an accurate estimate of ω.

Algorithm 14 EstimateSupport$_p$

1: **if** TestSmallSupport$_p(\varepsilon, \frac{1}{10})$ returns yes **then return** $\tilde{\omega} \leftarrow (1-\varepsilon^2)n$
2: **end if**
3: Call GetNonSupport$_p((1-\frac{\varepsilon}{2})n, \frac{1}{10})$ to obtain a non-support reference point r.
4: **for** j **from** 0 **to** $\log_{1+\varepsilon}\log_{1+\varepsilon} n$ **do**
5: Set $\tilde{\omega} \leftarrow (1+\varepsilon)^{(1+\varepsilon)^j}$.
6: Call IsAtMostSupportSize$_p(\tilde{\omega}, r, \varepsilon, \frac{1}{100\cdot(j+1)^2})$ to check if $\tilde{\omega}$ is an upper bound on ω.
7: **if** the call returned no **then**
8: Perform a binary search on $\{(1+\varepsilon)^{j-1}, \ldots, (1+\varepsilon)^j\}$ to find i^*, the smallest $i \geq 2$ such that IsAtMostSupportSize$_p((1+\varepsilon)^i, r, \varepsilon, \frac{1}{10(j+1)})$ returns no.
9: **return** $\tilde{\omega} \leftarrow (1+\varepsilon)^{i^*-1}$.
10: **end if**
11: **end for**

In the rest of this section, we formalize and rigorously argue the above. Conditioning on each of the calls to the subroutines TestSmallSupport, GetNonSupport and IsAtMostSupportSize being correct (which overall happens except with probability at most $1/10 + 1/10 + \sum_{j=1}^{\infty} 1/(100j^2) + 1/10 < 1/3$ by a union bound), we show that the output $\tilde{\omega}$ of EstimateSupport is indeed within a factor $(1+\varepsilon)$ of ω.

- If the test on Step 1 passes, then by Lemma 57 we must have $\text{supp}(p) > (1-\varepsilon)n$. Thus, the estimate we output is correct, as $[(1-\varepsilon)n, n] \subseteq [\tilde{\omega}/(1+\varepsilon), (1+\varepsilon)\tilde{\omega}]$.

- Otherwise, if it does not then by Lemma 57 it must be the case that $\text{supp}(p) < (1-\varepsilon/2)n$.

Therefore, if we reach Step 3 then $(1-\varepsilon/2)n$ is indeed an upper bound on ω, and GetNonSupport will return a point $r \notin \text{supp}(p)$ as expected. The analysis of the rest of the algorithm is straightforward: from the guarantee of IsAtMostSupportSize, the binary search will be performed for the first index j such that $\omega \in [(1+\varepsilon)^{(1+\varepsilon)^{j-1}}, (1+\varepsilon)^{(1+\varepsilon)^j}]$; and will be on a set of $(1+\varepsilon)^{j-1}$ values. Similarly, for the value i^* eventually obtained, it must be

the case that $(1+\varepsilon)^{i^*} > \omega$ (by contrapositive, as no was returned by the subroutine) but $(1+\varepsilon)^{i^*-1} \leq (1+\varepsilon)\omega$ (again, as the subroutine returned yes). But then, $\tilde{\omega} = (1+\varepsilon)^{i^*-1} \in (\omega/(1+\varepsilon), (1+\varepsilon)\omega]$ as claimed.

Query complexity. The query complexity of our algorithm originates from the following different steps:

- the call to TestSmallSupport, which from Lemma 57 costs $\tilde{O}\left(1/\varepsilon^2\right)$ queries;
- the call to GetNonSupport, on Step 3, that from the choice of the upper bound also costs $\tilde{O}\left(1/\varepsilon^2\right)$ queries;
- the (at most) $\log_{1+\varepsilon}\log_{1+\varepsilon} n = O\left((\log\log n)/\varepsilon\right)$ calls to IsAtMostSupportSize on Step 6. Observing that the query complexity of IsAtMostSupportSize is only $\tilde{O}\left(1/\varepsilon^2\right)\cdot\log(1/\delta)$, and from the choice of $\delta = \frac{1}{(j+1)^2}$ at the j-th iteration this step costs at most

  $$\tilde{O}\left(\frac{1}{\varepsilon^2}\right)\cdot\sum_{j=1}^{\log_{1+\varepsilon}\log_{1+\varepsilon} n} O\left(\log(j^2)\right) = \tilde{O}\left(\frac{1}{\varepsilon^2}\log_{1+\varepsilon}\log_{1+\varepsilon} n\right) = \tilde{O}\left(\frac{1}{\varepsilon^3}\log_{1+\varepsilon}\log_{1+\varepsilon} n\right)$$

  queries.
- Similarly, Step 8 results in at most $j \leq \log\log n$ calls to IsAtMostSupportSize with δ set to $1/(10(j+1))$, again costing $\tilde{O}\left(\frac{1}{\varepsilon^2}\right)\cdot\log j = \tilde{O}\left(\frac{1}{\varepsilon^2}\log_{1+\varepsilon}\log_{1+\varepsilon} n\right) = \tilde{O}\left(\frac{1}{\varepsilon^3}\log\log n\right)$ queries.

Gathering all terms, the overall query complexity is $\tilde{O}\left(\frac{\log\log n}{\varepsilon^3}\right)$, as claimed. □

Proof of Lemma 57. Hereafter, we assume without loss of generality that $\tau < 2$: indeed, if $\tau \geq 2$ then the support is of size at most $n/2$, and it suffices to output no to meet the requirements of the lemma. We will rely on the (easy) fact below, which ensures that any distribution with dense support and minimum non-zero probability τ/n put significant mass on "light" elements:

Fact 3. *Fix any $\varepsilon \in [0,1)$. Assume p satisfies both $\mathrm{supp}(p) \geq (1-\varepsilon)n$ and $p(x) \geq \tau/n$ for $x \in \mathrm{supp}(p)$. Then, setting $L_\varepsilon \triangleq \{x \in [n] \ : \ p(x) \in [\tau/n, 2/n]\}$, we have $|L_\varepsilon| \geq (1/2-\varepsilon)n$ and $p(L_\varepsilon) \geq (1/2-\varepsilon)\tau$.*

Proof. As the second claim follows directly from the first and the minimum mass of elements of L_ε, it suffices to prove that $|L_\varepsilon| \geq (1/2-\varepsilon)n$. This follows from observing that

$$1 = p([n]) \geq p([n] \setminus L_\varepsilon) \geq (|\,\mathrm{supp}(p)\,| - |L_\varepsilon|)\frac{2}{n} \geq 2(1-\varepsilon) - \frac{2|L_\varepsilon|}{n}$$

and rearranging the terms. □

Description and intuition. The algorithm (as described in Algorithm 15) works as follows: it first takes enough uniformly distributed samples $s_1, \ldots, s_\ell$ to get (with high probability) an accurate enough fraction of them falling in the support to distinguish between the two cases. The issue is now to detect those s_j's which indeed are support elements; note that we do not care about underestimating this fraction in case (b) (when the support is at most $(1-\varepsilon)n$, but importantly do not want to underestimate it in case (a) (when the support size is at least $(1-\varepsilon/2)n$). To perform this detection, we take constantly many samples *according to p* (which are therefore ensured to be in the support), and use pairwise conditional queries to sort them by increasing probability mass (up to approximation imprecision), and keep only the lightest of them, t. In case (a), we now from Fact 3 that with high probability our t has mass in $[1/n, 2/n]$, and will therefore be either much lighter than or comparable to *any* support element: this will ensure that in case (a) we do detect all of the s_j's that are in the support.

This also works in case (b), even though Fact 3 does not give us any guarantee on the mass of t. Indeed, either t turns out to be light (and then the same argument ensures us our estimate of the number of "support" s_j's is good), or t is too heavy – and then our estimate will end up being smaller than the true value. But this is fine, as the latter this only means we will reject the distribution (as we should, since we are in the small-support case).

Correctness. Let η be the fraction of the s_j's that are in the support of the distribution. By a multiplicative Chernoff bound and a suitable constant in our choice of ℓ, we get that

Algorithm 15 TESTSMALLSUPPORT$_p$

Require: COND access to p; accuracy parameter $\varepsilon \in (0, 1/2)$, threshold $\tau > 0$, probability of failure δ

Repeat the following $O(\log(1/\delta))$ times and output the majority vote.
loop
  Draw $\ell \triangleq \Theta\left(\frac{1}{\varepsilon^2}\right)$ independent samples $s_1, \ldots, s_\ell \sim \mathcal{U}_n$.
  Draw $k \triangleq \Theta\left(\frac{1}{\tau}\right)$ independent samples $t_1, \ldots, t_k \sim p$.
  for all $1 \le i < j \le k$ **do** ▷ Order the t_j's
    Call COMPARE$(\{t_i\}, \{t_j\}, \eta = \frac{1}{2}, K = 2, \frac{1}{4k^2})$ to get a 2-approx. ρ of $\frac{p(t_j)}{p(t_i)}$, High or Low.
    if COMPARE returned High or a value ρ **then**
      Record $t_i \preceq t_j$
    else
      Record $t_j \prec t_j$
    end if
  end for
  Set t to be (any of the) smallest t_j's, according to $\preceq$.
  for all $1 \le j \le \ell$ **do** ▷ Find the fraction of support elements among the s_j's
    Call COMPARE$(\{t\}, \{s_j\}, \eta = \frac{1}{2}, K = \frac{2}{\tau}, \frac{1}{4\ell})$ to get either a value ρ, High or Low.
    if COMPARE returned High or a value $\rho \ge 1/2$ **then**
      Record s_j as "support."
    end if
  end for
  if the number of s_j's marked "support" is at least $(1 - \frac{3}{4}\varepsilon)\ell$ **then return** yes
  else return no
  end if
end loop

(i) if $\mathrm{supp}(p) \ge 1 - \varepsilon/2$, then $\Pr[\,\eta < 1 - 3\varepsilon/4\,] \le 1/12$, while (ii) if $\mathrm{supp}(p) \le 1 - \varepsilon/2$, then $\Pr[\,\eta \ge 1 - 3\varepsilon/4\,] \le 1/12$. We hereafter condition on this (i.e., η being a good enough estimate). We also condition on all calls to COMPARE yielding results as per specified, which by a union bound overall happens except with probability $1/12 + 1/12 = 1/6$, and break the rest of the analysis in two cases.

(a) Since the support size ω is in this case at least $(1 - \varepsilon/2)n$, from Fact 3 we get that $p(L_{\varepsilon/2}) \ge \frac{1-\varepsilon}{2}\tau \ge \frac{\tau}{4}$. Therefore, except with probability at most $(1 - \tau/4)^k < 1/12$, at least one of the t_j's will belong to $L_{\varepsilon/2}$. When this happens, and by the choice of parameters in the calls to COMPARE, we get that $t \in L_{\varepsilon/2}$; that is $p(t) \in [\tau/n, 2/n]$. But then the calls to the routine on Step 15 will always return either a value (since t

is "comparable" to all $x \in L_{\varepsilon/2}$ – i.e., has probability within a factor $2/\tau$ of them) or High (possible for those s_j's that have weight greater than $2/n$), unless s_j has mass 0 (that is, is not in the support). Therefore, the fraction of points marked as support is exactly η, which by the foregoing discussion is at least $1 - 3\varepsilon/4$: the algorithm returns yes at Step 20.

(b) Conversely, if $\omega \leq (1-\varepsilon)n$, there will be a fraction $1 - \eta > 3\varepsilon/4$ of the s_j's having mass 0. However, no matter what t is it will still be in the support and therefore have $p(t) \geq \tau/n$: for these s_j's, the call to Compare on Step 15 can thus only return Low. This means that there can only be less than $(1 - \frac{3}{4}\varepsilon)\ell$ points marked "support" among the s_j's, and hence that the algorithm will output no as it should.

Overall, the inner loop of the algorithm thus only fails with probability at most $1/12 + 1/6 + 1/12 = 1/3$ (respectively for η failing to be a good estimate, the calls to Compare failing to yield results as guaranteed, or no t_j hitting $L_{\varepsilon/2}$ in case (a)). Repeating independently $\log(1/\delta)$ times and taking the majority vote boosts the probability of success to $1 - \delta$.

Query complexity. The sample complexity comes from the k^2 calls on Step 4 (each costing $O(\log k)$ queries) and the ℓ calls on Step 15 (each costing $O\left(\frac{1}{\tau}\log \ell\right)$ queries). By the setting of ℓ and because of the $\log(1/\delta)$ repetitions, this results in an overall query complexity $O\left(\left(\frac{1}{\tau^2}\log\frac{1}{\tau} + \frac{1}{\tau\varepsilon^2}\log\frac{1}{\varepsilon}\right)\log\frac{1}{\delta}\right)$. □

Proof of Lemma 58. As described in Algorithm 16, the subroutine is fairly simple: using its knowledge of an upper bound on the support size, it takes enough uniformly distributed samples to have (with high probability) at least one falling outside the support. Then, it uses the conditional oracle to "order" these samples according to their probability mass, and returns the lightest of them – i.e., one with zero probability mass.

Correctness. It is straightforward to see that provided at least one of the s_j's falls outside the support and that all calls to Compare behave as expected, then the procedure returns one of the "lightest" s_j's, i.e. a non-support element. By a union bound, the latter holds with probability at least $1 - \delta/2$; as for the former, since t is by assumption an upper bound

Algorithm 16 GETNONSUPPORT$_p(m, \delta)$

Require: COND access to p; upper bound t on $\mathrm{supp}(p)$, probability of failure δ
Ensure: Returns $r \in [n]$ such that, with probability at least $1 - \delta$, $r \notin \mathrm{supp}(p)$

Set $k \triangleq \lceil \log \frac{2}{\delta} \log^{-1} \frac{n}{t} \rceil$.
Draw independently k points $s_1, \ldots, s_k \sim \mathcal{U}_n$
for all $1 \leq i < j \leq k$ **do**
    Call COMPARE$(\{s_i\}, \{s_j\}, \eta = \frac{1}{2}, K = 2, \frac{\delta}{2k^2})$ to get a 2-approx. ρ of $\frac{p(s_j)}{p(s_i)}$, High or Low.
    if COMPARE returned High or a value ρ **then**
        Record $s_i \preceq s_j$
    else
        Record $s_j \prec s_j$
    end if
end for
return $\arg\min_{\preceq}\{s_1, \ldots, s_k\}$ ▷ Return (any) minimal element for $\preceq$.

on the support size it holds with probability at least $1 - (t/n)^k \geq 1 - \delta/2$ (from our setting of k). Overall, the procedure's output is correct with probability at least $1 - \delta$, as claimed.

Query complexity. The query complexity of GETNONSUPPORT is due to the $\binom{k}{2}$ calls to COMPARE, and is therefore $O\left(k^2 \log \frac{k}{\delta}\right)$ because of our setting for η and K (which is in turn $\tilde{O}\left(\log^2 \frac{1}{\delta} \log^{-2} \frac{n}{t}\right)$). (In our case, we shall eventually take $t = (1 - \varepsilon/2)n$ and $\delta = 1/10$, thus getting $k = O(1/\varepsilon)$ and a query complexity of $\tilde{O}(1/\varepsilon^2)$.) □

Proof of Lemma 59. Our final subroutine, described in Algorithm 17, essentially derives from the following observation: a random set S of size (approximately) σ obtained by including independently each element of the domain with probability $1/\sigma$ will intersect the support on ω/σ points on expectation. What we *can* test given our reference point $r \notin \mathrm{supp}(p)$, however, is only whether $S \cap \mathrm{supp}(p) = \emptyset$. But this is enough, as by repeating sufficiently many times (taking a random S and testing whether it intersects the support at all) we can distinguish between the two cases we are interested in. Indeed, the expected fraction of times S includes a support element in either cases is known to the algorithm and differs by roughly $\Omega(\varepsilon)$, so $O(1/\varepsilon^2)$ repetitions are enough to tell the two cases apart.

Algorithm 17 IsAtMostSupportSize$_p(\sigma, r, \varepsilon, \delta)$

Require: COND access to p; size $\sigma \geq 2$, non-support element r, accuracy ε, probability of failure δ

Ensure: Returns, with probability at least $1 - \delta$, yes if $\sigma \leq |\operatorname{supp}(p)|$ and no if $\sigma > (1+\varepsilon)|\operatorname{supp}(p)|$.

1: Set $\alpha \leftarrow \left(1 - \frac{1}{\sigma}\right)^{\sigma} \in [\frac{1}{4}, e^{-1}]$, $\tau \leftarrow \alpha(\alpha^{-\frac{\varepsilon}{2}} - 1) = \Theta(\varepsilon)$.

2: Repeat the following $O(\log(1/\delta))$ times and output the majority vote.

3: **loop**

4: **for** $k = O\left(\frac{1}{\tau^2}\right)$ times **do**

5: Draw a subset $S \subseteq [n]$ by including independently each $x \in [n]$ with probability $1/\sigma$.

6: Draw $x \sim p_S$.

7: Call Compare$(\{x\}, \{r\}, \eta = \frac{1}{2}, K = 1, \frac{1}{100k})$ ▷ Low if $S \cap \operatorname{supp}(p) \neq \emptyset$; $\rho \in [\frac{1}{2}, 2)$ o.w.

8: Record yes if Compare returned Low, no otherwise.

9: **end for**

10: **return** yes if at least $k\left(\alpha + \frac{\tau}{2}\right)$ "yes"'s were recorded, no otherwise. ▷ Thresholding.

11: **end loop**

Correctness. We condition on all calls to Compare being correct: by a union bound, this overall happens with probability at least 99/100. We shall consider the two cases $\sigma \leq \omega$ and $\sigma > (1+\varepsilon)\omega$, and focus on the difference of probability p of recording yes on Step 8 between the two, in any fixed of the k iterations. In both cases, note p is exactly $(1 - 1/\sigma)^{\omega}$.

- If $\sigma \leq \omega$, then we have $p \leq \left(1 - \frac{1}{\sigma}\right)^{\sigma} = \alpha$.
- If $\sigma > (1+\varepsilon)\omega$, then $p > \left(1 - \frac{1}{\sigma}\right)^{\sigma/(1+\varepsilon)} > \left(1 - \frac{1}{\sigma}\right)^{\sigma(1-\varepsilon/2)} = \alpha^{1-\varepsilon/2}$.

As $\alpha \in [\frac{1}{4}, e^{-1}]$, the difference between the two is $\tau = \alpha(\alpha^{-\varepsilon/2} - 1) = \Theta(\varepsilon)$. Thus, repeating the atomic test of Step 4 $O(1/\tau^2)$ times before thresholding at Step 10 yields the right answer with constant probability, then brought to $1 - \delta$ by the outer repeating and majority vote.

Query complexity. Each call to COMPARE at Step 7 costs $O(\log k)$ queries, and is in total repeated $O(k \log(1/\delta))$ times. By the setting of k and τ, the overall query complexity is therefore $O\left(\frac{1}{\varepsilon^2} \log \frac{1}{\varepsilon} \log \frac{1}{\delta}\right)$. □

6.5 Non-Adaptive Upper Bounds

In this section, we prove upper bounds for several distribution testing and property estimation problems in the non-adaptive model. At their core, all of our algorithms depend on our algorithm ANACONDA, which is presented in Section 6.5.2 as Algorithm 18. We then describe results for uniformity testing (Section 6.5.3), equivalence testing (Section 6.5.4), identity testing (Section 6.5.5), and support-size estimation (Section 6.5.6).

6.5.1 Additional Preliminaries

We will frequently use $z = (p - q)/\varepsilon$ to denote the "noise vector" between p and q, and $\bar{p} = (p + q)/2$. While the two cases in distribution testing that one considers are usually $p = q$ and $d_{\mathrm{TV}}(p, q) \geq \varepsilon$, for convenience of notation, we will generally assume the latter case to be $d_{\mathrm{TV}}(p, q) = \varepsilon$ – it is not hard to see that our analysis carries through whenever the algorithm is given a parameter ε which is less than the true total variation distance between p and q. With this in mind, when $p = q$, we have that $z = \vec{0}$, and when $d_{\mathrm{TV}}(p, q) = \varepsilon$, we have that $\|z\|_1 = 2$ and $\sum_{i \in [n]} z(i) = 0$. Let z^+ denote the "rectified" version of z, where $z^+(i) = \max(0, z(i))$ – here, in the latter case, $\|z^+\|_1 = \sum_{i \in [n]} z^+(i) = 1$. $z^-(i) = \max(0, -z(i))$ is defined similarly.

For our analysis, we will group indices into bins:

Definition 25. *The j-th bin for a vector x, denoted by $Bin_j(x)$, contains all indices whose values are in the range $[2^{-j}, 2^{-j+1})$, i.e. $Bin_j(x) \triangleq \{i \ : \ \frac{1}{2^j} \leq x(i) < \frac{1}{2^{j-1}}\}$.*

6.5.2 ANACONDA: A Non-Adaptive Algorithm for Distribution Testing

Our algorithm, ANACONDA, is presented in Algorithm 18. While it is phrased in terms of equivalence testing, it still works when a distribution q is explicitly given (i.e., identity testing), as one can simply simulate NACOND queries to q. It takes three parameters, T, τ, and ε', which we will instantiate differently (as required by our analysis) for uniformity and equivalence testing.

The algorithm's behavior can roughly be summarized as follows. The algorithm first chooses a random size for a query set. It then chooses a random subset of the domain of this size. Next, it draws several conditional samples from this set, from both p and q. Finally, if it detects that a single element from the query set has a significantly discrepant probability mass under p and q, it outputs that the two distributions are far. It repeats this process several times, eventually outputting that the distributions are equal if it never discovers a discrepant element.

Algorithm 18 ANACONDA: An algorithm for testing equivalence given NACOND oracle access to p, q

function ANACONDA(ε, NACOND_p oracle, NACOND_q oracle, parameters T, τ, ε')
  for $t = 1$ to T **do**
    Choose an integer $j \in \{1, \ldots, 2 \log n\}$ uniformly at random, and define $r \triangleq 2^j$.
    Choose a random set $S \subseteq [n]$, independently selecting each i to be in S with probability $1/r$.
    Perform τ queries to NACOND_p and NACOND_q on the set S.
    Using these queries, form the empirical distribution $\hat{p}_S$ and $\hat{q}_S$.
    if $\exists i \in S$ such that $|\hat{p}_S(i) - \hat{q}_S(i)| \geq \varepsilon'$ **then**
      return $d_{\mathrm{TV}}(p, q) \geq \varepsilon$
    end if
  end for
  return $p = q$
end function

6.5.3 Analysis of ANACONDA for Uniformity Testing

In this section, we will prove Theorem 53, by instantiating ANACONDA with parameters $T = \Theta(\log n)$, $\tau = \Theta(\log \log n / \varepsilon^2)$, and $\varepsilon' = \Theta(\varepsilon)$.

Theorem 53 (Non-Adaptive Uniformity Testing)**.** *There exists an algorithm which, given* NACOND *access to an unknown distribution p on $[n]$, makes $\tilde{O}\left(\frac{\log n}{\varepsilon^2}\right)$ queries to the oracle on p and distinguishes between the cases $p = \mathcal{U}_n$ and $d_{\mathrm{TV}}(p, \mathcal{U}_n) \geq \varepsilon$ with probability at least $2/3$, where $\mathcal{U}_n$ is the uniform distribution on $[n]$.*

Our strategy will be as follows. We will argue that, with probability $\Omega(1/\log n)$, ANACONDA will select a set S with a single element that has significantly different mass under the uniform distribution and the distribution p_S. In this way, we will reduce the problem from ℓ_1-testing to ℓ_∞-testing, the latter of which is solvable with very few samples, by Lemma 1.

More precisely, we compare the probability assigned to a particular symbol i when performing a conditional sample on S, in the two cases where $p = \mathcal{U}_n$, and when $d_{\mathrm{TV}}(p, \mathcal{U}_n) = \varepsilon$. In the former case, the probability is $\frac{\mathcal{U}_n(i)}{\mathcal{U}_n(S)}$, while in the latter, it is $\frac{\mathcal{U}_n(i)+\varepsilon z(i)}{\mathcal{U}_n(S)+\varepsilon z(S)}$. Therefore, the difference in probability assigned is

$$\left| \frac{\mathcal{U}_n(i) + \varepsilon z(i)}{\mathcal{U}_n(S) + \varepsilon z(S)} - \frac{\mathcal{U}_n(i)}{\mathcal{U}_n(S)} \right|. \tag{6.5}$$

In the following two subsections, we will show that the following lemma:

Lemma 60. *If $d_{\mathrm{TV}}(p, \mathcal{U}_n) = \varepsilon$, then for each t,* ANACONDA *will select a set S which causes (6.5) to be $\geq \Omega(\varepsilon)$ with probability $\geq \Omega(1/\log n)$.*

Assuming this to be true for the moment, we will show how to complete the proof. Repeating this process $T = \Theta(\log n)$ times will guarantee that at least one iteration will choose an S containing a sufficiently discrepant element with probability $\geq 9/10$. We focus on the iteration where such an S is selected.

Now if we draw $\Theta(\log \log n/\varepsilon^2)$ samples from p_S, Lemma 1 implies the empirical distribution $\hat{p}_S$ will approximate p_S in Kolmogorov distance up to an additive ε', with probability at least $1 - O\left(\frac{1}{\log n}\right)$, and thus Line 7 will correctly identify that $d_{\mathrm{TV}}(p, \mathcal{U}_n) = \varepsilon$. Therefore, with probability at least $4/5$, the algorithm will correctly detect in this case that $d_{\mathrm{TV}}(p, \mathcal{U}_n) = \varepsilon$.

We now examine what happens when $p = \mathcal{U}_n$. For each iteration t, the uniform distribution on S and p_S will be equal. We again invoke Lemma 1 with $\Theta(\log \log n/\varepsilon^2)$ samples, and use a union bound over all $T = \Theta(\log n)$ iterations. This implies that, with probability

at least 9/10, Line 7 will never identify an element which has $\geq \varepsilon'$ discrepancy, and thus the algorithm will output that $p = \mathcal{U}_n$ in Line 11.

It remains to prove Lemma 60. We break the analysis into two cases, which we address in the following two subsections. In Section 6.5.3.1, we handle the case where, for all $x \in \{z^-, z^+\}$, $\sum_{j=\log n/32+1}^{\log 5n} \sum_{i \in \mathrm{Bin}_j(x)} x(i) \geq 1/5$. This corresponds to the case where there are many symbols with small discrepancy from the uniform distribution, in both the positive and negative direction. In Section 6.5.3.2, we handle the complement of this case, where there exists an $x \in \{z^-, z^+\}$ for which $\sum_{j=1}^{\log n/32} \sum_{i \in \mathrm{Bin}_j(x)} x(i) \geq 3/5$. Roughly, this happens when there are not too many symbols which capture the discrepancy between the distributions.

6.5.3.1 Case I: Many Small Discrepancies

In this section, we prove Lemma 60 in the case where for all $x \in \{z^-, z^+\}$, $\sum_{j=\log n/32+1}^{\log 5n} \sum_{i \in \mathrm{Bin}_j(x)} x(i) \geq 1/5$. In short, the analysis can be summarized as follows: if the algorithm chooses a set S of size 2, it is likely to contain two elements with non-trivial discrepancy, and in both the positive and negative direction – this will suffice to make (6.5) be $\geq \Omega(\varepsilon)$.

We have the following proposition relating the size of a bin to the mass it contains, which is immediate from Definition 25.

Proposition 22. $2^{j-1} \sum_{i \in Bin_j} x(i) \leq |Bin_j(x)| \leq 2^j \sum_{i \in Bin_j} x(i)$.

This gives us the following lower bound on the number of symbols which are in bins $\log n/32 + 1$ through $\log 5n$:

$$\sum_{j=\log n/32+1}^{\log 5n} |\mathrm{Bin}_j(x)| \geq \sum_{j=\log n/32+1}^{\log 5n} 2^{j-1} \sum_{i \in \mathrm{Bin}_j(x)} x(i) \geq \frac{n}{32} \sum_{j=\log n/32+1}^{\log 5n} \sum_{i \in \mathrm{Bin}_j(x)} x(i) \geq \frac{n}{160} \quad (6.6)$$

In other words, for either $x \in \{z^-, z^+\}$, there are $\Omega(n)$ symbols with $x(i) \geq 1/5n$.

We complete the proof of Lemma 60 as follows. With probability $\frac{1}{2 \log n}$, Anaconda will select $r = \log n$ in Line 3. Conditioning on this, with constant probability, the set S selected in Line 4 will be of size exactly 2. Further conditioning on this, due to (6.6), with constant probability S will consist of two symbols $i_1 \in \mathrm{Bin}_{j'}(z^+)$ and $i_2 \in \mathrm{Bin}_{j''}(z^-)$ for $\log n/32 + 1 \leq j', j'' \leq \log 5n$.

Without loss of generality, suppose that $z(i_1) \geq 0$ and $z(i_2) \leq 0$. Then (6.5) is the following:

$$\left| \frac{\mathcal{U}_n(i) + \varepsilon z(i)}{\mathcal{U}_n(S) + \varepsilon z(S)} - \frac{\mathcal{U}_n(i)}{\mathcal{U}_n(S)} \right| = \frac{\varepsilon n(z(i_i) - z(i_2))}{2(2 + \varepsilon n(z(i_1) + z(i_2)))} \geq \frac{\varepsilon n \cdot \frac{2}{5n}}{2(2 + \varepsilon n \cdot \frac{32}{n})} \geq \frac{\varepsilon}{68}. \tag{6.7}$$

This expression is $\geq \Omega(\varepsilon)$, and this event happens with probability $\geq \Omega(1/\log n)$, thus proving Lemma 60 in this case.

6.5.3.2 Case II: Not So Many Small Discrepancies

In this section, we prove Lemma 60 in the case where there exists an $x \in \{z^-, z^+\}$ for which $\sum_{j=1}^{\log n/32} \sum_{i \in \mathrm{Bin}_j(x)} x(i) \geq 3/5$. Without loss of generality, assume that this holds for z^+. Furthermore, we focus our analysis on the case where ANACONDA picks an $r \leq \log n/32$. For the remainder of this proof, condition on this event, which happens with probability at least $1/4$.

We will need the following key lemma:

Lemma 61. *Suppose $d_{\mathrm{TV}}(p, \mathcal{U}_n) = \varepsilon$. For each iteration t, with probability $\geq \frac{3}{20 \log n/32}$, the algorithm will choose an r and a set S such that there exists $i \in S$ with $z^+(i) \geq 1/r$.*

Proof. For some fixed j, the probability of choosing j is $\frac{1}{\log n/32}$, and, conditioning on this j, the probability of picking any element from $\mathrm{Bin}_j(z^+)$ to be in S is $1 - \left(1 - \frac{1}{2^j}\right)^{|\mathrm{Bin}_j(z^+)|}$. By the law of total probability, we sum this over all bins to get the probability that the event of interest happens:

$$\begin{aligned}
\frac{1}{\log n/32} \sum_{j \in [\log n/32]} 1 - \left(1 - \frac{1}{2^j}\right)^{|\mathrm{Bin}_j(z^+)|} &\geq \frac{1}{\log n/32} \sum_{j \in [\log n/32]} 1 - \exp\left(-\frac{|\mathrm{Bin}_j(z^+)|}{2^j}\right) && (6.8) \\
&\geq \frac{1}{\log n/32} \sum_{j \in [\log n/32]} 1 - \exp\left(-\frac{1}{2} \sum_{i \in \mathrm{Bin}_j(z^+)} z^+(i)\right) && (6.9) \\
&\geq \frac{1}{\log n/32} \sum_{j \in [\log n/32]} \frac{1}{4} \sum_{i \in \mathrm{Bin}_j(z^+)} z^+(i) && (6.10) \\
&\geq \frac{3}{20 \log n/32}. && (6.11)
\end{aligned}$$

(6.8) follows from the inequality $1 - x \leq \exp(-x)$, (6.9) is due to Proposition 22, (6.10) is by the inequality $1 - \exp(-x) \geq x/2$ (which holds for all $x \in [0, 1]$), and (6.11) is by assumption $\square$

We will require the following lemmata to complete the proof:

Lemma 62. *For any i and j,*

$$\Pr\left[\frac{1}{2 \cdot 2^j} \leq \mathcal{U}_n\left(S \setminus i\right) \leq \frac{3}{2 \cdot 2^j}\right] \geq 1 - 2/e^2.$$

Proof. Observe that the size of $S \setminus i$ is a sum of $n-1$ i.i.d. Bernoulli random variables with parameter $1/2^j$, and thus has expectation $\mu = \frac{n-1}{2^j}$. Then, by Chernoff bound, we have

$$\Pr\left[\frac{9}{16}\frac{(n-1)}{2^j} \leq |S \setminus i| \leq \frac{23}{16}\frac{(n-1)}{2^j}\right] \geq 1 - 2\exp\left(-\frac{49\mu}{768}\right) \geq 1 - 2/e^2.$$

The last inequality follows since $j \leq \log n/32$ for n larger than some absolute constant. Similarly, the lemma follows for n larger than some absolute constant by rescaling the size of the set by a factor of n. $\square$

Lemma 63. *If $d_{\mathrm{TV}}(p, \mathcal{U}_n) = \varepsilon$, then for any i and j,*

$$\Pr\left[z\left(S \setminus i\right) \geq \frac{4}{2^j}\right] \leq 1/4.$$

Proof. Note that $z^+(S \setminus i)$ is a non-negative random variable. Its expectation $\mathbf{E}[z^+(S \setminus i)] \leq \mathbf{E}[z^+(S)] \leq 1/2^j$. The lemma follows by Markov's inequality, and by observing that the addition of any negative elements of z will only decrease $z(S \setminus i)$. $\square$

Note that, by Lemmas 61, 62, 63, if $d_{\mathrm{TV}}(p, \mathcal{U}_n) = \varepsilon$, with probability at least $\frac{1}{4} \cdot \frac{1}{\log n/32} \cdot \left(1 - \frac{1}{4} - 2/e^2\right) \geq \Omega(1/\log n)$, the following events happen simultaneously:

- $r \leq n/32$;
- $z(i) \geq 1/r$;
- $\mathcal{U}_n(i) = 1/n$;

- $z(S \setminus i) \leq 4/r$;
- $\frac{1}{2r} \leq \mathcal{U}_n(S \setminus i) \leq \frac{3}{2r}$;

We now show that a set S with all these properties will result in (6.5) being $\geq \Omega(\varepsilon)$:

$$\begin{aligned}
\left|\frac{\mathcal{U}_n(i)+\varepsilon z(i)}{\mathcal{U}_n(S)+\varepsilon z(S)} - \frac{\mathcal{U}_n(i)}{\mathcal{U}_n(S)}\right| &= \varepsilon \left|\frac{z(i)\mathcal{U}_n(S\setminus i) - z(S\setminus i)\mathcal{U}_n(i)}{\mathcal{U}_n(S)(\mathcal{U}_n(S)+\varepsilon z(S))}\right| \\
&\geq \varepsilon \cdot \frac{1}{\mathcal{U}_n(S)(\mathcal{U}_n(S)+\varepsilon z(S))}\left(\frac{z(i)}{2r} - \frac{4}{rn}\right) \\
&\geq \varepsilon \cdot \frac{r}{2}\frac{1}{\frac{2}{r}+\varepsilon\left(\frac{4}{r}+z(i)\right)}\left(\frac{z(i)}{2r} - \frac{4}{rn}\right) \\
&\geq \varepsilon \cdot \frac{1}{\frac{2}{r}+\varepsilon\left(\frac{4}{r}+z(i)\right)}\left(\frac{z(i)}{4} - \frac{2}{n}\right)
\end{aligned}$$

The analysis concludes by considering two cases. If $\varepsilon z(i) \geq \frac{2}{r} + \varepsilon \cdot \frac{4}{r}$, then we have the lower bound $\varepsilon \cdot \frac{1}{2\varepsilon z(i)}\left(\frac{z(i)}{4} - \frac{2}{n}\right) = \Omega(1) \geq \Omega(\varepsilon)$, as desired. Otherwise, we have the lower bound $\varepsilon \cdot \frac{r}{12}\left(\frac{z(i)}{4} - \frac{2}{n}\right) \geq \varepsilon \cdot \frac{r}{12}\left(\frac{1}{4r} - \frac{2}{n}\right) \geq \frac{\varepsilon}{96}$, which completes the proof.

6.5.4 Analysis of Anaconda for Equivalence Testing

In this section, we will prove Theorem 52 by instantiating Anaconda with parameters $T = \Theta(\log^6 n)$, $\tau = \tilde{\Theta}(\log^6 n/\varepsilon^2)$, and $\varepsilon' = \frac{\varepsilon}{\tilde{\Theta}(\log^3 n)}$.

Theorem 52 (Non-Adaptive Equivalence Testing)**.** *There exists an algorithm which, given* NACOND *access to unknown distributions p, q on $[n]$, makes $\tilde{O}\left(\frac{\log^{12} n}{\varepsilon^2}\right)$ queries to the oracle on each distribution and distinguishes between the cases $p = q$ and $d_{\mathrm{TV}}(p,q) \geq \varepsilon$ with probability at least $2/3$.*

We will require the following proposition, says if $d_{\mathrm{TV}}(p,q) = \varepsilon$ and Anaconda selects an appropriate set S, then it will detect the discrepancy.

Proposition 23. *Suppose that $d_{\mathrm{TV}}(p,q) = \varepsilon$ and that within the first T iterations a set S is identified such that for some $i \in S$ and some $c > 0$,*

$$\min\{z(i), z(i) - z(S)\} \geq \frac{p(S)+q(S)}{\tilde{O}(\log^c n)}.$$

Then, for $\varepsilon' = \frac{\varepsilon}{\tilde{O}(\log^c n)}$ *and* $\tau = \tilde{\Omega}\left(\frac{\log^{2c} n}{\varepsilon^2}\right)$*, the algorithm outputs that* $d_{\mathrm{TV}}(p,q) \geq \varepsilon$ *with probability at least* $1 - \frac{1}{\operatorname{poly}\log n}$*.*

Proof. We first argue that $|p_S(i) - q_S(i)| \geq \varepsilon \frac{\min\{z(i), z(i) - z(S)\}}{p(S)+q(S)}$.

We set $\bar{p} = \frac{p+q}{2}$. We have that $p = \bar{p} + z\frac{\varepsilon}{2}$, $q = \bar{p} - z\frac{\varepsilon}{2}$ and

$$\left|\frac{p(i)}{p(S)} - \frac{q(i)}{q(S)}\right| = \left|\frac{\bar{p}(i) + z(i)\frac{\varepsilon}{2}}{\bar{p}(S) + z(S)\frac{\varepsilon}{2}} - \frac{\bar{p}(i) - z(i)\frac{\varepsilon}{2}}{\bar{p}(S) - z(S)\frac{\varepsilon}{2}}\right| = \frac{\varepsilon}{2}\left|\frac{z(i)\bar{p}(S) - \bar{p}(i)z(S)}{\bar{p}^2(S) - (z(S)\frac{\varepsilon}{2})^2}\right| \geq \frac{\varepsilon}{2}\left|\frac{z(i)\bar{p}(S) - \bar{p}(i)z(S)}{\bar{p}^2(S)}\right|.$$

As $z(i)\bar{p}(S) - \bar{p}(i)z(S) \geq \bar{p}(S)\min\{z(i), z(i) - z(S)\}$, it follows that

$$|p_S(i) - q_S(i)| \geq \frac{\varepsilon}{2}\frac{\min\{z(i), z(i) - z(S)\}}{(p(S) + q(S))/2}$$

To complete the proof, we note that the condition implies that $|p_S(i) - q_S(i)| \geq \frac{\varepsilon}{\tilde{O}(\log^c n)}$ and thus by Lemma 1, $\tau = \tilde{\Omega}\left(\frac{\log^{2c} n}{\varepsilon^2}\right)$ suffices to detect (with failure probability $1/\operatorname{poly}\log n$) that $\|p_S - q_S\|_\infty > \varepsilon' = \frac{\varepsilon}{\tilde{O}(\log^c n)}$. □

To complete the proof, we will show that after $T = \operatorname{poly}\log n$ iterations, Algorithm 18 will choose a set S that satisfies the conditions of Proposition 23.

We define $\hat{z}$ to be the vector with $\hat{z}(i) = z(i)$ if $|z(i)| > \frac{p(i)+q(i)}{400\log n}$ and $\hat{z}(i) = 0$ otherwise. Roughly, this "zeroes out" the noise for any i where the noise vector z is too large in comparison to the signal vector $p + q$. Let b^+ be the measure on $\{1, \ldots, 2\log n\}$ with mass $\hat{z}^+(\mathrm{Bin}_j(z^+))$ and equivalently define b^-. Notice that $|b^+|, |b^-| \in [1 - \frac{1}{200\log n}, 1]$. This is because $\sum_{i:\hat{z}^+(i)=0} z^+(i) \leq \sum_i \frac{p(i)+q(i)}{400\log n} \leq \frac{1}{200\log n}$.

The next lemma shows that, if there are two bins (with respect to the positive and negative z vectors) which are both "heavy" and are close in index, then we will obtain an appropriate set S (for Proposition 23).

Lemma 64. *If* $b^+(j) > \frac{1}{\tilde{O}(\log^\alpha n)}$ *and* $b^-(j') > \frac{1}{\tilde{O}(\log^\beta n)}$*, for some* j *and* j' *with* $2^{|j-j'|} = \tilde{O}(\log^\gamma n)$*, then a single iteration of Algorithm 18 finds set* S *and* $i \in S$ *with* $\min\{z(i), z(i) - z(S)\} \geq \frac{p(S)+q(S)}{\tilde{O}(\log^{\gamma+1} n)}$ *with probability* $\frac{1}{\tilde{O}(\log^{\alpha+\beta+\gamma+1} n)}$*.*

Proof. With probability $\frac{1}{\tilde{O}(\log n)}$, an iteration of Algorithm 18 will choose $r = 2^{-\max\{j,j'\}-3}$. Given this value of r, a unique i with $\hat{z}^+(i) \in [2^{-j}, 2^{-j+1})$ and a unique i' with $\hat{z}^-(i') \in$

$[2^{-j'}, 2^{-j'+1})$ are selected with probability $\frac{1}{\tilde{O}(\log^{\alpha+\beta+\gamma} n)}$. It holds that $z^-(i'), z^+(i) \in [8, \tilde{O}(\log^{\gamma} n)] \cdot r$ and their corresponding $p(i) + q(i) \leq O(\log n) \cdot z(i) \leq \tilde{O}(\log^{1+\gamma} n) r$ and $p(i') + q(i') \leq \tilde{O}(\log^{1+\gamma} n) r$.

By Markov's inequality, with probability at least $3/4$, $z(S \setminus \{i, i'\}) \leq z^+(S \setminus \{i, i'\}) \leq 4r$. Similarly, with probability at least $3/4$, $p(S \setminus \{i, i'\}) + q(S \setminus \{i, i'\}) \leq 8r$. By a union bound with probability $1/2$ both hold simultaneously.

When all of these events occur, which happens with probability at least $\frac{1}{\tilde{O}(\log^{\alpha+\beta+\gamma+1} n)}$ we get that:

$$\min\{z(i), z(i) - z(S)\} \geq 4r \quad \text{since} \quad z(i) - z(S) \geq z^-(i') - z(S \setminus \{i, i'\}) \geq 4r$$

The lemma follows by noting that $p(S) + q(S) \leq \tilde{O}(\log^{\gamma+1} n) r$. □

Finally, we have our main lemma required for the analysis. It leverages Lemma 64 to show that we can obtain an appropriate set S with reasonable probability.

Lemma 65. *If $d_{\mathrm{TV}}(p, q) = \varepsilon$, then a single iteration of Algorithm 18 finds set S and $i \in S$ with $\min\{z(i), z(i) - z(S)\} \geq \frac{p(S)+q(S)}{\tilde{O}(\log^3 n)}$ with probability $\frac{1}{\tilde{O}(\log^6 n)}$.*

Proof. Before we begin, we require the following two simple concentration lemmas:

Lemma 66. *Let $0 < a < b$, $X_i \sim$ Bernoulli(2^{-a}) and let $1 > \sum_{i:x_i<2^{-b}} x_i \geq c$. Then, $\sum_{i:pxi<2^{-b}} X_i x_i > 2^{-a}(c - t2^{-(b-a)/2})$, with probability $1 - e^{-t}$.*

Proof. We apply the Chernoff bound on the variables $Z_i = X_i 2^b x_i$. We get that with probability $1 - e^{-t}$, $2^b \sum_{i:x_i<2^{-b}} X_i x_i > 2^{b-a} c - t2^{(b-a)/2}$. Thus, $2^a \sum_{i:x_i<2^{-b}} X_i x_i > c - t2^{-(b-a)/2}$ □

Lemma 67. *Let $a \geq 1$, $X_i \sim$ Bernoulli(2^{-a}) and let $1 > \sum_{i:x_i>2^{-a}} x_i$. Then, $\sum_{i:x_i>2^{-a}} X_i x_i = 0$, with probability $\frac{1}{4}$.*

Proof. There are at most 2^a elements x_i and every element is selected independently with probability 2^{-a}. The probability that no element is chosen is $(1 - 2^{-a})^{2^a} \geq \frac{1}{4}$. □

We continue with the main proof. Consider two cases:

1. $d_{\mathrm{K}}(b^+, b^-) \leq \frac{1}{8 \log n}$.

   In this case, as $\sum b^+(j) > 2/3$, there will be a bin j with $b^+(j) \geq \frac{2/3}{2 \log n}$. As the $d_{\mathrm{K}}(b^+, b^-) \leq \frac{1}{8 \log n}$, the corresponding $b^-(j) \geq \frac{1}{3 \log n} - \frac{2}{8 \log n} \geq \frac{1}{12 \log n}$. Then, Lemma 64 implies that a good set will be identified with high probability.

2. $d_{\mathrm{K}}(b^+, b^-) > \frac{1}{8 \log n}$.

   In this case, there will be a bin j_r with $|\sum_{j \geq j_r} b^-(j) - \sum_{j \geq j_r} b^+(j)| \geq \frac{1}{8 \log n}$. Without loss of generality, $\sum_{j \geq j_r} b^+(j) < \sum_{j \geq j_r} b^-(j)$.

   Let j_l be the largest index such that $\frac{1}{8 \log n} < \sum_{j=j_l}^{j_r} b^+(j)$. Then there must exist a $j^* \in [j_l, j_r]$ such that $b^+(j^*) > \frac{1}{16 \log^2 n}$ as $|[j_l, j_r]| \leq 2 \log n$.

   If there is a $j \in [j^*, j^* + 2 \log \log n]$, with $b^-(j) > \frac{1}{100 \log n \log \log n}$, Lemma 64 implies that with probability $\frac{1}{O(\log^6 n)}$, $\min\{z(i), z(i) - z(S)\} \geq \frac{p(S)+q(S)}{\tilde{O}(\log^3 n)}$.

   Otherwise, we have that $\sum_{j \geq j^* + 2 \log \log n} b^-(j) > \frac{1}{20 \log n} + \sum_{j \geq j^*} b^+(j)$. We will show that in this case, when the algorithm selects $r = 2^{-j^*}$, a good set is identified with non-trivial probability.

   With probability $\Omega(b^+(j^*)) = \frac{1}{O(\log^2 n)}$, a unique i with $\hat{z}^+(i) \in [2^{-j^*}, 2^{-j^*+1})$ is selected. It holds that $z^+(i) \in [1, 2] \cdot r$ and the corresponding $p(i) + q(i) \leq O(\log n) \cdot z(i) \leq \tilde{O}(\log n) r$.

   To complete the proof, we now provide bounds for $z(S \setminus \{i\})$ and $p(S \setminus \{i\}) + q(S \setminus \{i\})$.

   We decompose $z(S \setminus \{i\})$ into contributions from different sets of elements:

   (a) $\hat{z}^+((S \setminus \{i\}) \cap (\bigcup_{j \geq j^*} \mathrm{Bin}_j(z^+))\}) \leq r \sum_{j \geq j^*} b^+(j) + \frac{r}{200 \log n}$ with probability at least $\frac{1}{100 \log n}$. This holds by Markov's inequality.

   (b) $\hat{z}^+((S \setminus \{i\}) \cap (\bigcup_{j < j^*} \mathrm{Bin}_j(z^+))\}) = 0$ with probability 1/4. This holds by Lemma 67.

   (c) $z^+(S \setminus \{i\}) - \hat{z}^+(S \setminus \{i\}) \leq 3 \frac{r}{200 \log n}$ with probability 2/3. This holds by Markov's inequality.

   (d) $z^-(S \setminus \{i\}) \geq r \sum_{j \geq j^* + 2 \log \log n} b^-(j) - \frac{r}{200 \log n}$ with probability 15/16. This holds by a concentration bound presented in Lemma 66.

Applying a union bound on cases (b)-(d), we get that they hold simultaneously with probability 1/8. Noting that

Thus, overall $-z(S \setminus \{i\}) \geq r \sum_{j \geq j^* + 2\log\log n} b^-(j) - r \sum_{j \geq j^*} b^+(j) - \frac{5r}{200 \log n} \geq \frac{r}{20 \log n}$.

In addition, $z(i) \geq r$ and thus $\min\{z(i), z(i) - z(S)\} \geq \frac{r}{20 \log n}$. With constant probability, we also have that $p(S) + q(S) \leq O(\log n) \cdot r$.

Thus, with probability $\frac{1}{O(\log^4 n)}$, $\min\{z(i), z(i) - z(S)\} \geq \frac{p(S)+q(S)}{\tilde{O}(\log^2 n)}$.

□

Finally, with Lemma 65 in hand, we combine it with Proposition 23 to complete the proof of Theorem 52.

Proof of Theorem 52: Set $T = \Theta(\log^6(n))$. Then Lemma 65 implies that, with constant probability, after T iterations, a set S will be identified such that for some $i \in S$,

$$\min\{z(i), z(i) - z(S)\} \geq \frac{p(S) + q(S)}{\tilde{O}(\log^3 n)}.$$

Proposition 23 then implies that for $\varepsilon' = \frac{\varepsilon}{\tilde{O}(\log^3 n)}$ and $\tau = \tilde{\Omega}\left(\frac{\log^6 n}{\varepsilon^2}\right)$, the algorithm correctly outputs that $d_{\mathrm{TV}}(p, q) \geq \varepsilon$ with probability at least $1 - \frac{1}{\mathrm{poly}\log n}$.

In contrast, when $d_{\mathrm{TV}}(p, q) = 0$, the algorithm incorrectly correctly outputs that $d_{\mathrm{TV}}(p, q) \geq \varepsilon$ with probability at most $\frac{1}{\mathrm{poly}\log n}$. □

6.5.5 Analysis of ANACONDA for Identity Testing

In this section, we discuss how our results for uniformity testing imply Theorem 54 for identity testing.

Theorem 54 (Non-Adaptive Identity Testing)**.** *There exists an algorithm which, given* NACOND *access to an unknown distribution p on $[n]$ and a description of a distribution q over $[n]$, makes $\tilde{O}\left(\frac{\log^2 n}{\varepsilon^2}\right)$ queries to the oracle on p and distinguishes between the cases $p = q$ and $d_{\mathrm{TV}}(p, q) \geq \varepsilon$ with probability at least $2/3$.*

We adapt the reduction of [CFGM16], from non-adaptive identity testing to non-adaptive near-uniform identity testing. In particular, we use their Algorithm 4.2.2, with a few crucial

differences – to describe these differences, we assume familiarity with the terminology of their paper.

In Line 1, they partition the domain using $Bucket(q, [n], \frac{\varepsilon}{30})$. We perform a less fine-grained partitioning, using $Bucket(q, [n], \frac{1}{100})$. Their bucketing defines M_0 as all i such that $q(i) < \frac{1}{n}$. We define it as all i such that $q(i) < \frac{\varepsilon}{100n}$.[10] The first modification will require a stronger near-uniform identity tester than the one in their paper, which can handle identity testing to any distribution q such that $\|q - \mathcal{U}_n\|_\infty \leq \frac{1}{100n}$. The second change implies that we do not have to do a near-uniform identity test on M_0 – either $\|z(M_0)\|_1 > \varepsilon/50$ and the discrepancy will be discovered in Line 3, or $\|z(M_0)\|_1 \leq \varepsilon/50$, and this bucket can be ignored, as $\|z([n] \setminus M_0)\|_1 \geq 49\varepsilon/50$. As a result of these changes, there are only $\Theta(\log(n/\varepsilon))$ buckets in the partition, and we perform the tests in Line 2 with error bound $\frac{\delta \log(1+1/100)}{2\log(100n/\varepsilon)}$.

With these changes, mimicking the analysis of Theorem 4.2.1 of [CFGM16] gives the following theorem:

Theorem 57. *Suppose there exists a $k(n, \varepsilon, \delta)$-query algorithm, which, given* NACOND *access to an unknown distribution p over $[n]$ and a description of a distribution q over $[n]$ such that $\|q - \mathcal{U}_n\|_\infty \leq \frac{1}{100n}$, distinguishes between the cases $p = q$ versus $d_{\mathrm{TV}}(p, q) \geq \varepsilon$ with probability $1 - \delta$.*

Then there exists an algorithm which, given NACOND *access to an unknown distribution p on $[n]$ and a description of a distribution q, makes $\tilde{O}\left(\log(n/\varepsilon) \cdot k\left(n, \varepsilon/2, \frac{\log(1+1/100)}{6\log(100n/\varepsilon)}\right) + \frac{\sqrt{\log(n/\varepsilon)}}{\varepsilon^2}\right)$ queries to the oracle on p and distinguishes between the cases $p = q$ and $d_{\mathrm{TV}}(p, q) \geq \varepsilon$ with probability at least $2/3$.*

In the rest of this section, we will sketch how the analysis of Theorem 53 can be extended to apply to any distribution q such that $\|q - \mathcal{U}_n\|_\infty \leq \frac{1}{100n}$, while maintaining the same sample complexity:

Theorem 58 (Non-Adaptive Near-Uniform Identity Testing)**.** *There exists an algorithm which, given* NACOND *access to an unknown distribution p over $[n]$ and a description of a distribution q over $[n]$ such that $\|q - \mathcal{U}_n\|_\infty \leq \frac{1}{100n}$, makes $\tilde{O}\left(\frac{\log n}{\varepsilon^2}\right)$ queries to the oracle on p and distinguishes between the cases $p = q$ versus $d_{\mathrm{TV}}(p, q) \geq \varepsilon$ with probability at least $2/3$.*

[10] We note that the original definition of M_0 used in [CFGM13, CFGM16] appears to be an erratum, and a similar modification is required for the reduction to go through in their setting as well.

With this in hand, instantiating Theorem 57 with $k(n, \varepsilon, \delta) = \tilde{O}\left(\frac{\log n}{\varepsilon^2} \cdot \log(1/\delta)\right)$[11] gives Theorem 54.

Most of the analysis in Section 6.5.3 involves reasoning about the noise vector z, none of which changes for this setting. The exceptions are at the end of Sections 6.5.3.1 and 6.5.3.2, where we argue that (6.5) is large. We deal with the former case first – here, (6.5) can be written as

$$\varepsilon \cdot \left| \frac{z(i_1)q(i_2) - z(i_2)q(i_1)}{q(S)(q(S) + \varepsilon z(S))} \right| \geq \varepsilon \frac{2 \cdot \frac{1}{5n} \cdot \frac{99}{100n}}{\frac{202}{100n}\left(\frac{202}{100n} + \varepsilon \cdot \frac{32}{n}\right)} \geq \Omega(\varepsilon),$$

as desired. In the latter case, the proof follows with two minimal changes in the events that happen simultaneously (mentioned towards the end of the section). Instead of $\mathcal{U}_n(i) = 1/n$, we have that $q(i) \leq 101/100n$. Also, instead of $\frac{1}{2r} \leq \mathcal{U}_n(S \setminus i) \leq \frac{3}{2r}$, we have that $\frac{1}{2r} \leq q(S \setminus i) \leq \frac{3}{2r}$. This can be proved by essentially the same argument as Lemma 62, but rescaling at the end by a factor of $100n/99$ or $100n/101$. With these changes, the argument is identical, and thus we have Theorem 58, implying Theorem 54.

6.5.6 An Algorithm for Support Size Estimation

In this section, we sketch how similar – yet less involved – ideas as our algorithm for support size estimation (in Section 6.4) can be used to derive a non-adaptive upper bound for support-size estimation. For simplicity, we describe the algorithm for 2-approximation: adapting it to general $(1 + \varepsilon)$-approximation is straightforward.

The high-level idea is to perform a simple binary search (instead of the double exponential search from the preceding section) to identify the greatest lower bound on the support size of the form $k = 2^j$. For each guess $k \in \{2, 4, 8 \ldots, n\}$, we pick uniformly at random a set $S \subseteq [n]$ of cardinality k, and check whether p_S is uniform using the non-adaptive tester of Theorem 53. If p_S is found to be uniform for all values of k, we return n as our estimate (as the distribution is close to uniform on $[n]$); otherwise, we return n/k, for the smallest k on which p_S was found to be far from uniform. Indeed, p_S can only be far from uniform if S contains points from the support of p, which intuitively only happens if $n/k = \Omega(1)$.

[11] Note that a standard boosting applied to Theorem 58 gives a $1-\delta$ probability of success at a multiplicative cost of $\log(1/\delta)$.

To be more precise, the algorithm proceeds as follows, where $\tau > 0$ is an absolute constant:

for all $k \in \{2, 4, \ldots, n\}$ **do**
  Set a counter $c_k \leftarrow 0$.
  for $m = O(\log \log n)$ times **do**
    Pick uniformly at random a set $S \subseteq [n]$ of k elements.
    Test (non-adaptively) uniformity of p_S on S, with the tester of Theorem 53.
    if the tester rejects **then** increment c_k.
    end if
  end for
  if $c_k > \tau \cdot m$ **then return** $\tilde{\omega} \leftarrow \frac{n}{k}$.
  end if
end for
return $\tilde{\omega} \leftarrow n$.

The query complexity is easily seen to be $\operatorname{poly} \log n$, from the $\tilde{O}(\log n)$ calls to the $\tilde{O}(\log n)$ tester of Theorem 53. As for correctness, it follows from the fact that for any set S with mass $p(S) > 0$ which contains at least an η fraction of points outside the support, it holds that p_S is η-far from the uniform distribution on S.

6.6 Non-Adaptive Lower Bounds

In this section, we prove our lower bounds for non-adaptive support-size estimation and uniformity testing, Theorems 55 and 56.

Theorem 55 (Non-Adaptive Support-Size Estimation Lower Bound)**.** *Any algorithm which, given* NACOND *access to an unknown distribution p on $[n]$, estimates the size of the support up to a factor of $\gamma \geq \sqrt{2}$ must make at least $\Omega\left(\frac{\log n}{\log^2 \gamma}\right)$ queries.*

Theorem 56 (Non-Adaptive Uniformity Testing Lower Bound)**.** *Any algorithm which, given* NACOND *access to an unknown distribution p on $[n]$, distinguishes between the cases $p = \mathcal{U}_n$ and $d_{\mathrm{TV}}(p, \mathcal{U}_n) \geq \varepsilon$ with probability at least $2/3$ must make at least $\Omega(\log n/\varepsilon)$ queries.*

These two theorems follow from the same argument, which we outline below before

turning to the proof itself. Note that we will, in this section, establish Theorem 56 for *constant* ε, i.e. $\varepsilon = 1/4$; before explaining how to derive the general statement as a corollary at the end.

Structure of the proof. By Yao's Minimax Principle, we consider deterministic tests and study their performance over random distributions, chosen to be uniform over a random subset of carefully picked size. The proof of Theorem 55 then proceeds in three steps: in Lemma 68, we first argue that all queries made by the deterministic tester will (with high probability over the choice of the support size s) behave very "nicely" with regard to s, i.e. not be concentrated around it. Then, we condition on this to bound the total variation distance between the sequence of samples obtained in the two cases we "oppose," a random distribution from a family $\mathcal{P}$ and the corresponding one from a family $\mathcal{Q}$. In Lemma 69 we show that the part of total variation distance due to samples from the small queries is zero, except with probability $o(1)$ over the choice of s. Similarly, Lemma 69 allows us to say (comparing both cases to a third "reference" case, a honest-to-goodness uniform distribution over the whole domain, and applying a triangle inequality) that the remaining part of the total variation distance due to samples from the big queries is zero as well, except again with probability $o(1)$. Combining these three lets us conclude by properties of the total variation distance, as (since the queries are non-adaptive) the distribution over the sequence of samples is a product distribution. (Moreover, applying Lemma 69 as a stand-alone enables us, with little additional work, to obtain Theorem 56, as our argument in particular implies distributions from $\mathcal{P}$ are hard to distinguish from uniform.)

The families $\mathcal{P}$ and $\mathcal{Q}$. Fix $\gamma \geq \sqrt{2}$ as in Theorem 55; writing $\beta \triangleq \gamma^2$, we define the set

$$\mathcal{S} \triangleq \left\{ \beta^k n^{1/4} \ : \ 0 \leq k \leq \frac{\log n}{2 \log \beta} \right\} = \{n^{1/4}, \beta n^{1/4}, \beta^2 n^{1/4}, \ldots, , n^{3/4}\} \tag{6.12}$$

A no-instance $(p, q) \in \mathcal{P} \times \mathcal{Q}$ is then obtained by the following process:

- Draw s uniformly at random from $\mathcal{S}$.
- Pick a uniformly random set $S_1 \subseteq [n]$ of size s, and set p to be uniform on S_1.

- Pick a uniformly random set $S_2 \subseteq [n]$ of size βs, and set q to be uniform on S_2.

(Similarly, a yes-instance is obtained by first drawing a no-instance (p, q), and discarding q to keep only $(p, p) \in \mathcal{P} \times \mathcal{Q}$.)

We will argue that no algorithm can distinguish with high probability between the cases $(p, q) \sim \mathcal{P} \times \mathcal{Q}$ and $(p, q) \sim \mathcal{P} \times \mathcal{P}$, by showing that in both cases p and q generate transcripts indistinguishable from those the *uniform* distribution $\mathcal{U}_n$ would yield. This will imply Theorem 55, as any algorithm to estimate the support within a multiplicative γ would imply a distinguisher between instances of the form (p, p) and (p, q) (indeed, the support sizes of p and q differ by a factor $\beta = \gamma^2$). (As for Theorem 56, observe that any distribution $p \in \mathcal{P}$ has constant distance from the uniform distribution on $[n]$, so that a uniformity tester must be able to tell p apart from $\mathcal{U}_n$.)

Small and big query sets. Let $\mathcal{T}$ be any deterministic non-adaptive algorithm making $m_{\mathcal{T}} \leq m = \frac{1}{80000} \frac{\log n}{\log^2 \beta}$ queries. Without loss of generality, we can assume $\mathcal{T}$ makes exactly m queries, and denote them by $A_1, \ldots, A_m \subseteq [n]$. Moreover, we let $a_i = |A_i|$, and (again without loss of generality) write $a_1 \geq \cdots \geq a_m$.

As a preliminary observation, note that for any $A \subset [n]$ and $0 \leq s \leq n$ we have

$$\mathbf{E}_S \left[|S \cap A| \right] = \frac{|A| s}{n}$$

where the expectation is over a uniform choice of S among all $\binom{n}{s}$ subsets of size s. This observation will lead us to divide the query sets A_i in two groups, depending on the expected size of their intersection with the (random) support.

With this in mind, the following definition will be crucial to our proof. Intuitively, it captures the distribution of *sizes of intersection* of various query sets with the randomly chosen set S.

Definition 26. *Let $s \geq 1$, and $\mathcal{A} = \{a_1, \ldots, a_m\}$ be any set of q integers. For a real number $t > 0$, define*

$$C_t(s) \triangleq \left| \left\{ i \in [m] \ : \ \frac{a_i s}{n} \in \left(\beta^{-t}, \beta^t \right) \right\} \right| \tag{6.13}$$

to be the number of t-hit points of $\mathcal{A}$ (for s).

The next result will be crucial to prove our lower bounds: it roughly states that if we consider the set of a_i's and scale them by the random quantity s/n, then the distribution of the random variable generated has an exponential tail with respect to t.

Lemma 68 (Hitting Lemma)**.** *Fix $\mathcal{A}$ as in the previous definition. If s is drawn uniformly at random from $\mathcal{S}$, then with probability at least* 99/100*.*

$$\sup_{t>0} \frac{C_t(s)}{t} < \frac{2}{100}. \tag{6.14}$$

Proof. Without loss of generality, assume that the a_i's are in decreasing order. We will work in the logarithmic domain: a number a_i contributes to $C_t(s)$ if and only if $\log s \in [\log(n/a_i) - t\log\beta, \log(n/a_i) + t\log\beta]$. Indeed, we can restate the lemma in an additive form. Let $\mathcal{A} = \{\alpha_1, \ldots, \alpha_m\}$ be any set of numbers in $[0, \log n]$. These are defined as transformations of a_i's: $\alpha_i \triangleq \log(n/a_i)$. In the additive restating, x will play the role of $\log s$, or equivalently, $s = 2^x$. For a point $x \in [0, \log n]$, let ℓ_j^x (r_j^x respectively) denote the distance of x from the jth point to its left (right respectively) from the set $\mathcal{A}$. More precisely, if $\alpha_\gamma \le x \le \alpha_{\gamma+1}$, $\ell_j^x \triangleq x - \alpha_{\gamma+1-j}$. If we consider only points to the left of x, $\ell_j^x/\log\beta$ is the least value of t such that $C_t(2^x) = j$. Therefore, if $t_j^x \triangleq \frac{1}{\log\beta}\min\left\{\ell_j^x, r_j^x\right\}$, then we are guaranteed that $j \le C_{t_j^x}(2^x) \le 2j$.

Observe that $C_t(2^x)$ is a piecewise-constant function which is monotone non-decreasing in t. Therefore, the supremum of $\frac{C_t(2^x)}{t}$ is attained at one of these discontinuities:

$$\sup_{t>0} \frac{C_t(2^x)}{t} = \max_{\{t\,:\,\forall\delta>0, C_{t-\delta}(2^x)<C_t(2^x)\}} \frac{C_t(2^x)}{t}.$$

We can in turn upper bound this by looking at the set of all t_j^x. Note that this may ignore some discontinuous points: for instance, suppose that $\ell_j^x < r_j^x$, then $r_j^x/\log\beta$ will not be considered. However, we note that in this case, $C_{r_j^x/\log\beta}(2^x) \le 2C_{\ell_j^x/\log\beta}(2^x) = 2C_{t_j^x}(2^x)$:

$$\frac{C_{r_j^x/\log\beta}(2^x)}{r_j^x/\log\beta} \le \frac{C_{r_j^x/\log\beta}(2^x)}{\ell_j^x/\log\beta} \le \frac{2C_{\ell_j^x/\log\beta}(2^x)}{\ell_j^x/\log\beta} = \frac{2C_{t_j^x}(2^x)}{t_j^x}.$$

Therefore,

$$\sup_{t>0} \frac{C_t(2^x)}{t} \leq \max_{\{t_j^x \,:\, j\in[m]\}} \frac{2C_{t_j^x}(2^x)}{t_j^x} \leq \max_{\{t_j^x \,:\, j\in[m]\}} \frac{4j}{t_j^x} = \max_{\{t_j^x \,:\, j\in[m]\}} \frac{4j\log\beta}{\min\{\ell_j^x, r_j^x\}}$$

We would like to upper bound this term by $2/100$. Equivalently, we satisfy the lemma conditions if

$$\min_j \left\{ \frac{t_j^x}{j} \right\} \geq 200\log\beta.$$

For a constant $c > 0$, let S_c be the set of all points x (where recall $s = 2^x$ is selected according to the distribution $\mathcal{S}$) such that violate this inequality for c:

$$\min_j \left\{ \frac{t_j^x}{j} \right\} \leq c.$$

We would like to upper bound the probability that a randomly selected x violates this inequality for $c = 200\log\beta$ by $1/100$. Equivalently, since x is selected uniformly at random from $\frac{\log n}{2\log\beta}$ different values, we would like to upper bound the size of $S_{200\log\beta}$ by $\log n/200\log\beta$. We do this with the following claim:

Claim 12. $|S_c| \leq 2cm$.

Substituting in $c = 200\log\beta$ and $m = \log n/80000\log^2\beta$ will give the desired result.

Proof of Claim 12. We consider the set of points in $S_{c,\ell} \subset S_c$ that satisfy $\ell_j^x/j < c$ for some j, and show that their measure is at most cm. An identical bound holds for the set of points of S_c for which $r_j^x/j < c$. Let $S_{c,\ell}^i \subset S_{c,\ell}$ be the set of points in $S_{c,\ell}$ that satisfy $\min_j \left\{ \frac{t_j^x}{j} \right\} < c$ with respect to the set $\alpha_1, \ldots, \alpha_i$. We will show by induction that $|S_{c,\ell}^i| < ci$.

For the first point α_1, the set $S_{c,\ell}^1 = [\alpha_1, \alpha_1 + c]$. Suppose by induction that $|S_{c,\ell}^i| < ci$. Let x_i be the right-most point in the set $S_{c,\ell}^i$. Then it is clear that $x_i > \alpha_i$, in fact $x_i \geq \alpha_i + c$. Furthermore, either $x_i = \log n$, or $\ell_j^{x_i}/j = c$ for some j. Moreover, we claim that $[\alpha_i, x_i] \in S_{c,\ell}^i$. Indeed, for the same j that $\ell_j^{x_i}/j < c$, all points in $[\alpha_i, x_i]$ satisfy the condition. If $x_i = \log n$, then the result holds trivially. We therefore consider the point α_{i+1} and prove the inductive step for $x_i < \log n$. There are two cases:

If $\alpha_{i+1} \geq x_i$: In this case, $S^{i+1}_{c,\ell} = S^i_{c,\ell} \cup [\alpha_{i+1}, x_{i+1}]$. We have to show that $x_{i+1} \leq \alpha_{i+1} + c$. Suppose to the contrary that $x_{i+1} > \alpha_{i+1} + c \geq x_i + c$. Then there is a point α_h for $h \leq i$, such that $\frac{x_{i+1}-\alpha_h}{i+2-h} < c$, and then $\frac{\alpha_{i+1}+c-\alpha_h}{i+2-h} < c$, so that

$$\frac{\alpha_{i+1} - \alpha_h}{i+1-h} < c,$$

however, this implies that $\alpha_{i+1} \in S^i_{c,\ell}$, contradicting the assumption of this case.

If $\alpha_{i+1} < x_i$: In this case, $S^{i+1}_{c,\ell} = S^i_{c,\ell} \cup [x_i, x_{i+1}]$. We have to show that $x_{i+1} \leq x_i + c$. Suppose on the contrary that $x_{i+1} > x_i + c > \alpha_{i+1} + c$. Suppose h be the index such that $\frac{x_{i+1}-\alpha_h}{i+2-h} < c$, and therefore, $\frac{x_i+c-\alpha_h}{i+2-h} < c$, implying that

$$\frac{x_i - \alpha_h}{i+1-h} < c,$$

contradicting that x_i is the rightmost point of $S^i_{c,\ell}$.

□

This concludes the proof of the hitting lemma. □

We proceed to show how to use this lemma to bound the contribution of various types of queries to the distinguishability of p and q. In particular, we will apply Lemma 68 to the set of query sizes $\{a_1, \ldots, a_m\}$.

Recall that the a_i's are non-increasing. If $a_{m'}s/n \leq 1$ let $m' \triangleq m + 1$, otherwise define m' as the largest integer such that $a_{m'}s/n > 1$. We partition the queries made by $\mathcal{T}$ in two: $A_1, \ldots, A_{m'}$ are said to be *big*, while $A_{m'+1}, \ldots, A_m$ are *small queries*.

Lemma 69. *With probability at least $1 - 2^{-10}$, a random distribution from $\mathcal{P}$ or from $\mathcal{Q}$ does not intersect with any small query.*

Proof. Let s be the random size drawn for the definition of the instances. We first claim that $\mathbf{E}[|A_{m'+j} \cap S|] \leq \beta^{-50j}$ for all $j \geq 1$, where the expectation is over the uniform choice of set S_1 for p. Indeed, by contradiction suppose there is a $j \geq 1$ such that $\mathbf{E}[|A_{m'+j} \cap S|] =$

$\frac{a_{m'+j}s}{n} > \beta^{-50j}$. By definition of m', for $1 \le i \le j$,

$$1 \ge \frac{a_{m'+i}s}{n} > \beta^{-50j}.$$

Therefore, the queries $A_{m'}, A_{m'+1}, \ldots, A_{m'+j}$ contribute to C_{50j}, and we obtain $\frac{C_{50j}}{50j} \ge \frac{j}{50j} = \frac{2}{100}$, contradicting Lemma 68. Thus, the expected intersection can be bounded as follows:

$$\begin{aligned}\mathbf{E}[|(A_{m'+1} \cup A_{m'+2} \cdots \cup A_m) \cap S|] &\le \mathbf{E}[|A_{m'+1} \cap S|] + \mathbf{E}[|A_{m'+2} \cap S|] + \cdots + \mathbf{E}[|A_m \cap S|] \\ &\le \beta^{-50} + \beta^{-100} + \ldots \\ &\le 2^{-12},\end{aligned}$$

since $\beta \ge 2$. From this, we obtain the result holds for $\mathcal{P}$ by Markov's inequality. The same applies to $\mathcal{Q}$ with probability of intersection at most 2^{-10}, proving the lemma. □

We now turn our attention to the sets with *large* intersections. We will show that under $\mathcal{P}$ and $\mathcal{Q}$, the output of querying the sets $A_1, \ldots A_{m'}$ are indistinguishable from simply picking elements uniformly from the sets $A_1, \ldots, A_{m'}$. More precisely, we establish the following.

Lemma 70. *Let $\eta^* = 2^{-10}$ and $\eta_s = 1/100$; and fix $\ell \in \{1, 2\}$. At least an $1 - \eta_s$ fraction of elements $s_1, \ldots, s_{m'} \in A_1 \times A_2, \ldots, A_{m'}$ satisfy*

$$\Pr_{\ell}[\,(s_1, \ldots, s_{m'})\,] \in [1 - 5\eta^*, 1 + 5\eta^*] \cdot \frac{1}{|A_1| \ldots |A_{m'}|}, \tag{6.15}$$

for $\ell \in \{p, q\}$.

As this holds for most distributions in both $\mathcal{P}$ and $\mathcal{Q}$, this implies the queries are indistinguishable from the product distribution over $A_1 \times A_2, \ldots, A_{m'}$ (which is the one induced by the same queries on the uniform distribution over $[n]$) in either case, with probability at least $1 - \eta_s - 5\eta^*$.

Proof of Lemma 70. From standard Chernoff-like concentration bounds for hypergeometric random variables (Lemma 3), we obtain the claim below.

Claim 13. *Suppose A is a set of size a, and S is a uniformly chosen random set of size s. Then, for all $\eta \in (0,1]$, we have $\Pr\big[\,|A\cap S| > (1+\eta)\frac{as}{n}\,\big] < e^{-\eta^2\cdot\frac{as}{3n}}$ and $\Pr\big[\,|A\cap S| < (1-\eta)\frac{as}{n}\,\big] < e^{-\eta^2\cdot\frac{as}{3n}}$.*

We use this to show that indeed all the $|A_i \cap S|$ concentrate around their expected values for $1 \le j \le m'$. First note that, as a consequence of Lemma 68, it is the case that these expected values satisfy $a_{m'-j}s/n \ge \beta^{50(j+1)}$ for every $0 \le j \le m'-1$ (with probability at least 99/100). Conditioning on this, we first invoke Lemma 13 on A_j with $\eta = 3\cdot\beta^{20(j+1)}$, and then apply a union bound to obtain

$$\Pr\Big[\,\exists j \in [m'] \text{ s.t. } |A_j\cap S| \notin \Big[1-4\cdot\beta^{-20(j+1)}, 1+4\cdot\beta^{-20(j+1)}\Big]\cdot\frac{a_j s}{n}\,\Big] < e^{-\beta^{10}} \tag{6.16}$$

i.e., with high probability all intersections simultaneously concentrate around their expected values.

Note that since s is at most $n^{3/4}$, each A_i under consideration has size at least $n\beta^{50}/n^{3/4} > n^{1/4}$. Therefore, the probability that a random selection of elements from $A_1,\ldots,A_{m'}$ exhibits no collision is at least

$$\prod_{i=1}^{m'}\frac{|A_i|-m'}{|A_i|} \ge \left(1-\frac{m'}{n^{1/4}}\right)^{m'} \ge 1-\frac{(m')^2}{n^{1/4}} > 1-\frac{\log^2 n}{n^{1/4}}.$$

We henceforth condition on this event.

Let $N = \binom{n}{s}$ be the number of outcomes for the set S. We write $N_0 \ge N(1-e^{-\beta^{10}})$ for the number of such sets for which (6.16) holds. Let $s_1^{m'}$ denote $s_1\ldots s_{m'}$. For a set of distinct $(s_1,\ldots,s_{m'}) \in A_1\times\cdots\times A_{m'}$, let $N(s_1^{m'}) = \binom{n-m'}{s-m'}$ be the number of sets of size s that contain $s_1^{m'}$, and let $N_0(s_1^{m'})$ of them satisfy (6.16).

By Markov's inequality, with probability at least $1-e^{-\beta^9}$, for a randomly chosen $s_1^{m'}$ we have $N_0(s_1^{m'})/N(s_1^{m'}) > 1-e^{2-\beta^9}$. For any such $s_1^{m'}$,

$$\begin{aligned}\Pr\Big[\,s_1^{m'}\,\Big] &\ge \frac{N_0(s_1^{m'})}{N}\cdot\prod_{i=1}^{m'}\frac{1}{|A_i\cap S|} \ge (1-e^{2-\beta^9})\frac{s(s-1)\ldots(s-m'+1)}{n(n-1)\ldots(n-m'+1)}\cdot(1-4\cdot\beta^{-19})\prod_{i=1}^{m'}\frac{n}{a_i s}\\ &\ge (1-6\cdot\beta^{-19})\prod_{i=1}^{m'}\frac{1}{a_i},\end{aligned}$$

for large n and as $|S| > n^{1/4}$. Since the sum of probabilities of elements is at most 1, the other side of the inequality in Lemma 70 follows. □

Proof of Theorem 55 and Theorem 56 : Let T_1 (resp. T_2, T_U) be the distribution over transcripts generated by the queries $A_1, \ldots, A_m$ when given conditional access to the distribution p from a no-instance (resp. q, resp. uniform $\mathcal{U}_n$); that is, a distribution over m-tuples in $A_1 \times \cdots \times A_m$. Since the queries were non-adaptive, we can break T_1 (and similarly for T_2, T_U) in $T_1^{\text{big}} \times T_1^{\text{small}}$ according to m', and use Lemma 70 and Lemma 69 separately to obtain $d_{\mathrm{TV}}(T_1, T_U) \leq \eta_s + \eta^* + 2^{-10} < 1/50$ and $d_{\mathrm{TV}}(T_1, T_U) \leq \eta_s + \eta^* + 2^{-10} < 1/50$ (for the latter, recalling that queries that do not intersect the support receive samples exactly uniformly distributed in the query set) – thus establishing both theorems. □

On the dependence on ε in Theorem 56. We remark that Theorem 56, by establishing a lower bound of $\Omega(\log n)$ queries for non-adaptive testing of uniformity with constant distance parameter $1/4$, immediately implies, by a standard argument, an $\Omega((\log n)/\varepsilon)$ lower bound for distance parameter $\varepsilon \in (0, 1/4)$. In more detail, this is a consequence of the following reduction: any $m(n,\varepsilon)$-query non-adaptive tester for uniformity $\mathcal{T}$ can be used, given conditional access to some distribution p on $[n]$, on the mixture distribution

$$p_\varepsilon \triangleq 4\varepsilon p + (1 - 4\varepsilon)\mathcal{U}_n \,, \tag{6.17}$$

for which a conditional oracle can be easily simulated given a conditional oracle for p. Moreover, answering $m(n,\varepsilon)$ to p_ε can be done with an expected $4\varepsilon m(n,\varepsilon)$ conditional queries to p. As it is immediate to see that $d_{\mathrm{TV}}(p_\varepsilon, \mathcal{U}_n) = 4\varepsilon d_{\mathrm{TV}}(p, \mathcal{U}_n)$, we get that $\mathcal{T}$ can be used to obtain a tester for non-adaptive testing of uniformity with constant distance parameter $1/4$, with query complexity $O(\varepsilon m(n,\varepsilon))$ for every $\varepsilon < 1/4$ (converting the expected query complexity to a worst-case one is straightforward via Markov's inequality followed by success probability amplification by a constant number of repetitions). Therefore, the lower bound of Theorem 56 implies that $m(n,\varepsilon) = \Omega((\log n)/\varepsilon)$, as claimed.

It is also worth noting that the above argument does *not* yield an analogue statement for support-size estimation via Theorem 55. Indeed, mixing the distribution p with the uniform

distribution does not preserve the support size in that case (nor the guarantee that every point of the support has probability mass at least τ/n).

Chapter 7

Other Directions in Distribution Testing

This thesis focused on several recent directions in distribution testing which the author has worked on – it does not attempt to be a comprehensive resource for all problems worth studying. Below, we highlight a few additional directions which may be of interest. We hope that some of these suggestions inspire additional work in this field.

Communication Constrained Testing. Often, data may be collected on a number of different devices simultaneously, and we wish to perform some statistical procedure on the dataset as a whole. Some examples include data collected via a sensor network, or user data generated on a number of remote devices like cellphones. One approach is to send all the data to a central authority who then performs the statistical procedure, but this involves communicating a significant amount of data over the network. The pertinent question is whether one can reduce the amount of communication required. This appears to be an emerging area of interest within distribution testing [ACT18, FMO18], and there are likely to be connections with the field of sketching [SS02].

Memory Constrained Testing. In a similar vein to the previous direction, we may wish to perform some statistical task on a computing device with restricted memory. For instance, a cellphone may need to run some background processes with limited memory, in order to avoid slowing the phone significantly and ruining the user experience. This type of memory-restricted statistical inference has been studied recently in the learning community [Raz16,

KRT17a, Raz17, MM17, GRT18, MM18], but has not yet seen significant study in the distribution testing and property estimation community. [GMV06] is one classic result which studies the estimation of entropy and various distribution divergences in this setting. There are likely to be significant connections with the streaming literature [AMS99, Mut05].

Sampling Correctors. The work of Canonne, Gouleakis, and Rubinfeld [CGR18] initiates the study of *sampling correctors*. A sampling corrector is an algorithm which receives samples from a distribution which is close to having some property $\mathcal{C}$, and outputs samples from a nearby distribution which has the property $\mathcal{C}$. Observe that this can be viewed as a generalization of the agnostic learning problem. The authors investigate various connections between sampling correctors, agnostic learning, and distribution testing. While this paper sets the stage for study, there is much room for investigating sampling correctors for a number of classes of interest.

Quantum Tomography, Certification, and Estimation. As quantum devices become more prevalent, methods to estimate and test properties of quantum states will become increasingly important. There has recently been significant interest in studying the copy complexity required to estimate various properties of mixed quantum states. Quantum tomography is the quantum equivalent of distribution learning, and is studied in [FGLE12, Wri16, OW16a, OW17, OW16b, HHJ$^+$17, KRT17b, Aar18, ACHN18]. There are also quantum equivalents of distribution testing (quantum state certification), support size testing (rank testing) and entropy estimation (entropy estimation). See [MdW16] for a survey of this field, and [HM13a, OW15, HLM17, GNW17, BOW17, AISW17, PLM18] for some recent works in this area. The above works focus on quantum algorithms for testing of quantum states. A slightly different direction is to use quantum algorithms for classical distribution testing problems, as considered in [BHH11, LW17]

Tolerant Testing. In Chapter 2, we investigated tolerant distribution testing with respect to a number of different distances. While we obtained a tight understanding for all the problems we considered, there are still a number of fundamental questions which one can ask when in search of strongly-sublinear tolerant testers. First, we only considered a

limited number of distribution distances, and additional distance measures might give rise to testing problems with new sample complexities. For instance, we conjecture that there exists a hierarchy of distances between χ^2-distance and KL divergence, and that testing with tolerance to these intermediate distances might allow one to interpolate between the $\Theta(n^{1/2})$ and $\Theta(n/\log n)$ sample complexities at either end. Second, our focus in Chapter 2 was on testing problems with a constant factor gap between the soundness and completeness cases. It is likely that one can reduce the sample complexity of these problems by increasing this gap, thereby reducing the amount of tolerance provided by the testing algorithm.

High-Dimensional Testing. As covered in Chapter 4, one can efficiently test distributions over a multivariate domain if the underlying density is from an Ising model. Similar results exist for testing Bayesian networks [DP17, CDKS17, ABDK18]. This is far from a comprehensive list of all multivariate distributions, and there are many natural structures which may permit distribution testing with complexity which is polynomial in the dimension. One clear open problem is to understand the complexity of testing Markov Random Fields (MRFs), the broad class of distributions for which the Ising model is the prototypical example.

Bibliography

[AA01] Dakshi Agrawal and Charu C. Aggarwal. On the design and quantification of privacy preserving data mining algorithms. In *Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '01, pages 247–255, New York, NY, USA, 2001. ACM.

[AAK+07] Noga Alon, Alexandr Andoni, Tali Kaufman, Kevin Matulef, Ronitt Rubinfeld, and Ning Xie. Testing k-wise and almost k-wise independence. In *Proceedings of the 39th Annual ACM Symposium on the Theory of Computing*, STOC '07, pages 496–505, New York, NY, USA, 2007. ACM.

[Aar18] Scott Aaronson. Shadow tomography of quantum states. In *Proceedings of the 50th Annual ACM Symposium on the Theory of Computing*, STOC '18, pages 325–338, New York, NY, USA, 2018. ACM.

[ABDK18] Jayadev Acharya, Arnab Bhattacharyya, Constantinos Daskalakis, and Saravanan Kandasamy. Learning and testing causal models with interventions. *arXiv preprint arXiv:1805.09697*, 2018.

[ACFT18] Jayadev Acharya, Clément L. Canonne, Cody Freitag, and Himanshu Tyagi. Test without trust: Optimal locally private distribution testing. *arXiv preprint arXiv:1808.02174*, 2018.

[ACHN18] Scott Aaronson, Xinyi Chen, Elad Hazan, and Ashwin Nayak. Online learning of quantum states. *arXiv preprint arXiv:1802.09025*, 2018.

[ACK15a] Jayadev Acharya, Clément L. Canonne, and Gautam Kamath. Adaptive estimation in weighted group testing. In *Proceedings of the 2015 IEEE International Symposium on Information Theory*, ISIT '15, pages 2116–2120, Washington, DC, USA, 2015. IEEE Computer Society.

[ACK15b] Jayadev Acharya, Clément L. Canonne, and Gautam Kamath. A chasm between identity and equivalence testing with conditional queries. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques.*, RANDOM '15, pages 449–466, Dagstuhl, Germany, 2015. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[ACS10] Michat Adamaszek, Artur Czumaj, and Christian Sohler. Testing monotone continuous distributions on high-dimensional real cubes. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, pages 56–65, Philadelphia, PA, USA, 2010. SIAM.

[ACT18] Jayadev Acharya, Clément L. Canonne, and Himanshu Tyagi. Distributed simulation and distributed inference. *arXiv preprint arXiv:1804.06952*, 2018.

[AD11] Anne Auger and Benjamin Doerr. *Theory of Randomized Search Heuristics: Foundations and Recent Developments*. World Scientific, 2011.

[AD15] Jayadev Acharya and Constantinos Daskalakis. Testing Poisson binomial distributions. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 1829–1840, Philadelphia, PA, USA, 2015. SIAM.

[ADJ+11] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, and Shengjun Pan. Competitive closeness testing. In *Proceedings of the 24th Annual Conference on Learning Theory*, COLT '11, pages 47–68, 2011.

[ADJ+12] Jayadev Acharya, Hirakendu Das, Ashkan Jafarpour, Alon Orlitsky, Shengjun Pan, and Ananda Suresh. Competitive classification and closeness testing. In *Proceedings of the 25th Annual Conference on Learning Theory*, COLT '12, pages 22.1–22.18, 2012.

[ADK15] Jayadev Acharya, Constantinos Daskalakis, and Gautam Kamath. Optimal testing for properties of distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 3577–3598. Curran Associates, Inc., 2015.

[ADLS17] Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, pages 1278–1289, Philadelphia, PA, USA, 2017. SIAM.

[ADOS17] Jayadev Acharya, Hirakendu Das, Alon Orlitsky, and Ananda Theertha Suresh. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 11–21. JMLR, Inc., 2017.

[ADR18] Maryam Aliakbarpour, Ilias Diakonikolas, and Ronitt Rubinfeld. Differentially private identity and closeness testing of discrete distributions. In *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pages 169–178. JMLR, Inc., 2018.

[Agr12] Alan Agresti. *Categorical Data Analysis*. Wiley, 2012.

[AISW17] Jayadev Acharya, Ibrahim Issa, Nirmal V. Shende, and Aaron B. Wagner. Measuring quantum entropy. *arXiv preprint arXiv:1711.00814*, 2017.

[AJ06] José A. Adell and Pedro Jodrá. Exact Kolmogorov and total variation distances between some familiar discrete distributions. *Journal of Inequalities and Applications*, 2006(1):1–8, 2006.

[AJOS13] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. A competitive test for uniformity of monotone distributions. In *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics*, AISTATS '13, pages 57–65. JMLR, Inc., 2013.

[AJOS14a] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Efficient compression of monotone and m-modal distributions. In *Pro-*

ceedings of the 2014 IEEE International Symposium on Information Theory, ISIT '14, pages 1867–1871, Washington, DC, USA, 2014. IEEE Computer Society.

[AJOS14b] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Sublinear algorithms for outlier detection and generalized closeness testing. In *Proceedings of the 2014 IEEE International Symposium on Information Theory*, ISIT '14, pages 3200–3204, Washington, DC, USA, 2014. IEEE Computer Society.

[AKN06] Pieter Abbeel, Daphne Koller, and Andrew Y. Ng. Learning factor graphs in polynomial time and sample complexity. *Journal of Machine Learning Research*, 7(Aug):1743–1788, 2006.

[AKSZ18] Jayadev Acharya, Gautam Kamath, Ziteng Sun, and Huanyu Zhang. Inspectre: Privately estimating the unseen. In *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pages 30–39. JMLR, Inc., 2018.

[AMS99] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.

[AOST17] Jayadev Acharya, Alon Orlitsky, Ananda Theertha Suresh, and Himanshu Tyagi. Estimating Rényi entropy of discrete distributions. *IEEE Transactions on Information Theory*, 63(1):38–56, 2017.

[Arb10] John Arbuthnott. Ii. an argument for divine providence, taken from the constant regularity observ'd in the births of both sexes. by Dr. John Arbuthnott, physitian in ordinary to her majesty, and fellow of the college of physitians and the royal society. *Philosophical Transactions of the Royal Society of London*, 27:186–190, 1710.

[ASZ17] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private testing of identity and closeness of discrete distributions. *arXiv preprint arXiv:1707.05128*, 2017.

[ASZ18] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Communication efficient, sample optimal, linear time locally private discrete distribution estimation. *arXiv preprint arXiv:1802.04705*, 2018.

[AW89] Nabil R. Adam and John C. Worthmann. Security-control methods for statistical databases: A comparative study. *ACM Computing Surveys (CSUR)*, 21(4):515–556, 1989.

[BBBB72] Richard E. Barlow, David J. Bartholomew, J. M. Bremner, and Hugh D. Brunk. *Statistical Inference Under Order Restrictions: Theory and Application of Isotonic Regression*. Wiley, 1972.

[BBBY12] Maria-Florina Balcan, Eric Blais, Avrim Blum, and Liu Yang. Active property testing. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '12, pages 21–30, Washington, DC, USA, 2012. IEEE Computer Society.

[BBKN14] 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.

[BC17] Tuğkan Batu and Clément L. Canonne. Generalized uniformity testing. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '17, pages 880–889, Washington, DC, USA, 2017. IEEE Computer Society.

[BC18] Rishiraj Bhattacharyya and Sourav Chakraborty. Property testing of joint distributions using conditional samples. *Transactions on Computation Theory*, 10(4):16:1–16:20, 2018.

[BCG17] Eric Blais, Clément L. Canonne, and Tom Gur. Distribution testing lower bounds via reductions from communication complexity. In *Proceedings of the 32nd Computational Complexity Conference*, CCC '17, pages 28:1–28:40, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[BDKR05] Tuğkan Batu, Sanjoy Dasgupta, Ravi Kumar, and Ronitt Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 35(1):132–150, 2005.

[BDL+17] Maria-Florina Balcan, Travis Dick, Yingyu Liang, Wenlong Mou, and Hongyang Zhang. Differentially private clustering in high-dimensional euclidean spaces. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 322–331. JMLR, Inc., 2017.

[BFF+01] Tuğkan Batu, Eldar Fischer, Lance Fortnow, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '01, pages 442–451, Washington, DC, USA, 2001. IEEE Computer Society.

[BFR+00] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '00, pages 259–269, Washington, DC, USA, 2000. IEEE Computer Society.

[BFR+13] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing closeness of discrete distributions. *Journal of the ACM*, 60(1):4:1–4:25, 2013.

[BFRV11] Arnab Bhattacharyya, Eldar Fischer, Ronitt Rubinfeld, and Paul Valiant. Testing monotonicity of distributions over general partial orders. In *Proceedings of the 2nd Conference on Innovations in Computer Science*, ICS '11, pages 239–252, Beijing, China, 2011. Tsinghua University Press.

[BGS14] Guy Bresler, David Gamarnik, and Devavrat Shah. Structure learning of antiferromagnetic Ising models. In *Advances in Neural Information Processing Systems 27*, NIPS '14, pages 2852–2860. Curran Associates, Inc., 2014.

[Bha16] Bhaswar B. Bhattacharya. Power of graph-based two-sample tests. *arXiv preprint arXiv:1508.07530*, 2016.

[BHH11] Sergey Bravyi, Aram H. Harrow, and Avinatan Hassidim. Quantum algorithms for testing properties of distributions. *IEEE Transactions on Information Theory*, 57(6):3971–3981, 2011.

[BIKK12] Christina Brandt, Nicole Immorlica, Gautam Kamath, and Robert Kleinberg. An analysis of one-dimensional Schelling segregation. In *Proceedings of the 44th Annual ACM Symposium on the Theory of Computing*, STOC '12, pages 789–804, New York, NY, USA, 2012. ACM.

[Bir87] Lucien Birgé. Estimating a density under order restrictions: Nonasymptotic minimax risk. *The Annals of Statistics*, 15(3):995–1012, 1987.

[BJRP13] Fadoua Balabdaoui, Hanna Jankowski, Kaspar Rufibach, and Marios Pavlides. Asymptotics of the discrete log-concave maximum likelihood estimator and related applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):769–790, 2013.

[BK16] Guy Bresler and Mina Karzand. Learning a tree-structured Ising model in order to make predictions. *arXiv preprint arXiv:1604.06749*, 2016.

[BK18] Quentin Berthet and Varun Kanade. Statistical windows in testing for the initial distribution of a reversible markov chain. *arXiv preprint arXiv:1808.01857*, 2018.

[BKR04] Tuğkan Batu, Ravi Kumar, and Ronitt Rubinfeld. Sublinear algorithms for testing monotone and unimodal distributions. In *Proceedings of the 36th Annual ACM Symposium on the Theory of Computing*, STOC '04, New York, NY, USA, 2004. ACM.

[BLM13] Stephane Boucheron, Gabor Lugosi, and Pierre Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[BM16] Bhaswar B. Bhattacharya and Sumit Mukherjee. Inference in Ising models. *Bernoulli*, 2016.

[BN18] Guy Bresler and Dheeraj Nagaraj. Optimal single sample tests for structured versus unstructured network data. In *Proceedings of the 31st Annual Conference on Learning Theory*, COLT '18, pages 1657–1690, 2018.

[BNS15] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Learning privately with labeled and unlabeled examples. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 461–477, Philadelphia, PA, USA, 2015. SIAM.

[BNS16a] Amos Beimel, Kobbi Nissim, and Uri Stemmer. Private learning and sanitization: Pure vs. approximate differential privacy. *Theory of Computing*, 12(1):1–61, 2016.

[BNS16b] Mark Bun, Kobbi Nissim, and Uri Stemmer. Simultaneous private learning of multiple concepts. In *Proceedings of the 7th Conference on Innovations in Theoretical Computer Science*, ITCS '16, pages 369–380, New York, NY, USA, 2016. ACM.

[BNSV15] Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 634–649, Washington, DC, USA, 2015. IEEE Computer Society.

[BOW17] Costin Bădescu, Ryan O'Donnell, and John Wright. Quantum state certification. *arXiv preprint arXiv:1708.06002*, 2017.

[Bre15] Guy Bresler. Efficiently learning Ising models on arbitrary graphs. In *Proceedings of the 47th Annual ACM Symposium on the Theory of Computing*, STOC '15, pages 771–782, New York, NY, USA, 2015. ACM.

[BS16] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Proceedings of the 14th Conference on Theory of Cryptography*, TCC '16-B, pages 635–658, Berlin, Heidelberg, 2016. Springer.

[BV15] Bhaswar Bhattacharya and Gregory Valiant. Testing closeness with unequal sized samples. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 2611–2619. Curran Associates, Inc., 2015.

[BW10] Fadoua Balabdaoui and Jon A. Wellner. Estimation of a k-monotone density: Characterizations, consistency and minimax lower bounds. *Statistica Neerlandica*, 64(1):45–70, 2010.

[BW17a] Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *arXiv preprint arXiv:1706.10003*, 2017.

[BW17b] Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for high-dimensional multinomials: A selective review. *arXiv preprint arXiv:1712.06120*, 2017.

[BWM97] Michael J. Berry, David K Warland, and Markus Meister. The structure and precision of retinal spike trains. *Proceedings of the National Academy of Sciences*, 94(10):5411–5416, 1997.

[Can15a] Clément L. Canonne. Big data on the rise? - testing monotonicity of distributions. In *Proceedings of the 42nd International Colloquium on Automata, Languages, and Programming*, ICALP '15, pages 294–305, 2015.

[Can15b] Clément L. Canonne. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22(63), 2015.

[Can16] Clément L. Canonne. Are few bins enough: Testing histogram distributions. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '16, pages 455–463, New York, NY, USA, 2016. ACM.

[Can17] Clément L. Canonne. A short note on Poisson tail bounds. `http://www.cs.columbia.edu/~ccanonne/files/misc/2017-poissonconcentration.pdf`, 2017.

[CCG+12] Robert K. Colwell, Anne Chao, Nicholas J. Gotelli, Shang-Yi Lin, Chang Xuan Mao, Robin L. Chazdon, and John T. Longino. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, 5(1):3–21, 2012.

[CDGR16] Clément L. Canonne, Ilias Diakonikolas, Themis Gouleakis, and Ronitt Rubinfeld. Testing shape restrictions of discrete distributions. In *Proceedings of the 33rd Symposium on Theoretical Aspects of Computer Science*, STACS '16, pages 25:1–25:14, Dagstuhl, Germany, 2016. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[CDK17] Bryan Cai, Constantinos Daskalakis, and Gautam Kamath. Priv'it: Private and sample efficient identity testing. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 635–644. JMLR, Inc., 2017.

[CDKS17] Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing Bayesian networks. In *Proceedings of the 30th Annual Conference on Learning Theory*, COLT '17, pages 370–448, 2017.

[CDKS18] Clément L. Canonne, Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Testing conditional independence of discrete distributions. In *Proceedings of*

the 50th Annual ACM Symposium on the Theory of Computing, STOC '18, pages 735–748, New York, NY, USA, 2018. ACM.

[CDS17] Clément L. Canonne, Ilias Diakonikolas, and Alistair Stewart. Fourier-based testing for families of distributions. *arXiv preprint arXiv:1706.05738*, 2017.

[CDSS14] Siu On Chan, Ilias Diakonikolas, Rocco A. Servedio, and Xiaorui Sun. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the 46th Annual ACM Symposium on the Theory of Computing*, STOC '14, pages 604–613, New York, NY, USA, 2014. ACM.

[CDVV14] Siu On Chan, Ilias Diakonikolas, Gregory Valiant, and Paul Valiant. Optimal algorithms for testing closeness of discrete distributions. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 1193–1203, Philadelphia, PA, USA, 2014. SIAM.

[CFGM13] Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the power of conditional samples in distribution testing. In *Proceedings of the 4th Conference on Innovations in Theoretical Computer Science*, ITCS '13, pages 561–580, New York, NY, USA, 2013. ACM.

[CFGM16] Sourav Chakraborty, Eldar Fischer, Yonatan Goldhirsh, and Arie Matsliah. On the power of conditional samples in distribution testing. *SIAM Journal on Computing*, 45(4):1261–1296, 2016.

[CGR18] Clément L. Canonne, Themis Gouleakis, and Ronitt Rubinfeld. Sampling correctors. *SIAM Journal on Computing*, 47(4):1373–1423, 2018.

[CH11] Kamalika Chaudhuri and Daniel Hsu. Sample complexity bounds for differentially private learning. In *Proceedings of the 24th Annual Conference on Learning Theory*, COLT '11, pages 155–186, 2011.

[Cha05] Sourav Chatterjee. *Concentration Inequalities with Exchangeable Pairs*. PhD thesis, Stanford University, June 2005.

[Chv79] Vasek Chvátal. The tail of the hypergeometric distribution. *Discrete Mathematics*, 25(3):285–287, 1979.

[CL68] C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

[CR14] Clément L. Canonne and Ronitt Rubinfeld. Testing probability distributions underlying aggregated data. In *Proceedings of the 41st International Colloquium on Automata, Languages, and Programming*, ICALP '14, pages 283–295, 2014.

[CRS14] Clément L. Canonne, Dana Ron, and Rocco A. Servedio. Testing equivalence between distributions using conditional samples. In *Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '14, pages 1174–1192, Philadelphia, PA, USA, 2014. SIAM.

[CRS15] Clément L. Canonne, Dana Ron, and Rocco A. Servedio. Testing probability distributions using conditional samples. *SIAM Journal on Computing*, 44(3):540–616, 2015.

[CS10] Madeleine Cule and Richard Samworth. Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electronic Journal of Statistics*, 4:254–270, 2010.

[CSS13] Kamalika Chaudhuri, Anand D. Sarwate, and Kaushik Sinha. A near-optimal algorithm for differentially-private principal components. *Journal of Machine Learning Research*, 14(Sep):2905–2943, 2013.

[CT06a] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2006.

[CT06b] Imre Csiszár and Zsolt Talata. Consistent estimation of the basic neighborhood of Markov random fields. *The Annals of Statistics*, 34(1):123–145, 2006.

[Dal77] Tore Dalenius. Towards a methodology for statistical disclosure control. *Statistisk Tidskrift*, 15:429–444, 1977.

[DBNNR11] Khanh Do Ba, Huy L. Nguyen, Huy N. Nguyen, and Ronitt Rubinfeld. Sublinear time algorithms for earth moverâĂŹs distance. *Theory of Computing Systems*, 48(2):428–442, 2011.

[DDG18] Constantinos Daskalakis, Nishanth Dikkala, and Nick Gravin. Testing symmetric markov chains from a single trajectory. In *Proceedings of the 31st Annual Conference on Learning Theory*, COLT '18, pages 385–409, 2018.

[DDK17] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Concentration of multilinear functions of the Ising model with applications to network data. In *Advances in Neural Information Processing Systems 30*, NIPS '17. Curran Associates, Inc., 2017.

[DDK18] Constantinos Daskalakis, Nishanth Dikkala, and Gautam Kamath. Testing Ising models. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, Philadelphia, PA, USA, 2018. SIAM.

[DDKT16] Constantinos Daskalakis, Anindya De, Gautam Kamath, and Christos Tzamos. A size-free CLT for Poisson multinomials and its applications. In *Proceedings of the 48th Annual ACM Symposium on the Theory of Computing*, STOC '16, pages 1074–1086, New York, NY, USA, 2016. ACM.

[DDS+13] Constantinos Daskalakis, Ilias Diakonikolas, Rocco A. Servedio, Gregory Valiant, and Paul Valiant. Testing k-modal distributions: Optimal algorithms via reductions. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '13, pages 1833–1852, Philadelphia, PA, USA, 2013. SIAM.

[DFH+15] 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, 2015.

[DGJ08] Martin Dyer, Leslie Ann Goldberg, and Mark Jerrum. Dobrushin conditions and systematic scan. *Combinatorics, Probability and Computing*, 17(6):761–779, 2008.

[DGPP16] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Collision-based testers are optimal for uniformity and closeness. *arXiv preprint arXiv:1611.03579*, 2016.

[DGPP18] Ilias Diakonikolas, Themis Gouleakis, John Peebles, and Eric Price. Sample-optimal identity testing with high probability. In *Proceedings of the 45th International Colloquium on Automata, Languages, and Programming*, ICALP '18, pages 41:1–41:14, 2018.

[DHS15] Ilias Diakonikolas, Moritz Hardt, and Ludwig Schmidt. Differentially private learning of structured discrete distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 2566–2574. Curran Associates, Inc., 2015.

[Dif17] Differential Privacy Team, Apple. Learning with privacy at scale. `https://machinelearning.apple.com/docs/learning-with-privacy-at-scale/appledifferentialprivacysystem.pdf`, December 2017.

[DJW17] John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 2017.

[DK14] Constantinos Daskalakis and Gautam Kamath. Faster and sample near-optimal algorithms for proper learning mixtures of Gaussians. In *Proceedings of the 27th Annual Conference on Learning Theory*, COLT '14, pages 1183–1213, 2014.

[DK16] Ilias Diakonikolas and Daniel M. Kane. A new approach for testing properties of discrete distributions. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 685–694, Washington, DC, USA, 2016. IEEE Computer Society.

[DKK+16] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robust estimators in high dimensions without the computational intractability. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 655–664, Washington, DC, USA, 2016. IEEE Computer Society.

[DKK+17] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being robust (in high dimensions) can be practical. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 999–1008. JMLR, Inc., 2017.

[DKK+18a] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Robustly learning a Gaussian: Getting optimal error, efficiently. In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, Philadelphia, PA, USA, 2018. SIAM.

[DKK+18b] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*, 2018.

[DKN15a] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Optimal algorithms and lower bounds for testing closeness of structured distributions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 1183–1202, Washington, DC, USA, 2015. IEEE Computer Society.

[DKN15b] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Testing identity of structured distributions. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '15, pages 1841–1854, Philadelphia, PA, USA, 2015. SIAM.

[DKN17] Ilias Diakonikolas, Daniel M. Kane, and Vladimir Nikishkin. Near-optimal closeness testing of discrete histogram distributions. In *Proceedings of the 44th*

International Colloquium on Automata, Languages, and Programming, ICALP '17, pages 8:1–8:15, 2017.

[DKP18] Ilias Diakonikolas, Daniel M. Kane, and Eric Price. Testing identity of multidimensional histograms. *arXiv preprint arXiv:1804.03636*, 2018.

[DKS16] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Efficient robust proper learning of log-concave distributions. *arXiv preprint arXiv:1606.03077*, 2016.

[DKS17] Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart. Sharp bounds for generalized uniformity testing. *arXiv preprint arXiv:1709.02087*, 2017.

[DKT15] Constantinos Daskalakis, Gautam Kamath, and Christos Tzamos. On the structure, covering, and learning of Poisson multinomial distributions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 1203–1217, Washington, DC, USA, 2015. IEEE Computer Society.

[DKW56] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 09 1956.

[DKW18] Constantinos Daskalakis, Gautam Kamath, and John Wright. Which distribution distances are sublinearly testable? In *Proceedings of the 29th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, pages 2747–2764, Philadelphia, PA, USA, 2018. SIAM.

[DLS+17] Aref N. Dajani, Amy D. Lauger, Phyllis E. Singer, Daniel Kifer, Jerome P. Reiter, Ashwin Machanavajjhala, Simson L. Garfinkel, Scot A. Dahl, Matthew Graham, Vishesh Karwa, Hang Kim, Philip Lelerc, Ian M. Schmutte, William N. Sexton, Lars Vilhuber, and John M. Abowd. The modernization of statistical disclosure limitation at the U.S. census bureau, 2017. Presented at the September 2017 meeting of the Census Scientific Advisory Committee.

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.

[DMR11] Constantinos Daskalakis, Elchanan Mossel, and Sébastien Roch. Evolutionary trees and the Ising model on the Bethe lattice: A proof of Steel's conjecture. *Probability Theory and Related Fields*, 149(1):149–189, 2011.

[DMR18] Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The minimax learning rate of normal and Ising undirected graphical models. *arXiv preprint arXiv:1806.06887*, 2018.

[DN03] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '03, pages 202–210, New York, NY, USA, 2003. ACM.

[Dob56] Roland L. Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory of Probability & Its Applications*, 1(1):65–80, 1956.

[DP17] Constantinos Daskalakis and Qinxuan Pan. Square Hellinger subadditivity for Bayesian networks and its applications to identity testing. In *Proceedings of the 30th Annual Conference on Learning Theory*, COLT '17, pages 697–703, 2017.

[DR96] Devdatt Dubhashi and Desh Ranjan. Balls and bins: A study in negative dependence. *Random Structures and Algorithms*, 13(2):99–124, 1996.

[DR14] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Machine Learning*, 9(3–4):211–407, 2014.

[DTTZ14] Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze Gauss: Optimal bounds for privacy-preserving principal component analysis. In

Proceedings of the 46th Annual ACM Symposium on the Theory of Computing, STOC '14, pages 11–20, New York, NY, USA, 2014. ACM.

[Dwo08] Cynthia Dwork. Differential privacy: A survey of results. In *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation*, TAMC '08, pages 1–19, Berlin, Heidelberg, 2008. Springer.

[Ell93] Glenn Ellison. Learning, local interaction, and coordination. *Econometrica*, 61(5):1047–1071, 1993.

[EPK14] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM Conference on Computer and Communications Security*, CCS '14, pages 1054–1067, New York, NY, USA, 2014. ACM.

[Fel04] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates Sunderland, 2004.

[FGLE12] Steven T. Flammia, David Gross, Yi-Kai Liu, and Jens Eisert. Quantum tomography via compressed sensing: Error bounds, sample complexity and efficient estimators. *New Journal of Physics*, 14(9):095022, 2012.

[Fis25] Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, 1925.

[Fis14] Eldar Fischer. Problem 66: Distinguishing distributions with conditional samples. `https://sublinear.info/index.php?title=Open_Problems:66`, May 2014. Open Problem from Bertinoro Workshop on Sublinear Algorithms 2014.

[FJO+15] Moein Falahatgar, Ashkan Jafarpour, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. Faster algorithms for testing under conditional sampling. In *Proceedings of the 28th Annual Conference on Learning Theory*, COLT '15, pages 607–636, 2015.

[FLNP00] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.

[FLV17] Eldar Fischer, Oded Lachish, and Yadu Vasudev. Improving and extending the testing of distributions for shape-restricted properties. In *Proceedings of the 34th Symposium on Theoretical Aspects of Computer Science*, STACS '17, pages 31:1–31:14, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[FMO18] Orr Fischer, Uri Meir, and Rotem Oshman. Distributed uniformity testing. In *Proceedings of the 2018 ACM Symposium on Principles of Distributed Computing*, PODC '18, pages 455–464, New York, NY, USA, 2018. ACM.

[FOS08] Jon Feldman, Ryan O'Donnell, and Rocco A. Servedio. Learning mixtures of product distributions over discrete domains. *SIAM Journal on Computing*, 37(5):1536–1564, 2008.

[FX15] Vitaly Feldman and David Xiao. Sample complexity bounds on differentially private learning via communication complexity. *SIAM Journal on Computing*, 44(6):1740–1764, 2015.

[Geo11] Hans-Otto Georgii. *Gibbs Measures and Phase Transitions*. Walter de Gruyter, 2011.

[GG86] Stuart Geman and Christine Graffigne. Markov random field image models and their applications to computer vision. In *Proceedings of the International Congress of Mathematicians*, pages 1496–1517. American Mathematical Society, 1986.

[GGB18] Alon Gonem and Ram Gilad-Bachrach. Smooth sensitivity based approach for differentially private PCA. In *Algorithmic Learning Theory*, ALT '18, pages 438–450. JMLR, Inc., 2018.

[GGR96] Oded Goldreich, Shafi Goldwasser, and Dana Ron. Property testing and its connection to learning and approximation. In *Proceedings of the 37th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '96, pages 339–348, Washington, DC, USA, 1996. IEEE Computer Society.

[GLP17] Reza Gheissari, Eyal Lubetzky, and Yuval Peres. Concentration inequalities for polynomials of contracting Ising models. *arXiv preprint arXiv:1706.00121*, 2017.

[GLRV16] Marco Gaboardi, Hyun-Woo Lim, Ryan M. Rogers, and Salil P. Vadhan. Differentially private chi-squared hypothesis testing: Goodness of fit and independence testing. In *Proceedings of the 33rd International Conference on Machine Learning*, ICML '16, pages 1395–1403. JMLR, Inc., 2016.

[GM18] Anna Gilbert and Audra McMillan. Property testing for differential privacy. In *56th Annual Allerton Conference on Communication, Control, and Computing*, Allerton '18, Washington, DC, USA, 2018. IEEE Computer Society.

[GMV06] Sudipto Guha, Andrew McGregor, and Suresh Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '06, pages 733–742, Philadelphia, PA, USA, 2006. SIAM.

[GNS17] Aditya Gangrade, Bobak Nazer, and Venkatesh Saligrama. Lower bounds for two-sample structural change detection in Ising and Gaussian models. In *55th Annual Allerton Conference on Communication, Control, and Computing*, Allerton '17, pages 1016–1025, Washington, DC, USA, 2017. IEEE Computer Society.

[GNW17] David Gross, Sepehr Nezami, and Michael Walter. Schur-Weyl duality for the Clifford group with applications: Property testing, a robust Hudson theorem, and de Finetti representations. *arXiv preprint arXiv:1712.08628*, 2017.

[Gol14] Oded Goldreich. On multiple input problems in property testing. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques.*, RANDOM '14, pages 704–720, Dagstuhl, Germany, 2014. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[Gol16] Oded Goldreich. The uniform distribution is complete with respect to testing identity to a fixed distribution. *Electronic Colloquium on Computational Complexity (ECCC)*, 23(15), 2016.

[Gol17] Oded Goldreich. *Introduction to Property Testing*. Cambridge University Press, 2017.

[GR00] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. *Electronic Colloquium on Computational Complexity (ECCC)*, 7(20), 2000.

[GR18] Marco Gaboardi and Ryan Rogers. Local private hypothesis testing: Chi-square tests. In *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pages 1626–1635. JMLR, Inc., 2018.

[GRT18] Sumegha Garg, Ran Raz, and Avishay Tal. Testing conditional independence of discrete distributions. In *Proceedings of the 50th Annual ACM Symposium on the Theory of Computing*, STOC '18, pages 990–1002, New York, NY, USA, 2018. ACM.

[GS02] Alison L. Gibbs and Francis E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, dec 2002.

[GSS18] Friedrich Götze, Holger Sambale, and Arthur Sinulis. Higher order concentration for functions of weakly dependent random variables. *arXiv preprint arXiv:1801.06348*, 2018.

[GT56] I.J. Good and G.H. Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

[GTZ17] Themistoklis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Faster sublinear algorithms using conditional sampling. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, pages 1743–1757, Philadelphia, PA, USA, 2017. SIAM.

[GTZ18] Themis Gouleakis, Christos Tzamos, and Manolis Zampetakis. Certified computation from unreliable datasets. In *Proceedings of the 31st Annual Conference on Learning Theory*, COLT '18, pages 3271–3294, 2018.

[Han14] Steve Hanneke. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2–3):131–309, 2014.

[Hay06] Thomas P. Hayes. A simple condition implying rapid mixing of single-site dynamics on spin systems. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '06, pages 39–46, Washington, DC, USA, 2006. IEEE Computer Society.

[HHJ+17] Jeongwan Haah, Aram W. Harrow, Zhengfeng Ji, Xiaodi Wu, and Nengkun Yu. Sample-optimal tomography of quantum states. *IEEE Transactions on Information Theory*, 63(9):5628–5641, 2017.

[HJW16] Yanjun Han, Jiao Jiantao, and Tsachy Weissman. Minimax rate-optimal estimation of divergences between discrete distributions. *arXiv preprint arXiv:1605.09124*, 2016.

[HKKT18] Steve Hanneke, Adam Kalai, Gautam Kamath, and Christos Tzamos. Actively avoiding nonsense in generative models. In *Proceedings of the 31st Annual Conference on Learning Theory*, COLT '18, pages 209–227, 2018.

[HKL+17] Daniel Hsu, Aryeh Kontorovich, David A. Levin, Yuval Peres, and Csaba Szepesvári. Mixing time estimation in reversible Markov chains from a single sample path. *arXiv preprint arXiv:1708.07367*, 2017.

[HKM17] Linus Hamilton, Frederic Koehler, and Ankur Moitra. Information theoretic properties of Markov random fields, and their algorithmic applications. In

Advances in Neural Information Processing Systems 30, NIPS '17. Curran Associates, Inc., 2017.

[HKS15] Daniel Hsu, Aryeh Kontorovich, and Csaba Szepesvári. Mixing time estimation in reversible Markov chains from a single sample path. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 1459–1467. Curran Associates, Inc., 2015.

[HLM17] Aram W. Harrow, Cedric Yen-Yu Lin, and Ashley Montanaro. Sequential measurements, disturbance, and property testing. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, pages 1598–1611, Philadelphia, PA, USA, 2017. SIAM.

[HM13a] Aram W. Harrow and Ashley Montanaro. Testing product states, quanum Merlin-Arthur games and tensor optimization. *Journal of the ACM*, 60(1):3:1–3:43, 2013.

[HM13b] Dayu Huang and Sean Meyn. Generalized error exponents for small sample universal hypothesis testing. *IEEE Transactions on Information Theory*, 59(12):8157–8181, 2013.

[HP14] Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. In *Advances in Neural Information Processing Systems 27*, NIPS '14, pages 2861–2869. Curran Associates, Inc., 2014.

[HSR+08] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8):1–9, 2008.

[HVK05] Peter Hall and Ingrid Van Keilegom. Testing for monotone increasing hazard rate. *The Annals of Statistics*, 33(3):1109–1137, 2005.

[ILR12] Piotr Indyk, Reut Levi, and Ronitt Rubinfeld. Approximating and testing k-histogram distributions in sub-linear time. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '12, pages 15–22, New York, NY, USA, 2012. ACM.

[Ing94] Yuri Izmailovich Ingster. Minimax detection of a signal in ℓ_p metrics. *Journal of Mathematical Sciences*, 68(4):503–515, 1994.

[Ing97] Yuri Izmailovich Ingster. Adaptive chi-square tests. *Zapiski Nauchnykh Seminarov POMI*, 244:150–166, 1997.

[IS03] Yuri Izmailovich Ingster and Irina A. Suslina. *Nonparametric Goodness-of-fit Testing Under Gaussian Models*, volume 169 of *Lecture Notes in Statistics*. Springer, 2003.

[Isi25] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik A Hadrons and Nuclei*, 31(1):253–258, 1925.

[JHW16] Jiantao Jiao, Yanjun Han, and Tsachy Weissman. Minimax estimation of the ℓ_1 distance. In *Proceedings of the 2016 IEEE International Symposium on Information Theory*, ISIT '16, pages 750–754, Washington, DC, USA, 2016. IEEE Computer Society.

[JJR11] Ali Jalali, Christopher C. Johnson, and Pradeep K. Ravikumar. On learning discrete graphical models using greedy methods. In *Advances in Neural Information Processing Systems 24*, NIPS '11, pages 1935–1943. Curran Associates, Inc., 2011.

[Jor10] Michael Jordan. Lecture notes for Bayesian modeling and inference, 2010.

[JS93] Mark Jerrum and Alistair Sinclair. Polynomial-time approximation algorithms for the Ising model. *SIAM Journal on Computing*, 22(5):1087–1116, 1993.

[JS13] Aaron Johnson and Vitaly Shmatikov. Privacy-preserving data exploration in genome-wide association studies. In *Proceedings of the 19th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, KDD '13, pages 1079–1087, New York, NY, USA, 2013. ACM.

[JVHW17] Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2017.

[JW09] Hanna K. Jankowski and Jon A. Wellner. Estimation of a discrete monotone density. *Electronic Journal of Statistics*, 3:1567–1605, 2009.

[KBR16] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *Proceedings of the 33rd International Conference on Machine Learning*, ICML '16, pages 2436–2444. JMLR, Inc., 2016.

[Kla00] Bernhard Klar. Bounds on tail probabilities of discrete distributions. *Probability in the Engineering and Informational Sciences*, 14(02):161–171, 2000.

[KLN+11] Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

[KLSU18] Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. *arXiv preprint arXiv:1805.00216*, 2018.

[KM17] Adam Klivans and Raghu Meka. Learning graphical models using multiplicative weights. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '17, pages 343–354, Washington, DC, USA, 2017. IEEE Computer Society.

[KNS07] Jeongwoo Ko, Eric Nyberg, and Luo Si. A probabilistic graphical model for joint answer ranking in question answering. In *Proceedings of the 30th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '07, pages 343–350, New York, NY, USA, 2007. ACM.

[KOPS15] Sudeep Kamath, Alon Orlitsky, Venkatadheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Proceedings of the 28th Annual Conference on Learning Theory*, COLT '15, pages 1066–1100, 2015.

[KR17] Daniel Kifer and Ryan M. Rogers. A new class of private chi-square tests. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, AISTATS '17, pages 991–1000. JMLR, Inc., 2017.

[KRT17a] Gillat Kol, Ran Raz, and Avishay Tal. Time-space hardness of learning sparse parities. In *Proceedings of the 49th Annual ACM Symposium on the Theory of Computing*, STOC '17, pages 1067–1080, New York, NY, USA, 2017. ACM.

[KRT17b] Richard Kueng, Holger Rauhut, and Ulrich Terstiege. Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis*, 42(1):88–116, 2017.

[KS16] Vishesh Karwa and Aleksandra Slavković. Inference using noisy degrees: Differentially private β-model and synthetic graphs. *The Annals of Statistics*, 44(1):87–112, 2016.

[KSF17] Kazuya Kakizaki, Jun Sakuma, and Kazuto Fukuchi. Differentially private chi-squared test by unit circle mechanism. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 1761–1770. JMLR, Inc., 2017.

[KT13] Michael Kapralov and Kunal Talwar. On differentially private low rank approximation. In *Proceedings of the 24th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '13, pages 1395–1414, Philadelphia, PA, USA, 2013. SIAM.

[KT18] Gautam Kamath and Christos Tzamos. Anaconda: A non-adaptive conditional sampling algorithm for distribution testing. *arXiv preprint arXiv:1807.06168*, 2018.

[KV18] Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pages 44:1–44:9, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[KXS+16] Albert Kim, Liqi Xu, Tarique Siddiqui, Silu Huang, Samuel Madden, and Aditya Parameswaran. Optimally leveraging density and locality to support LIMIT queries. *arXiv preprint arXiv:1611.04705*, 2016.

[LAFH01] Charles Lagor, Dominik Aronsky, Marcelo Fiszman, and Peter J. Haug. Automatic identification of patients eligible for a pneumonia guideline: comparing the diagnostic accuracy of two decision support models. *Studies in Health Technology and Informatics*, 84(1):493–497, 2001.

[Lap78] Pierre-Simon Laplace. Mémoire sur les probabilités. *Mémoirs de l'Académie royale des Sciences de Paris*, 1778.

[LC73] Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

[LC06] Erich Leo Lehmann and George Casella. *Theory of Point Estimation*. Springer, 2006.

[LP16] David A. Levin and Yuval Peres. Estimating the spectral gap of a reversible markov chain from a short trajectory. *arXiv preprint arXiv:1612.05330*, 2016.

[LPW09] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, 2009.

[LR05] Erich Leo Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, 2005.

[LRR13] Reut Levi, Dana Ron, and Ronitt Rubinfeld. Testing properties of collections of distributions. *Theory of Computing*, 9(8):295–347, 2013.

[LW17] Tongyang Li and Xiaodi Wu. Quantum query complexity of entropy estimation. *arXiv preprint arXiv:1710.06025*, 2017.

[Mas90] P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 07 1990.

[MdCCU16] Abraham Martín del Campo, Sarah Cepeda, and Caroline Uhler. Exact goodness-of-fit testing for the Ising model. *Scandinavian Journal of Statistics*, 2016.

[MdW16] Ashley Montanaro and Ronald de Wolf. A survey of quantum property testing. *Theory of Computing Library, Graduate Surveys*, 7:1–81, 2016.

[Mil55] George A. Miller. Note on the bias of information estimates. *Information Theory in Psychology: Problems and Methods*, 2:95–100, 1955.

[MM17] Dana Moshkovitz and Michal Moshkovitz. Mixing implies lower bounds for space bounded learning. In *Proceedings of the 30th Annual Conference on Learning Theory*, COLT '17, pages 1516–1566, 2017.

[MM18] Dana Moshkovitz and Michal Moshkovitz. Entropy samplers and strong generic lower bounds for space bounded learning. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pages 28:1–28:20, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[MR95] Rajeev Motwani and Prabhakar Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[MS10] Andrea Montanari and Amin Saberi. The spread of innovations in social networks. *Proceedings of the National Academy of Sciences*, 107(47):20196–20201, 2010.

[Mut05] Shanmugavelayutham Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends® in Machine Learning*, 1(2):117–236, 2005.

[NBdRvS04] Ilya Nemenman, William Bialek, and Rob de Ruyter van Steveninck. Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, 69(5):056111:1–056111:6, 2004.

[Now12] Sebastian Nowozin. Improved information gain estimates for decision tree induction. In *Proceedings of the 29th International Conference on Machine Learning*, ICML '12, pages 571–578. JMLR, Inc., 2012.

[NS18] Kobbi Nissim and Uri Stemmer. Clustering algorithms for the centralized and local models. In *Algorithmic Learning Theory*, ALT '18, pages 619–653. JMLR, Inc., 2018.

[OS15] Alon Orlitsky and Ananda Theertha Suresh. Competitive distribution estimation: Why is Good-Turing good. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 2143–2151. Curran Associates, Inc., 2015.

[OS17] Maciej Obremski and Maciej Skorski. Rényi entropy estimation revisited. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques.*, APPROX '17, pages 20:1–20:15, Dagstuhl, Germany, 2017. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[OS18] Krzysztof Onak and Xiaorui Sun. Probability-revealing samples. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, AISTATS '18, pages 84:2018–2026. JMLR, Inc., 2018.

[OSW16] Alon Orlitsky, Ananda Theerta Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016.

[OW15] Ryan O'Donnell and John Wright. Quantum spectrum testing. In *Proceedings of the 47th Annual ACM Symposium on the Theory of Computing*, STOC '15, pages 529–538, New York, NY, USA, 2015. ACM.

[OW16a] Ryan O'Donnell and John Wright. Efficient quantum tomography. In *Proceedings of the 48th Annual ACM Symposium on the Theory of Computing*, STOC '16, pages 899–912, New York, NY, USA, 2016. ACM.

[OW16b] Ryan O'Donnell and John Wright. Guest column: A primer on the statistics of longest increasing subsequences and quantum states (shortened version). *ACM SIGACT News*, 48(3):37–59, 2016.

[OW17] Ryan O'Donnell and John Wright. Efficient quantum tomography ii. In *Proceedings of the 49th Annual ACM Symposium on the Theory of Computing*, STOC '17, pages 962–974, New York, NY, USA, 2017. ACM.

[OZ18] Ryan O'Donnell and Yu Zhao. On closeness to k-wise uniformity. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques.*, RANDOM '18, pages 54:1–54:19, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[Pan03] Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, 2003.

[Pan08] Liam Paninski. A coincidence-based test for uniformity given very sparsely sampled discrete data. *IEEE Transactions on Information Theory*, 54(10):4750–4755, 2008.

[Pea00] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.

[PLM18] Sam Pallister, Noah Linden, and Ashley Montanaro. Optimal verification of entangled states with local measurements. *Physical Review Letters*, 120(17):170502, 2018.

[Pol15] David Pollard. A few good inequalities. `http://www.stat.yale.edu/~pollard/Books/Mini/Basic.pdf`, 2015.

[PRR06] Michal Parnas, Dana Ron, and Ronitt Rubinfeld. Tolerant property testing and distance approximation. *Journal of Computer and System Sciences*, 72(6):1012–1042, 2006.

[RAS15] Firas Rassoul-Agha and Timo Seppäläinen. *A Course on Large Deviations with an Introduction to Gibbs Measures*. American Mathematical Society, 2015.

[Raz16] Ran Raz. Fast learning requires good memory: A time-space lower bound for parity learning. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 266–275, Washington, DC, USA, 2016. IEEE Computer Society.

[Raz17] Ran Raz. A time-space lower bound for a large class of learning problems. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '17, pages 732–742, Washington, DC, USA, 2017. IEEE Computer Society.

[Rog17] Ryan Michael Rogers. *Leveraging Privacy in Data Analysis*. PhD thesis, University of Pennsylvania, May 2017.

[RRSS09] Sofya Raskhodnikova, Dana Ron, Amir Shpilka, and Adam Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

[RRST16] Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 487–494, Washington, DC, USA, 2016. IEEE Computer Society.

[RS81] Jon N.K. Rao and Alastair J. Scott. The analysis of categorical data from complex sample surveys: Chi-squared tests for goodness of fit and independence in two-way tables. *Journal of the Americal Statistical Association*, 76(374):221–230, 1981.

[RS09] Ronitt Rubinfeld and Rocco A. Servedio. Testing monotone high-dimensional distributions. *Random Structures and Algorithms*, 34(1):24–44, 2009.

[RT16] Dana Ron and Gilad Tsur. The power of an example: Hidden set size approximation using group queries and conditional sampling. *Transactions on Computation Theory*, 8(4):15:1 – 15:19, 2016.

[Rub12] Ronitt Rubinfeld. Taming big probability distributions. *XRDS*, 19(1):24–28, 2012.

[RWL10] Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[RX14] Ronitt Rubinfeld and Ning Xie. Testing non-uniform k-wise independent distributions over product spaces. In *Proceedings of the 37th International Colloquium on Automata, Languages, and Programming*, ICALP '10, pages 565–581, 2014.

[Set12] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1 – 114, 2012.

[She17] Or Sheffet. Differentially private ordinary least squares. In *Proceedings of the 34th International Conference on Machine Learning*, ICML '17, pages 3105–3114. JMLR, Inc., 2017.

[She18] Or Sheffet. Locally private hypothesis testing. In *Proceedings of the 35th International Conference on Machine Learning*, ICML '18, pages 4605–4614. JMLR, Inc., 2018.

[Smi11] Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing*, STOC '11, pages 813–822, New York, NY, USA, 2011. ACM.

[SS02] Michael Saks and Xiaodong Sun. Space lower bounds for distance approximation in the data stream model. In *Proceedings of the 34th Annual ACM Symposium on the Theory of Computing*, STOC '02, pages 360–369, New York, NY, USA, 2002. ACM.

[SSB16] Sean Simmons, Cenk Sahinalp, and Bonnie Berger. Enabling privacy-preserving GWASs in heterogeneous human populations. *Cell Systems*, 3(1):54–61, 2016.

[SSJ17] Imdad S. B. Sardharwalla, Sergii Strelchuk, and Richard Jozsa. Quantum conditional query complexity. *Quantum Information & Computation*, 17(7& 8):541–566, 2017.

[Sto85] Larry Stockmeyer. On approximation algorithms for #P. *SIAM Journal on Computing*, 14(4):849–861, 1985.

[STW10] Sujay Sanghavi, Vincent Tan, and Alan Willsky. Learning graphical models for hypothesis testing and classification. *IEEE Transactions on Signal Processing*, 58(11):5481–5495, 2010.

[SW12] Narayana P. Santhanam and Martin J. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012.

[SW14] Adrien Saumard and Jon A. Wellner. Log-concavity and strong log-concavity: A review. *Statistics Surveys*, 8:45–114, 2014.

[Tib96] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

[USF13] Caroline Uhler, Aleksandra Slavković, and Stephen E. Fienberg. Privacy-preserving data sharing for genome-wide association studies. *The Journal of Privacy and Confidentiality*, 5(1):137–166, 2013.

[Vad17] Salil Vadhan. The complexity of differential privacy. In Yehuda Lindell, editor, *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*, chapter 7, pages 347–450. Springer International Publishing AG, Cham, Switzerland, 2017.

[Val11] Paul Valiant. Testing symmetric properties of distributions. *SIAM Journal on Computing*, 40(6):1927–1968, 2011.

[VMLC16] Marc Vuffray, Sidhant Misra, Andrey Lokhov, and Michael Chertkov. Interaction screening: Efficient and sample-optimal learning of Ising models. In *Advances in Neural Information Processing Systems 29*, NIPS '16, pages 2595–2603. Curran Associates, Inc., 2016.

[VS09] Duy Vu and Aleksandra Slavković. Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, pages 138–143. IEEE, 2009.

[VV10a] Gregory Valiant and Paul Valiant. A CLT and tight lower bounds for estimating entropy. *Electronic Colloquium on Computational Complexity (ECCC)*, 17(179), 2010.

[VV10b] Gregory Valiant and Paul Valiant. Estimating the unseen: A sublinear-sample canonical estimator of distributions. *Electronic Colloquium on Computational Complexity (ECCC)*, 17(180), 2010.

[VV11a] Gregory Valiant and Paul Valiant. Estimating the unseen: An $n/\log n$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing*, STOC '11, pages 685–694, New York, NY, USA, 2011. ACM.

[VV11b] Gregory Valiant and Paul Valiant. The power of linear estimators. In *Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science*, FOCS '11, pages 403–412, Washington, DC, USA, 2011. IEEE Computer Society.

[VV13] Gregory Valiant and Paul Valiant. Estimating the unseen: Improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems 26*, NIPS '13, pages 2157–2165. Curran Associates, Inc., 2013.

[VV14] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 51–60, Washington, DC, USA, 2014. IEEE Computer Society.

[VV15] Gregory Valiant and Paul Valiant. Instance optimal learning. In *Proceedings of the 47th Annual ACM Symposium on the Theory of Computing*, STOC '15, pages 142–155, New York, NY, USA, 2015. ACM.

[VV17a] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

[VV17b] Gregory Valiant and Paul Valiant. Estimating the unseen: Improved estimators for entropy and other properties. *Journal of the ACM*, 64(6):37:1–37:41, 2017.

[Wag15] Bo Waggoner. l_p testing and learning of discrete distributions. In *Proceedings of the 6th Conference on Innovations in Theoretical Computer Science*, ITCS '15, pages 347–356, New York, NY, USA, 2015. ACM.

[WHW+16] Shaowei Wang, Liusheng Huang, Pengzhan Wang, Yiwen Nie, Hongli Xu, Wei Yang, Xiang-Yang Li, and Chunming Qiao. Mutual information optimally local private discrete distribution estimation. *arXiv preprint arXiv:1607.08025*, 2016.

[WLK15] Yue Wang, Jaewoo Lee, and Daniel Kifer. Revisiting differentially private hypothesis tests for categorical data. *arXiv preprint arXiv:1511.03376*, 2015.

[Wri16] John Wright. *How to Learn a Quantum State*. PhD thesis, Carnegie Mellon University, May 2016.

[WWS15] Yining Wang, Yu-Xiang Wang, and Aarti Singh. Differentially private subspace clustering. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 1000–1008. Curran Associates, Inc., 2015.

[WY16] Yihong Wu and Pengkun Yang. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

[WY18] Yihong Wu and Pengkun Yang. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 2018.

[WZ10] Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

[YB18] Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64(8):5662–5676, 2018.

[YFSU14] Fei Yu, Stephen E. Fienberg, Aleksandra B. Slavković, and Caroline Uhler. Scalable privacy-preserving data sharing methodology for genome-wide association studies. *Journal of Biomedical Informatics*, 50:133–141, 2014.

[Zel87] Daniel Zelterman. Goodness-of-fit tests for large sparse multinomial distributions. *Journal of the American Statistical Association*, 82(398):624–629, 1987.