TESTING PROPERTIES OF DISTRIBUTIONS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Tugkan Batu

August 2001

TESTING PROPERTIES OF DISTRIBUTIONS

Tugkan Batu, Ph.D.

Cornell University 2001

We study the sample complexity of several basic statistical inference tasks as a function of the domain size for the underlying discrete probability distributions.

Given access only to samples from two distributions over an n-element set, we want to distinguish identical pairs of distributions from pairs of distributions that have large statistical distance. We give an algorithm that uses $O(n^{2/3} \log n)$ independent samples from each distribution, runs in time linear in the sample size, makes no assumptions about the structure of the distributions, and distinguishes the case that the statistical distance between the distributions is small from the case that it is large. We also prove a lower bound of $\Omega(n^{2/3})$ for the sample complexity.

Under a related model, we show how to test, given access to samples from a distribution over an n-element set, whether the distribution is statistically close to an explicitly specified distribution. Our test uses $\tilde{O}(n^{1/2})$ samples, which matches the known tight bounds for the case when the explicit distribution is uniform.

Given access to independent samples of a distribution $\mathbf{A}$ over the product space of two sets with n and m elements, respectively, we show how to test whether

the distributions induced by $\mathbf{A}$ restricted to each component are independent, i.e., whether $\mathbf{A}$ is statistically close to $\mathbf{A}_1 \times \mathbf{A}_2$ for some $\mathbf{A}_1$ over an n-element set and $\mathbf{A}_2$ over an m-element set. The sample complexity of our test is $\tilde{O}(n^{2/3}m^{1/3})$, assuming without loss of generality that $m \leq n$. We also give a matching lower bound up to polylogarithmic factors.

We consider the problem of approximating the entropy of a black-box discrete distribution in sublinear time. We show that a γ-multiplicative approximation to the entropy can be obtained in $\tilde{O}(n^{(1+\zeta)/\gamma^2})$ time for distributions with sufficiently high entropy where n is the size of the domain of the distribution and ζ is an arbitrarily small positive constant. We show that one cannot get a multiplicative approximation to the entropy in general. Even for the class of distributions to which our upper bound applies, we show a lower bound of $\Omega(n^{1/(2\gamma^2)})$.

Biographical Sketch

Tuğkan Batu made his first debut on May 23rd, 1974 in Ankara, Turkey. After a happy childhood, a carefree adolescence, two schools, three generations of computers, many good friends, and a short and not-so-bright volleyball career, he decided to settle on computers. He completed his B.S. degree in the Department of Computer Engineering and Information Science at Bilkent University in 1996. He then relocated to gorge(ou)s Ithaca, and was introduced to ice hockey to which he immediately got addicted. Despite his heavy hockey schedule all year around, he got his M.S. degree in Computer Science at Cornell in May 2000. After completing his Ph.D. studies, he is joining the Department of Computer and Information Science at University of Pennsylvania as a postdoctoral researcher.

To my family and Balam

Acknowledgments

Although it is only my name that appears on the title page, there are others who contributed greatly to the making of this dissertation. I am grateful to my advisor Ronitt Rubinfeld in so many ways that I cannot possibly express. While she was a great advisor and educator, Ronitt never made me feel less than a colleague. The pep talks were inspritational, the criticism useful. The effort she put into my writing and presentation skills was enormous. I still miss a good fraction of the "the"s though.

I have also learned a lot from my other collaborators Sanjoy Dasgupta, Lance Fortnow, Eldar Fischer, Ravi Kumar, Warren Smith, Patrick White. The work described in this dissertation is based on various joint works with them.

Patrick is much more than a mere co-author to me. He is a very good friend. In addition to being an invaluable mathematical reference, he painstakingly attended cultural, linguistic, and idiosyncratic questions of mine. He was very much fun to share with, whether it was boredom, a hotel room, a meal, or a drink that we shared.

The work presented in this dissertation was partially supported by ONR grant N00014–97–1–0505.

I cannot overlook many beautiful people who made my graduate school experience a blissful one. Any list of names that I come up with would undoubtedly be

incomplete, so I would like to thank all those who remain unnamed here—students I met at Cornell, the faculty and staff.

But above all, I owe much to my family. Their continual support and love, and the appreciation of knowledge that they planted in me made me what I am today. If there is anything good in me, that is because they put it there.

My best discovery at Cornell is not mentioned on these pages. Balam enchanted me with the gift of love. The beauty that she ingrained in me will be there forever. Thank you my love.

Table of Contents

List of Figures

Chapter 1

Introduction

The objective of a typical statistical application is to make inferences about a population using samples taken from the population. Such inferences might be used in a wide variety of ways, including to infer causalities, or to base strategies on. The types of inferences can include determining whether two populations have similar characteristics, whether certain events tend to occur together, or whether the occurrence of an event is distributed according to a specific kind of distribution. One such inference might be to unveil any correlation between the zip code area that people live in and their health conditions. Another such inference might be to validate a large-scale plant experiment on pesticide combinations by comparing the results against theoretical predictions.

Tests such as χ^2 test have been proposed for such central questions in statistics. Unfortunately, these tests seem to require a lot of samples. As the tasks become more complicated, more samples are needed to discover the patterns that are revealing. Samples are often expensive to obtain. As a consequence, in many statistical applications, the lack of a sufficient number of samples from the population is a perennial problem. The shortage of samples prevents one from reaching a

conclusion that one is confident about.

In order to reduce the required number of samples while keeping the results relevant, one might try to simplify the inference task. One approach for such a simplification is to make assumptions about the population, or equivalently, about the underlying probability distributions. For example, one can assume that the samples come from a normal distribution. Although such an assumption can make the task easier, it may not always be valid to make a simplifying assumption. A second way to simplify the inference task is most applicable when the distributions have large domains. The required number of samples in a statistical application depends on the size of the domain over which the distribution under study is defined. Thus, reducing the domain size translates immediately into a reduction in the required number of samples. A common approach to obtaining such a reduction is by coalescing domain elements of the distribution into categories that constitute a smaller domain. Note that in some cases, there may not be a justifiably good partition of the domain with which to do the coalescing. Even in the cases where the domain elements could legitimately be coalesced, this grouping introduces additional error.

We study several basic statistical tasks in terms of the required number of samples. Each task we study has a straight-forward algorithm that is based on the following idea: after taking enough samples from a distribution, the frequency counts for the elements with sufficiently high probability give accurate estimates of the respective probability values when normalized by the total number of samples. For each task we consider, it is not hard to show that these estimates allow one to reach reliable conclusions. Unfortunately, this approach requires at least linear (in the domain size) number of samples.

We give nearly tight characterizations of sample complexity for these tasks as a

function of the domain size. The algorithms we present have sublinear sample complexity. Although the straight-forward algorithms do not yield such performances, we use the approach of the straight-forward algorithm to handle the high-probability events of the distributions. We then give specially tailored algorithms that perform the tasks efficiently on events that are of low probability.

We make no assumptions about the distributions. We assume that our algorithms only get samples from the distributions as input. We analyze the number of samples required by the algorithms as a function of the domain size. We formalize this model of access by giving the algorithm an oracle that, upon request, outputs a sample distributed according to the distribution. We call such an oracle a black-box distribution.

In this dissertation, we study the following four properties of discrete distributions.

Testing Closeness of Distributions A marketing survey can use data on the shopping habits of people living in two different geographic areas to tailor advertising strategies. For example, if samples of shopping data indicate that shopping habits from the two areas are similar, then the same advertising strategy might be used for these areas.

We abstract the task of detecting similarities of two data-generating sources by the problem of testing closeness of distributions. An algorithm for testing closeness of two distributions has access to two black-box distributions. It distinguishes identical pairs of distributions from pairs of distributions that have large statistical distance. For pairs with small statistical distance, no performance guarantee is made.

In Chapter 3, we present an algorithm that requires $\tilde{O}(n^{2/3})$ samples for testing

closeness of two distributions over an n-element set and prove a lower bound of $\Omega(n^{2/3})$ samples.

Testing Identity of a Distribution Consider the task of checking whether a lottery is fair. In a fair lottery, each number has an equal chance of winning, that is, the winning number is chosen from a uniform distribution. The history of winning numbers in a lottery can be used to test if the lottery is fair.

We capture the verification problem of the source of given randomly-generated data as testing identity of a distribution. Similar to the previous problem of testing closeness of distributions, an algorithm for testing identity of a distribution also takes two distributions as input and distinguishes identical pairs of distributions from pairs of distributions that have large statistical distance. However, the algorithm is provided with the description of one of the distributions in this case. We call an oracle an explicit distribution if, upon being given the name of a domain element, it outputs the probability value assigned to that element. An algorithm for testing identity of a distribution has access to one black-box distribution and one explicit distribution.

In Chapter 3, we present an algorithm that requires $\tilde{O}(\sqrt{n})$ samples for testing identity of a distribution over an n-element set and prove a lower bound of $\Omega(\sqrt{n})$ samples.

It is interesting to note that when one of the distributions is given as an explicit distribution instead of a black-box distribution, the problem of testing closeness becomes provably easier.

Testing Independence of a Distribution In the evaluation of the results of a medical experiment, one can study the correlation of the genetic makeup of an

individual and the outcome of the treatment for different doses of the tested drug.

The task of detecting correlations in the data can be viewed as testing independence of joint distributions. We define the problem of testing independence of distributions on joint black-box distributions over a set of pairs. A joint distribution is called independent if for all possible a and b, the probability of sampling the pair (a, b) is equal to the product of the probability of sampling a pair with a in the first coordinate and the probability of sampling a pair with b in the second coordinate.

An algorithm for testing independence of a joint black-box distribution distinguishes independent joint distributions from distributions with large statistical distance to all independent joint distributions.

In Chapter 4, we present an algorithm that requires $\tilde{O}(n^{2/3}m^{1/3})$ samples for testing independence of a joint distribution over the product space of two sets with n and m elements, respectively, for $n \geq m$ and prove a lower bound of $\Omega(n^{2/3}m^{1/3})$ samples.

Approximating the Entropy of a Distribution The Shannon entropy is a measure of randomness of a distribution. The notion of entropy plays a central role in statistics, physics, information theory, and data compression and has been studied extensively. For example, knowing the entropy of a random source can shed light on the compressibility of data produced by such a source.

In Chapter 5, we present an algorithm that can approximate the entropy of a black-box distribution over an n-element set within a multiplicative factor of $\gamma > 1$ using $O(n^{(1+\zeta)/\gamma^2})$ samples for arbitrarily small $\zeta > 0$, provided that the entropy value is sufficiently high. Such a restriction on the class of distributions of which the entropy can be approximated is justified once we show that one cannot get a multiplicative approximation to the entropy in general. We show lower bounds on

the sample complexity of approximating entropy even for the class of distributions to which our algorithm applies.

Chapter 2

Preliminaries

For any natural number n, we denote the set $\{1, \ldots, n\}$ by $[n]$. A discrete probability distribution $\mathbf{p}$ over $[n]$ is identified with the vector notation $\mathbf{p} = (p_1, \ldots, p_n)$.

The notation $x \in_R [n]$ denotes that x is chosen uniformly at random from the set $[n]$. The L_1 norm of a vector $\mathbf{v}$ is denoted by $|\mathbf{v}|$ and is equal to $\sum_{i=1}^{n} |v_i|$. Similarly, the L_2 norm is denoted by $\|\mathbf{v}\|$ and is equal to $\sqrt{\sum_{i=1}^{n} v_i^2}$, and $\|\mathbf{v}\|_\infty = \max_i |v_i|$. If $|\mathbf{p} - \mathbf{q}| \leq \epsilon$, we say that $\mathbf{p}$ is **ϵ-close** to $\mathbf{q}$.

The **collision probability** of two distributions $\mathbf{p}$ and $\mathbf{q}$ over the set R is the probability that a sample from each of $\mathbf{p}$ and $\mathbf{q}$ yields the same element. Note that for two distributions $\mathbf{p}, \mathbf{q}$ over R, the collision probability is $\mathbf{p} \cdot \mathbf{q} = \sum_{i \in R} p_i q_i$. To avoid ambiguity, we refer to the collision probability of $\mathbf{p}$ with itself as the **self-collision probability** of $\mathbf{p}$. Note that the self-collision probability of $\mathbf{p}$ is $\|\mathbf{p}\|^2$.

We use the $\tilde{O}$ notation to hide dependencies on the logarithm of any of the quantities in the expression, i.e., $f = \tilde{O}(g)$ if $f = O(g\text{poly}(\log g))$. To simplify the exposition, we assume all tests are repeated so that the confidence is sufficiently high. Since a confidence of $1 - \delta$ can be achieved with $O(\log \frac{1}{\delta})$ trials, an additional multiplicative factor of $O(\text{poly}(\log n))$ is all that is required.

For a set R, let U_R denote the uniform distribution over R. For a distribution $\mathbf{p}$ over the set R and a subset R' of R, let $\mathbf{p}(R') \stackrel{\text{def}}{=} \sum_{i \in R'} p_i$.

We assume that a distribution can be specified in one of the two following ways.

Definition 2.1 (Black-box distribution) *A **black-box distribution** $\mathbf{p}$ is an oracle such that, upon request from an algorithm that has access to it, outputs a sample distributed according to $\mathbf{p}$.*

Definition 2.2 (Explicit distribution) *An **explicit distribution** $\mathbf{q}$ is an oracle that on input i outputs q_i.*

Definition 2.3 (Distinguishing random variables) *Let X and Y be discrete random variables taking values from a set $\mathcal{D}$. An algorithm $\mathcal{A}$ **distinguishes** X and Y, if given a sample from one of X or Y, $\mathcal{A}$ can correctly identify the source of the sample with probability at least $2/3$. The success probability is taken over the randomness of the random variables and the internal coin tosses of $\mathcal{A}$.*

The following observation is well-known:

Observation 2.4 *Let X and Y be as in Definition 2.3 and $\epsilon \stackrel{\text{def}}{=} \sum_{a \in \mathcal{D}} |\Pr[X = a] - \Pr[Y = a]|$. Then, no algorithm can distinguish X and Y with success probability more than $\frac{1}{2} + \frac{\epsilon}{4}$.*

2.1 Some useful theorems

We shall use the following inequalities often.

Theorem 2.5 (Markov's inequality) *Let X be a random variable assuming only non-negative values. Then for $\kappa > 0$,*

$$\Pr[X \geq \kappa \mathrm{E}[X]] \leq \frac{1}{\kappa}.$$

Theorem 2.6 (Chebyshev inequality) *Let X be a random variable with expectation $\mathrm{E}[X]$ and standard deviation σ_X. Then for $\kappa > 0$,*

$$\Pr\left[|X - \mathrm{E}[X]| \geq \kappa\sigma_X\right] \leq \frac{1}{\kappa^2}.$$

Theorem 2.7 (Chernoff bounds) *Let $X_1, X_2, \ldots, X_m$ be m independent random variables where $X_i \in [0,1]$. Let $\rho \stackrel{\text{def}}{=} \frac{1}{m}\sum_i \mathrm{E}[X_i]$. Then, for every $\gamma \in [0,1]$, the following bounds hold:*

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} X_i \;>\; (1+\gamma)\rho\right] < \exp(-\gamma^2\rho m/3)$$

and

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} X_i \;<\; (1-\gamma)\rho\right] < \exp(-\gamma^2\rho m/2).$$

The following theorem states that all sufficiently large entries of a probability vector can be estimated efficiently from frequency counts in a sample set.

Theorem 2.8 *Given a black-box distribution $\mathbf{p}$ over R, a threshold t and an accuracy parameter $\epsilon > 0$, there is an algorithm that requires $O(t^{-1}\epsilon^{-2}\log|R|\log(1/\delta))$ samples and outputs an estimate $\tilde{\mathbf{p}}$ such that with probability at least $1-\delta$, $\forall i \in R, p_i \geq t$ we have $(1-\epsilon)p_i \leq \tilde{p}(i) \leq (1+\epsilon)p_i$; the algorithm also outputs a set R' of members of R which includes $\{i \in R | p_i \geq t\}$ and on which the above approximation is guaranteed.*

The proof (omitted) of the above theorem is a simple application of a Chernoff bounds to independent samples from $\mathbf{p}$.

2.2 Restriction and coarsening

In this section, we define distributions induced by a distribution and a partition of its domain. The first one, restriction, is the conditional distribution on a subset of

the domain. The second one, coarsening, is obtained by coalescing all the elements in each set in a partitioning of the domain. We, then, prove some propositions regarding these induced distributions.

Definition 2.9 (Restriction) *Given a distribution* $\mathbf{p}$ *over* R*, and* $R' \subseteq R$*, the* ***restriction*** $(\mathbf{p}^{\downarrow R'})$ *is the distribution over* R' *such that, for all* $i \in R'$*,* $(\mathbf{p}^{\downarrow R'})_i = p_i/\mathbf{p}(R')$*.*

Definition 2.10 (Coarsening) *Given a distribution* $\mathbf{p}$ *over* R*, and a partition* $\mathcal{R} = \{R_1, \ldots, R_k\}$ *of* R*, the* ***coarsening*** $(\mathbf{p}^{\langle \mathcal{R} \rangle})$ *is the distribution over* $[k]$ *with distribution defined by* $(\mathbf{p}^{\langle \mathcal{R} \rangle})_i = \mathbf{p}(R_i)$*.*

We have the following:

Observation 2.11 *If* $\mathbf{p}$ *is a distribution over* R *and* $\mathcal{R} = \{R_1, \ldots, R_k\}$ *is a partition of* R*, then for all* i *in* $[k]$ *and* j *in* R_i*,* $p_j = (\mathbf{p}^{\langle \mathcal{R} \rangle})_i \cdot (\mathbf{p}^{\downarrow R_i})_j$*.*

In words, the probability of picking an element j belonging to the partition R_i according to $\mathbf{p}$ is equivalent to the probability of picking the partition R_i times the probability of picking j when restricted to the partition R_i.

The following lemma shows that two distributions are close if they are close with respect to restrictions and coarsening.

Lemma 2.12 *Let* $\mathbf{p}, \mathbf{q}$ *be distributions over* R *and let* $\mathcal{R} = \{R_1, \ldots, R_k\}$ *be a partition of* R*. If for all* i *in* $[k]$*,* $|(\mathbf{p}^{\downarrow R_i}) - (\mathbf{q}^{\downarrow R_i})| \leq \epsilon_1$ *and* $|(\mathbf{p}^{\langle \mathcal{R} \rangle}) - (\mathbf{q}^{\langle \mathcal{R} \rangle})| \leq \epsilon_2$*, then* $|\mathbf{p} - \mathbf{q}| \leq \epsilon_1 + \epsilon_2$*.*

PROOF: Using the hypothesis and Observation 2.11,

$$\begin{aligned}
|\mathbf{p}-\mathbf{q}| &= \sum_{i\in[k]}\sum_{j\in R_i}|p_j-q_j| = \sum_{i\in[k]}\sum_{j\in R_i}|(\mathbf{p}^{\langle\mathcal{R}\rangle})_i\cdot(\mathbf{p}^{\downarrow R_i})_j-(\mathbf{q}^{\langle\mathcal{R}\rangle})_i\cdot(\mathbf{q}^{\downarrow R_i})_j| \\
&\leq \sum_{i\in[k]}\sum_{j\in R_i}|(\mathbf{p}^{\langle\mathcal{R}\rangle})_i\cdot(\mathbf{p}^{\downarrow R_i})_j-(\mathbf{p}^{\langle\mathcal{R}\rangle})_i\cdot(\mathbf{q}^{\downarrow R_i})_j| \\
&\quad+\sum_{i\in[k]}\sum_{j\in R_i}|(\mathbf{p}^{\langle\mathcal{R}\rangle})_i\cdot(\mathbf{q}^{\downarrow R_i})_j-(\mathbf{q}^{\langle\mathcal{R}\rangle})_i\cdot(\mathbf{q}^{\downarrow R_i})_j| \\
&= \sum_{i\in[k]}\sum_{j\in R_i}(\mathbf{p}^{\langle\mathcal{R}\rangle})_i\cdot|(\mathbf{p}^{\downarrow R_i})_j-(\mathbf{q}^{\downarrow R_i})_j|+\sum_{i\in[k]}\sum_{j\in R_i}(\mathbf{q}^{\downarrow R_i})_j\cdot|(\mathbf{p}^{\langle\mathcal{R}\rangle})_i-(\mathbf{q}^{\langle\mathcal{R}\rangle})_i| \\
&\leq \epsilon_1\sum_{i\in[k]}(\mathbf{p}^{\langle\mathcal{R}\rangle})_i+\sum_{i\in[k]}|(\mathbf{p}^{\langle\mathcal{R}\rangle})_i-(\mathbf{q}^{\langle\mathcal{R}\rangle})_i| \leq \epsilon_1+\epsilon_2.
\end{aligned}$$

□

Note that for all $i \in [k]$, $(1-\epsilon)\mathbf{p}(R_i) \leq \mathbf{q}(R_i) \leq (1+\epsilon)\mathbf{p}(R_i)$ implies that $|(\mathbf{p}^{\langle\mathcal{R}\rangle})-(\mathbf{q}^{\langle\mathcal{R}\rangle})| \leq \epsilon$. The following lemma shows a partial converse: if $\mathbf{p}$ and $\mathbf{q}$ are close, then they are close when restricted to partitions of the domain with sufficiently large probability mass.

Lemma 2.13 *Let* $\mathbf{p}$ *and* $\mathbf{q}$ *be distributions over* R *and* $R' \subseteq R$. *Then* $|(\mathbf{p}^{\downarrow R'})-(\mathbf{q}^{\downarrow R'})| \leq 2\frac{|\mathbf{p}-\mathbf{q}|}{\mathbf{p}(R')}$.

PROOF:

$$\begin{aligned}
|(\mathbf{p}^{\downarrow R'})-(\mathbf{q}^{\downarrow R'})| &= \sum_{i\in R'}\left|\frac{p_i}{\mathbf{p}(R')}-\frac{q_i}{\mathbf{q}(R')}\right| = \sum_{i\in R'}\left|\frac{p_i}{\mathbf{p}(R')}-\frac{q_i}{\mathbf{p}(R')}+\frac{q_i}{\mathbf{p}(R')}-\frac{q_i}{\mathbf{q}(R')}\right| \\
&\leq \frac{1}{\mathbf{p}(R')}\sum_{i\in R'}|p_i-q_i|+\left|\frac{1}{\mathbf{p}(R')}-\frac{1}{\mathbf{q}(R')}\right|\sum_{i\in R'}q_i \\
&\leq \frac{1}{\mathbf{p}(R')}|\mathbf{p}-\mathbf{q}|+\left|\frac{1}{\mathbf{p}(R')}-\frac{1}{\mathbf{q}(R')}\right|\mathbf{q}(R') \\
&\leq \frac{1}{\mathbf{p}(R')}\left(|\mathbf{p}-\mathbf{q}|+|\mathbf{q}(R')-\mathbf{p}(R')|\right) \leq \frac{1}{\mathbf{p}(R')}\left(|\mathbf{p}-\mathbf{q}|+|\mathbf{p}-\mathbf{q}|\right) = 2\frac{|\mathbf{p}-\mathbf{q}|}{\mathbf{p}(R')}
\end{aligned}$$

□

2.3 Bucketing

In this section, we introduce a tool that we use in multiple occasions in the subsequent chapters. We use this tool, we call bucketing, to reduce the general problem to a more specific case: in particular, to the case where the input distributions are close to the uniform distribution. The inherent homogeneous nature of the uniform distribution makes the problem easier, at least qualitatively if not quantitatively.

Bucketing is a general tool which decomposes an arbitrary probability distribution into a collection of distributions that are almost uniform. The restriction of a distribution to a set of elements that have similar probabilities is close to uniform.

We define *Bucket*$(\mathbf{p}, R, \epsilon)$ as a partition $(R_0, R_1, \ldots, R_k)$ of R with $k \leq (2/\log(1+\epsilon)) \cdot \log|R|$ such that $R_0 = \{i \mid p_i \leq 1/(|R|\log|R|)\}$, and for all i in $[k]$,

$$R_i = \left\{ j \;\middle|\; \frac{(1+\epsilon)^{i-1}}{|R|\log|R|} \leq p_j \leq \frac{(1+\epsilon)^i}{|R|\log|R|} \right\}.$$

Lemma 2.14 *Let* $\mathbf{p}$ *be a distribution over* R *and let* $(R_0, \ldots, R_k) = \mathit{Bucket}(\mathbf{p}, R, \epsilon)$. *We have* $|(\mathbf{p}^{\downarrow R_i}) - U_{R_i}| \leq \epsilon$, $\|(\mathbf{p}^{\downarrow R_i}) - U_{R_i}\| \leq \epsilon^2/|R_i|$ *for* $i \in [k]$, *and* $\mathbf{p}(R_0) \leq 1/\log|R|$.

Proof: Clearly, $\mathbf{p}(R_0) \leq 1/\log|R|$. For $i \geq 1$, consider an arbitrary (non-empty) subset R_i. Without loss of generality, let $R_i = \{1, \ldots, \ell\}$ with $p_1 \leq \cdots \leq p_\ell$. Let $\mathbf{q} = (\mathbf{p}^{\downarrow R_i})$. Then, $q_\ell/q_1 < 1+\epsilon$. Also, by averaging, $q_1 \leq 1/\ell \leq q_\ell$. Hence $q_\ell \leq (1+\epsilon)q_1 \leq (1+\epsilon)/\ell$. Similarly it can be shown that $q_1 \geq 1/(\ell(1+\epsilon)) > (1-\epsilon)/\ell$. Thus, it follows that $|q_j - 1/\ell| \leq \epsilon/\ell$ for all $j = 1, \ldots, \ell$ and therefore, $|\mathbf{q} - U_{R_i}| \leq \epsilon$ and $\|\mathbf{q} - U_{R_i}\| \leq \epsilon^2/\ell$. □

Given an approximation $\tilde{\mathbf{p}}$ of $\mathbf{p}$, the bucketing of $\tilde{\mathbf{p}}$ has similar properties as the bucketing of $\mathbf{p}$.

Corollary 2.15 *Let* $\mathbf{p}$ *and* $\tilde{\mathbf{p}}$ *be distributions over* R *such that* $\tilde{\mathbf{p}}$ *approximates* $\mathbf{p}$*, i.e.,* $\forall i \in R, (1-\epsilon)p_i \leq \tilde{p}_i \leq (1+\epsilon)p_i$ *for some* $\epsilon > 0$*. Let* $Bucket(\tilde{\mathbf{p}}, R, \epsilon)$ *be the partition* $\{R_0, \ldots, R_k\}$ *of* R *with* $k = O(\epsilon^{-1} \log |R|)$*. Then, for all* $i \geq 1$*,* $|(\mathbf{p}^{\downarrow R_i}) - U_{R_i}| \leq 3\epsilon$ *and* $\mathbf{p}(R_0) \leq (1+\epsilon)/\log |R|$*.*

In our applications of bucketing, we usually ignore the bucket R_0 since the probability mass on this bucket would be negligible for our purposes.

Let $\mathcal{R}$ be a partition of the domain R generated by bucketing according to distribution $\mathbf{q}$. In the next lemma, we show that if another distribution $\mathbf{p}$ is close to uniform when restricted to a set R_i in $\mathcal{R}$, we can infer that distributions $\mathbf{p}$ and $\mathbf{q}$ are close when restricted to R_i.

Lemma 2.16 *Let* $\mathbf{p}, \mathbf{q}$ *be distributions over* R *and let* $(R_0, \ldots, R_k) = Bucket(\mathbf{q}, R, \epsilon)$*. For each* i *in* $[k]$*, if* $\|(\mathbf{p}^{\downarrow R_i})\|^2 \leq (1+\epsilon^2)/|R_i|$ *then* $|(\mathbf{p}^{\downarrow R_i}) - U_{R_i}| \leq \epsilon$ *and* $|(\mathbf{p}^{\downarrow R_i}) - (\mathbf{q}^{\downarrow R_i})| \leq 2\epsilon$*.*

The proof of this lemma uses the following fact.

Fact 2.17 *For any distribution* $\mathbf{p}$ *over* R*,* $\|\mathbf{p}\|^2 - \|U_R\|^2 = \|\mathbf{p} - U_R\|^2$*.*

PROOF: $\|\mathbf{p}\|^2 - \|U_R\|^2 = (\sum_{j \in R} p_j^2) - \frac{1}{|R|} = (\sum_{j \in R} p_j^2) + \frac{1}{|R|} - \frac{2}{|R|}\sum_{j \in R} p_j = \sum_{j \in R}(p_j - \frac{1}{|R|})^2 = \|\mathbf{p} - U_R\|^2$. □

PROOF: (of Lemma 2.16) Using Fact 2.17, note that for a distribution $\mathbf{p}$ over R, if $\|\mathbf{p} - U_R\|^2 = \|\mathbf{p}\|^2 - \|U_R\|^2 \leq \alpha$ then $|\mathbf{p} - U_R| \leq \sqrt{\alpha |R|}$. Since we are given $\|(\mathbf{p}^{\downarrow R_i})\|^2 - \|U_{R_i}\|^2 = \|(\mathbf{p}^{\downarrow R_i})\|^2 - 1/|R_i| \leq \epsilon^2/|R_i|$, the first part of the lemma follows. We also know that $|(\mathbf{q}^{\downarrow R_i}) - U_{R_i}| \leq \epsilon$ by Lemma 2.14. This gives $|(\mathbf{p}^{\downarrow R_i}) - (\mathbf{q}^{\downarrow R_i})| \leq 2\epsilon$. □

Chapter 3

Testing closeness and identity of distributions

In this chapter, we present two tests that distinguish identical pairs of distributions from pairs of distributions that have large L_1 distance. Both of the tests take a threshold parameter ϵ and have access to two distributions $\mathbf{p}$ and $\mathbf{q}$ over $[n]$. In the case of the closeness test, both $\mathbf{p}$ and $\mathbf{q}$ are black-box distributions, whereas the identity test has access to a black-box distribution $\mathbf{p}$ and an explicit distribution $\mathbf{q}$. The tests satisfy the following conditions: (1) if $\mathbf{p} = \mathbf{q}$, then the test outputs PASS with probability at least 2/3, (2) if $|\mathbf{p} - \mathbf{q}| \geq \epsilon$, then the test outputs FAIL with probability at least 2/3.

We then present lower bounds on the sample complexity of these tasks as a function of the domain size of the distributions. We show $\Omega(n^{2/3})$ samples are required for testing closeness, whereas $\Omega(\sqrt{n})$ samples are required for testing identity. The lower bounds match the respective upper bounds up to polylogarithmic factors. Besides showing the near optimality of our tests, these tight results establish the relative hardness of testing closeness with respect to testing identity.

Related works arise in several different contexts. Goldreich and Ron [16] give methods for testing that the L_2 distance between a given black-box distribution and the uniform distribution is small in time $O(\sqrt{n})$. Their method is based on the number of collisions in the sample set. We also use collisions in the tests presented here.

In an interactive setting, Sahai and Vadhan [23] show that given distributions $\mathbf{p}$ and $\mathbf{q}$, generated by polynomial-size circuits, the problem of distinguishing whether $\mathbf{p}$ and $\mathbf{q}$ are close or far in L_1 norm, is complete for statistical zero-knowledge.

A related work of Kannan and Yao [19] outlines a program checking framework for certifying the randomness of a program's output. Their model differs in that one does not assume that samples from the input distribution are independent.

There is much work on the problem estimating the distance between distributions in data streaming models where space is limited rather than time (cf. [14, 2, 10, 12]). Another line of work [6] estimates the distance in frequency count distributions on words between various documents, where again space is limited.

There is a vast literature on testing statistical hypotheses. In these works, one is given examples chosen from the same distribution out of two possible choices, say $\mathbf{p}$ and $\mathbf{q}$. The goal is to decide which of two distributions the examples are coming from. More generally, the goal can be stated as deciding which of two known classes of distributions contains the distribution generating the examples. This can be seen to be a generalization of our model as follows: Let the first class of distributions be the set of distributions of the form $\mathbf{q} \times \mathbf{q}$. Let the second class of distributions be the set of distributions of the form $\mathbf{p} \times \mathbf{q}$ where the L_1 distance of $\mathbf{p}$ and $\mathbf{q}$ is at least ϵ. Then, given examples from two distributions $\mathbf{u}, \mathbf{v}$, create a set of example pairs (x, y) where x is chosen according to $\mathbf{u}$ and y according to $\mathbf{v}$. Bounds and an

optimal algorithm for the general problem for various distance measures are given in [7, 22, 8, 9, 20, 28]. None of these give sublinear bounds in the domain size for our problem.

3.1 Testing closeness of distributions

We develop a closeness test such that, given ϵ and access to two black-box distributions over $[n]$, it distinguishes identical pairs of distributions from pairs of distributions that have L_1 distance larger than ϵ. Actually, we can prove a slightly stronger conditions: (1) if the distributions have L_1 distance at most $\max(\frac{\epsilon^2}{32\sqrt[3]{n}}, \frac{\epsilon}{4\sqrt{n}})$ then the algorithm will accept with probability at least 2/3, and (2) if the distributions have L_1 distance more than ϵ then the algorithm will accept with probability at most 1/3. The number of samples used is $O(n^{2/3}\epsilon^{-4}\log n)$. By repeating the the algorithm $O(\log(1/\delta)$ times and outputting the majority vote, the error probability can be brought down to δ. In Section 3.1.4, we give an $\Omega(n^{2/3})$ lower bound for testing L_1 distance.

Our closeness test for L_1 distance relies on a test for the L_2 distance, which is considerably easier to test: we give an algorithm that uses a number of samples that is independent of n. However, the L_2 distance does not in general give a good measure of the closeness of two distributions. For example, two distributions can have disjoint support and still have small L_2 distance; two distributions each uniform on disjoint halves of the domain have L_2 distance $4/n$. Still, we can estimate the L_2 distance of the distributions to within $\epsilon/\sqrt{n}$ and then use the fact that the L_1 distance is at most $\sqrt{n}$ times the L_2 distance for n-vectors. Unfortunately, the number of queries required by this approach is too large in general. Because of this, our L_1 test is forced to distinguish between two cases.

For distributions with small L_2 norm, we show how to use the L_2 distance to get a good approximation of the L_1 distance. For distributions with larger L_2 norm, we use the fact that such distributions must have elements which occur with relatively high probability. We create a filtering test that estimates the L_1 distance due to these high probability elements, and then approximates the L_1 distance due to the low probability elements using the test for L_2 distance. Optimizing the notion of "high probability" yields our $O(n^{2/3}\epsilon^{-4}\log n)$-time algorithm.

Theorem 3.1 *Given parameters ϵ, δ, and black-box distributions $\mathbf{p}, \mathbf{q}$ over a set of n elements, there is a test which runs in time $O(\epsilon^{-4}n^{2/3}\log n \log\frac{1}{\delta})$ such that if $|\mathbf{p}-\mathbf{q}| \leq \max(\frac{\epsilon^2}{32\sqrt[3]{n}}, \frac{\epsilon}{4\sqrt{n}})$, then the test outputs* `pass` *with probability at least $1-\delta$ and and if $|\mathbf{p}-\mathbf{q}| > \epsilon$, then the test outputs* `fail` *with probability at least $1-\delta$.*

The proof of this theorem is given in Section 3.1.2. In order to prove this theorem, we give a test which determines whether $\mathbf{p}$ and $\mathbf{q}$ are close in L_2 norm. The test is based on estimating the self-collision and collision probabilities of $\mathbf{p}$ and $\mathbf{q}$. In particular, if $\mathbf{p}$ and $\mathbf{q}$ are close, one would expect that the self-collision probabilities of each are close to the collision probability of the pair. Formalizing this intuition, in Section 3.1.1, we prove:

Theorem 3.2 *Given parameters ϵ, δ, and black-box distributions $\mathbf{p}$ and $\mathbf{q}$ over a set of n elements, there exists a test such that if $\|\mathbf{p}-\mathbf{q}\| \leq \epsilon/2$ then the test passes with probability at least $1-\delta$. If $\|\mathbf{p}-\mathbf{q}\| > \epsilon$ then the test passes with probability less than δ. The running time of the test is $O(\epsilon^{-4}\log\frac{1}{\delta})$.*

The test used to prove Theorem 3.2 is given in Figure 3.1. The number of pairwise self-collisions in sample set F is the count of $i < j$ such that the i^{th} sample in F is same as the j^{th} sample in F. Similarly, the number of collisions between Q_p and

Q_q is the count of (i, j) such that the i^{th} sample in Q_p is same as the j^{th} sample in Q_q. We use the parameter m to indicate the number of samples needed by the test

```
L_2-Distance-Test(p, q, m, ε, δ)
Repeat O(log(1/δ)) times
  Let F_p = a set of m samples from p
  Let F_q = a set of m samples from q
    Let r_p be the number of pairwise
      self-collisions in F_p.
    Let r_q be the number of pairwise
      self-collisions in F_q.
  Let Q_p = a set of m samples from p
  Let Q_q = a set of m samples from q
    Let s_pq be the number of collisions
      between Q_p and Q_q.
  Let r = (2m/(m-1))(r_p + r_q)
  Let s = 2s_pq
  If r - s > m^2 ε^2/2 then reject
Reject if the majority of iterations reject,
accept otherwise
```

Figure 3.1: Algorithm L_2-Distance-Test

to get constant confidence. In order to bound the L_2 distance between $\mathbf{p}$ and $\mathbf{q}$ by ϵ, we show that setting $m = O(\frac{1}{\epsilon^4})$ suffices. By maintaining arrays which count the number of times that each element is sampled in F_p, F_q, one can achieve the claimed running time bounds. Thus essentially m^2 estimations of the collision probability can be performed in $O(m)$ time.

Since $|\mathbf{v}| \leq \sqrt{n}\|\mathbf{v}\|$, a simple way to extend the above test to an L_1-distance test is by setting $\epsilon' = \epsilon/\sqrt{n}$. Unfortunately, due to the order of the dependence on ϵ in the L_2-distance test, the resulting running time is prohibitive. It is possible, though, to achieve sublinear running times if the input vectors are known to be reasonably evenly distributed. We make this precise by a closer analysis of the variance of the decision variable of the test in Lemma 3.5. In particular, we analyze the dependence of the variance of s on the parameter $b = \max(\|\mathbf{p}\|_\infty, \|\mathbf{q}\|_\infty)$. There we show that given $\mathbf{p}$ and $\mathbf{q}$ such that $\max(\|\mathbf{p}\|_\infty, \|\mathbf{q}\|_\infty) \leq n^{-\alpha}$ for some α, one can call L_2**-Distance-Test** with an error parameter of $\frac{\epsilon}{\sqrt{n}}$ and achieve a running time of $O(\epsilon^{-4}(n^{1-\alpha/2} + n^{2-2\alpha}))$.

We use the following definition to identify the elements with large weights.

Definition 3.3 *An element i is called* ***heavy*** *with respect to a distribution* $\mathbf{p}$ *if* $p_i \geq \frac{1}{n^{2/3}}$.

Our L_1-distance tester calls the L_2-distance testing algorithm as a subroutine. When both input distributions have no heavy elements, the input is passed to the L_2-distance test unchanged. If the input distributions have a large self-collision probability, the distances induced respectively by the heavy and non-heavy elements are measured in two steps. The first step measures the distance corresponding to the heavy elements via straightforward sampling and estimates probabilities using normalized frequency counts, and the second step modifies the distributions so that the distance attributed to the non-heavy elements can be measured using the L_2-distance test. The complete test is given in Figure 3.2. The proof of Theorem 3.1 is described in Section 3.1.2.

L_1-Distance-Test($\mathbf{p}, \mathbf{q}, \epsilon, \delta$)

Sample $\mathbf{p}$ and $\mathbf{q}$ for

  $M = O(\max(\epsilon^{-2}, 4)n^{2/3} \log n)$ times

Let S_p and S_q be the sample sets obtained

  by discarding elements that occur less

  than $(1 - \epsilon/63)Mn^{-2/3}$ times

If S_p and S_q are empty

    L_2-Distance-Test($\mathbf{p}, \mathbf{q}, O(n^{2/3}/\epsilon^4), \frac{\epsilon}{2\sqrt{n}}, \delta/2$)

else

  $\ell_i^{\mathbf{p}} = \#$ times element i appears in S_p

  $\ell_i^q = \#$ times element i appears in S_q

  Fail if $\sum_i |\ell_i^{\mathbf{p}} - \ell_i^q| > \epsilon M/8$.

  Define $\mathbf{p}'$ as follows:

    sample an element from $\mathbf{p}$

    if this sample is not in S_p output it,

    otherwise output an $x \in_R [n]$.

  Define $\mathbf{q}'$ similarly.

  L_2-Distance-Test($\mathbf{p}', \mathbf{q}', O(n^{2/3}/\epsilon^4), \frac{\epsilon}{2\sqrt{n}}, \delta/2$)

Figure 3.2: Algorithm L_1-Distance-Test

3.1.1 Testing closeness in L_2 norm

In this section we analyze the test in Figure 3.1 and prove Theorem 3.2. The statistics r_p, r_q and s in Algorithm L_2**-Distance-Test** are estimators for the self-collision probability of $\mathbf{p}$, of $\mathbf{q}$, and of the collision probability between $\mathbf{p}$ and $\mathbf{q}$, respectively. If $\mathbf{p}$ and $\mathbf{q}$ are statistically close, then the self-collision probabilities of each are close to the collision probability of the pair. These collision probabilities are exactly the inner products of these vectors. In particular if the set F_p of samples from $\mathbf{p}$ is given by $\{F_p^1, \ldots, F_p^m\}$ then for any pair $i, j \in [m], i \neq j$, we have that $\Pr\left[F_p^i = F_p^j\right] = \mathbf{p} \cdot \mathbf{p} = \|\mathbf{p}\|^2$. Similarly, if $O_p = \{Q_p^1, \ldots, Q_p^m\}$ and $O_q = \{Q_q^1, \ldots, Q_q^m\}$ are m samples from $\mathbf{p}$ and $\mathbf{q}$, respectively, then for any $i, j \in [m]$, we have that $\Pr\left[Q_p^i = Q_q^j\right] = \mathbf{p} \cdot \mathbf{q}$. One distinction to make between self-collisions and $\mathbf{p}, \mathbf{q}$ collisions is that for the self-collisions, we can only consider samples for which $i \neq j$, but this is not necessary for $\mathbf{p}, \mathbf{q}$ collisions. We accommodate this in our algorithm by scaling r_p and r_q appropriately. By this scaling and from the above discussion we see that $\mathrm{E}[s] = 2m^2(\mathbf{p} \cdot \mathbf{q})$ and that $\mathrm{E}[r - s] = m^2(\|\mathbf{p}\|^2 + \|\mathbf{q}\|^2 - 2(\mathbf{p} \cdot \mathbf{q})) = m^2(\|\mathbf{p} - \mathbf{q}\|^2)$.

A complication which arises from this scheme is that the pairwise samples are not independent. To analyze our test, we use Chebyshev's inequality. That is, for any random variable A, and $\rho > 0$, the probability $\Pr[|A - E[A]| > \rho]$ is bounded above by $\frac{\mathrm{Var}[A]}{\rho^2}$. To use this theorem, we require a bound on the variance, which we give in this section.

Our techniques extend the work of Goldreich and Ron [16], where self-collision probabilities are used to estimate norm of a vector, and the deviation of a distribution from uniform. In particular, their work provides an analysis of the statistics r_p and r_q above through the following lemma.

Lemma 3.4 (Goldreich Ron) *Let A be one of r_p or r_q in algorithm L_2-***Distance-Test***. Then $\mathrm{E}[A] = \binom{m}{2} \cdot \|\mathbf{p}\|^2$ and $\mathrm{Var}[A] \leq 2(\mathrm{E}[A])^{3/2}$**

It turns out that the bound on the variance is not sufficiently tight for our purposes. We extend their bounds and get a tighter analysis in terms of the infinity norms of the of the distributions.

Lemma 3.5 *There is a constant c such that $\mathrm{Var}[r-s] \leq c(m^3b^2 + m^2b)$, where $b = \max(\|\mathbf{p}\|_\infty, \|\mathbf{q}\|_\infty)$.*

PROOF: For $(i,j) \in [m] \times [m]$, define the indicator variable $C_{i,j} = 1$ if the i^{th} element of Q_p and the j^{th} element of Q_q are the same. Then the variable from the algorithm $s_{pq} = \sum_{i,j} C_{i,j}$. Also define the notation $\bar{C}_{i,j} = C_{i,j} - \mathrm{E}[C_{i,j}]$.

We can write $\mathrm{Var}\left[\sum_{F\times F} C_{i,j}\right] = \mathrm{E}\left[(\sum_{F\times F} \bar{C}_{i,j})^2\right]$, and unfold the summation and use the linearity of expectation to get

$$\mathrm{Var}[s_{pq}] = \mathrm{E}\left[\sum_{i,j}(\bar{C}_{i,j})^2 + 2\sum_{(i,j)\neq(k,l)} \bar{C}_{i,j}\bar{C}_{k,l}\right] \leq m^2(\mathbf{p}\cdot\mathbf{q}) + 2\mathrm{E}\left[\sum_{(i,j)\neq(k,l)} \bar{C}_{i,j}\bar{C}_{k,l}\right].$$

To analyze the last expectation, we use two facts. First, it is easy to see, by the definition of covariance, that $\mathrm{E}\left[\bar{C}_{i,j}\bar{C}_{k,l}\right] \leq \mathrm{E}[C_{i,j}C_{k,l}]$. Secondly, we note that $C_{i,j}$ and $C_{k,l}$ are not independent only when $i = k$ or $j = l$. Expanding the sum we get

$$\begin{aligned}
\mathrm{E}\left[\sum_{\substack{(i,j),(k,l)\in F\times F\\(i,j)\neq(k,l)}} \bar{C}_{i,j}\bar{C}_{k,l}\right] &= \mathrm{E}\left[\sum_{\substack{(i,j),(i,l)\in F\times F\\j\neq l}} \bar{C}_{i,j}\bar{C}_{i,l} + \sum_{\substack{(i,j),(k,j)\in F\times F\\i\neq k}} \bar{C}_{i,j}\bar{C}_{k,j}\right] \\
&\leq \mathrm{E}\left[\sum_{\substack{(i,j),(i,l)\in F\times F\\j\neq l}} C_{i,j}C_{i,l} + \sum_{\substack{(i,j),(k,j)\in F\times F\\i\neq k}} C_{i,j}C_{k,j}\right] \\
&\leq cm^3\sum_{\ell\in[n]} p_\ell q_\ell^2 + p_\ell^2 q_\ell \leq cm^3b^2\sum_{\ell\in[n]} q_\ell \leq cm^3b^2
\end{aligned}$$

for some constant c. In order to bound $\mathrm{Var}[r-s]$ we use Lemma 3.4. Since $\mathrm{Var}[r] \leq cm^2b$ and the variance is additive for independent random variables, we can write $\mathrm{Var}[r-s] \leq c(m^3b^2 + m^2b)$. □

Now using Chebyshev's inequality, it follows that if we choose $m = O(\epsilon^{-4})$, we can achieve an error probability less than 1/3. It follows from standard techniques that with $O(\log \frac{1}{\delta})$ iterations we can achieve an error probability at most δ.

Lemma 3.6 *For two distributions* $\mathbf{p}$ *and* $\mathbf{q}$ *such that* $b = \max(\|\mathbf{p}\|_\infty, \|\mathbf{q}\|_\infty)$ *and* $m = O((b^2 + \epsilon^2\sqrt{b})/\epsilon^4)$, *if* $\|\mathbf{p}-\mathbf{q}\| \leq \epsilon/2$, *then* L_2***-Distance-Test****(*$\mathbf{p}, \mathbf{q}, m, \epsilon, \delta$*) passes with probability at least* $1-\delta$. *If* $\|\mathbf{p}-\mathbf{q}\| > \epsilon$, *then the probability that* L_2***-Distance-Test****(*$\mathbf{p}, \mathbf{q}, m, \epsilon, \delta$*) passes is less than* δ. *The running time is* $O(m \log(\frac{1}{\delta}))$.

Proof: For our statistic $A = (r-s)$ we can say, using Chebyshev's inequality, that for some constant c,

$$\Pr[|A - \mathrm{E}[A]| > \rho] \leq \frac{c(m^3b^2 + m^2b)}{\rho^2}$$

Then when $\|\mathbf{p}-\mathbf{q}\| \leq \epsilon/2$, for one iteration,

$$\begin{aligned} \Pr[\text{pass}] &= \Pr[(r-s) < m^2\epsilon^2/2] \\ &\geq \Pr[|(r-s) - \mathrm{E}[r-s]| < m^2\epsilon^2/4] \\ &\geq 1 - \tfrac{4c(m^3b^2+m^2b)}{m^4\epsilon^4} \end{aligned}$$

It can be shown that this probability will be at least 2/3 whenever $m > k(b^2 + \epsilon^2\sqrt{b})/\epsilon^4$ for some constant k. A similar analysis can be used to show the other direction. □

3.1.2 Testing closeness in L_1 norm

The L_1-closeness test proceeds in two stages. The first phase of the algorithm filters out heavy elements (as defined in Definition 3.3) while estimating their contribu-

tion to the distance $|\mathbf{p}-\mathbf{q}|$. The second phase invokes the L_2 test on the filtered distribution, with closeness parameter $\frac{\epsilon}{2\sqrt{n}}$. The correctness of this subroutine call is given by Lemma 3.6 with $b = n^{-2/3}$. With these substitutions, the number of samples m is $O(\epsilon^{-4}n^{2/3})$. The choice of threshold $n^{-2/3}$ for the weight of the heavy elements arises from optimizing the running-time trade-off between the two phases of the algorithm.

We need to show that by using a sample of size $O(\epsilon^{-2}n^{2/3}\log n)$, we can estimate the weights of the heavy elements to within a multiplicative factor of $O(\epsilon)$.

Lemma 3.7 *Let $\epsilon \leq 1/2$. In L_1-**Distance-Test**, after performing $M = O(\frac{n^{2/3}\log n}{\epsilon^2})$ samples from a distribution $\mathbf{p}$, we define $\bar{p}_i = \ell_i^p/M$. Then, with probability at least $1-\frac{1}{n}$, the following hold for all i: (1) if $p_i \geq \epsilon^2 n^{-2/3}$ then $|\bar{p}_i - p_i| < \frac{\epsilon}{63}\max(p_i, n^{-2/3})$, (2) if $p_i < \epsilon^2 n^{-2/3}$, $\bar{p}_i < (1-\epsilon/63)n^{-2/3}$.*

PROOF: Using Chernoff bounds, one sees that for each i, with probability at least $1-\frac{1}{n^2}$, the following holds: (1a) If $p_i > n^{-2/3}$ then $|\bar{p}_i - p_i| < \epsilon p_i/63$. (1b) If $\epsilon^2 n^{-2/3} < p_i \leq n^{-2/3}$ then $|\bar{p}_i - p_i| < \epsilon n^{-2/3}/63$. (2) If $p_i < \epsilon^2 n^{-2/3}$ then $\bar{p}_i < 3\epsilon^2 n^{-2/3}$. Since, for $\epsilon \leq 1/2$, $3\epsilon^2 \leq (1-\epsilon/63)$, the lemma follows. □

Once the heavy elements are identified, we use the following fact to prove the gap in the distances of accepted and rejected pairs of distributions.

Fact 3.8 *For any vector $\mathbf{v}$, $\|\mathbf{v}\|^2 \leq |\mathbf{v}| \cdot \|\mathbf{v}\|_\infty$.*

Theorem 3.9 *L_1-**Distance-Test** passes distributions $\mathbf{p}, \mathbf{q}$ such that $|\mathbf{p}-\mathbf{q}| \leq \max(\frac{\epsilon^2}{32\sqrt[3]{n}}, \frac{\epsilon}{4\sqrt{n}})$, and fails when $|\mathbf{p}-\mathbf{q}| > \epsilon$. The error probability is δ. The running time of the whole test is $O(\epsilon^{-4}n^{2/3}\log n \log(\frac{1}{\delta}))$.*

PROOF: Suppose items (1) and (2) from Lemma 3.7 hold for all i, and for both $\mathbf{p}$ and $\mathbf{q}$. By Lemma 3.7, this event happens with probability at least $1-\frac{2}{n}$.

Let $S = S_p \cup S_q$. By our assumption, all the heavy elements of both $\mathbf{p}$ and $\mathbf{q}$ are in S, and no element with weight less than $\epsilon^2 n^{-2/3}$ (in either distribution) is in S.

Let Δ_1 be the L_1 distance attributed to the elements in S. Let $\Delta_2 = |\mathbf{p}' - \mathbf{q}'|$ (in the case that S is empty, $\Delta_1 = 0$, $\mathbf{p} = \mathbf{p}'$ and $\mathbf{q} = \mathbf{q}'$).

Notice that $\Delta_1 \leq |\mathbf{p}-\mathbf{q}|$. We can show that $\Delta_2 \leq |\mathbf{p}-\mathbf{q}|$, and $|\mathbf{p}-\mathbf{q}| \leq 2\Delta_1+\Delta_2$.

The algorithm estimates Δ_1 in a brute-force manner to within an additive error of $\epsilon/9$. The error on the i^{th} term of the sum is bounded by $\frac{\epsilon}{63}(\max(p_i, n^{-2/3}) + \max(q_i, n^{-2/3})) \leq \frac{\epsilon}{63}(p_i + q_i + 2n^{-2/3})$. Consider the sum over i of these error terms. Notice that this sum is over at most $2n^{2/3}/(1 - \epsilon/63)$ elements in S. Hence, the total additive error is bounded by

$$\sum_{i \in S} \frac{\epsilon}{63}(p_i + q_i + 2n^{-2/3}) \leq \frac{\epsilon}{63}(2 + 4/(1 - \epsilon/63)) \leq \epsilon/9.$$

Note that $\max(\|\mathbf{p}'\|_\infty, \|\mathbf{q}'\|_\infty) < n^{-2/3} + n^{-1}$. So, we can use the L_2**-Distance-Test** on $\mathbf{p}'$ and $\mathbf{q}'$ with $m = O(\epsilon^{-4} n^{2/3})$ as shown by Lemma 3.6.

If $|\mathbf{p} - \mathbf{q}| < \frac{\epsilon^2}{32\sqrt[3]{n}}$ then so are Δ_1 and Δ_2. The first phase of the algorithm passes with probability at least $1 - (2/n)$. By Fact 3.8, $\|\mathbf{p}' - \mathbf{q}'\| \leq \frac{\epsilon}{4\sqrt{n}}$. Therefore, the L_2**-Distance-Test** passes with probability at least $\delta/2$. Similarly, if $|\mathbf{p} - \mathbf{q}| > \epsilon$ then either $\Delta_1 > \epsilon/4$ or $\Delta_2 > \epsilon/2$. Either the first phase of the algorithm or the L_2**-Distance-Test** will fail.

To get the running time, note that the time for the first phase is $O(\epsilon^{-2} n^{2/3} \log n)$ and that the time for L_2**-Distance-Test** is $O(n^{2/3}\epsilon^{-4} \log \frac{1}{\delta})$. It is easy to see that our algorithm makes an error either when it makes a bad estimation of Δ_1 or when L_2**-Distance-Test** makes an error. So, the probability of error is bounded by δ. □

The next theorem improves this result by looking at the dependence of the variance calculation in Section 3.1.1 on L_∞ norms of the distributions separately.

Theorem 3.10 *Given two black-box distributions* $\mathbf{p}, \mathbf{q}$ *over* $[n]$*, with* $\|\mathbf{p}\|_\infty \leq \|\mathbf{q}\|_\infty$*, there is a test requiring* $O((n^2\|\mathbf{p}\|_\infty\|\mathbf{q}\|_\infty\epsilon^{-4} + \sqrt{n}\|\mathbf{p}\|_\infty\epsilon^{-2})\log(1/\delta))$ *samples that (1) if* $|\mathbf{p} - \mathbf{q}| \leq \frac{\epsilon^2}{\sqrt[3]{n}}$*, it outputs PASS with probability at least* $1 - \delta$ *and (2) if* $|\mathbf{p} - \mathbf{q}| > \epsilon$*, it outputs FAIL with probability at least* $1 - \delta$*.*

Finally, by similar methods to the proof of Theorem 3.10 (in conjunction with those of [16]), we can show the following (proof omitted):

Theorem 3.11 *Given a black-box distribution* $\mathbf{p}$ *over* $[n]$*, there is a test that takes* $O(\epsilon^{-4}\sqrt{n}\log(n)\log(1/\delta))$ *samples, outputs PASS with probability at least* $1 - \delta$ *if* $\mathbf{p} = U_{[n]}$*, and outputs FAIL with probability at least* $1 - \delta$ *if* $|\mathbf{p} - U_{[n]}| > \epsilon$*.*

3.1.3 Characterization of canonical algorithms for testing properties of distributions

In this section, we characterize canonical algorithms for testing properties of distributions defined by permutation-invariant functions. The argument hinges on the irrelevance of the labels of the domain elements for such a function. We obtain this canonical form in two steps, corresponding to the two lemmas below. The first step makes explicit the intuition that such an algorithm should be symmetric, that is, the algorithm would not benefit from discriminating among the labels. In the second step, we remove the use of labels altogether, and show that we can present the sample to the algorithm in an aggregate fashion.

Characterizations of property testing algorithms have been studied in other settings. For example, using similar techniques, Alon et al. [1] show a canonical form for algorithms for testing graph properties. Later, Goldreich and Trevisan [15] formally prove the result by Alon et al. In a different setting, Bar-Yossef et al. [3] show

a canonical form for sampling algorithms that approximate symmetric functions of the form $f : A^n \to B$ where A and B are arbitrary sets. In the latter setting, the algorithm is given oracle access to the input vector and takes samples from the coordinate values of this vector.

Definition 3.12 (Permutation of a distribution) *For a distribution* $\mathbf{p}$ *over* $[n]$ *and a permutation* π *on* $[n]$*, define* $\pi(\mathbf{p})$ *to be the distribution such that for all* i*,* $\pi(\mathbf{p})_{\pi(i)} = p_i$*.*

Definition 3.13 (Symmetric Algorithm) *Let* $\mathcal{A}$ *be an algorithm that takes samples from* k *discrete black-box distributions over* $[n]$ *as input. We say that* $\mathcal{A}$ *is* ***symmetric*** *if, once the distributions are fixed, the output distribution of* $\mathcal{A}$ *is identical for any permutation of the distributions.*

Definition 3.14 (Permutation-invariant function) *A* k*-ary function* f *on distributions over* $[n]$ *is* ***permutation-invariant*** *if for any permutation* π *on* $[n]$*, and all distributions* $(\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(k)})$*,*

$$f(\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(k)}) = f(\pi(\mathbf{p}^{(1)}), \ldots, \pi(\mathbf{p}^{(k)})).$$

Lemma 3.15 *Let* $\mathcal{A}$ *be an arbitrary testing algorithm for a* k*-ary property* $\mathcal{P}$ *defined by a permutation-invariant function. Suppose* $\mathcal{A}$ *has sample complexity* $s(n)$*, where* n *is the domain size of the distributions. Then, there exists a symmetric algorithm that tests the same property of distributions with sample complexity* $s(n)$*.*

PROOF: Given the algorithm $\mathcal{A}$, construct a symmetric algorithm $\mathcal{A}'$ as follows: Choose a random permutation of the domain elements. Upon taking $s(n)$ samples, apply this permutation to each sample. Pass this (renamed) sample set to $\mathcal{A}$ and output according to $\mathcal{A}$.

It is clear that the sample complexity of the algorithm does not change. We need to show that the new algorithm also maintains the testing features of $\mathcal{A}$. Suppose that the input distributions $(\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(k)})$ have the property $\mathcal{P}$. Since the property is defined by a permutation-invariant function, any permutation of the distributions maintains this property. Therefore, the permutation of the distributions should be accepted as well. Then,

$$\Pr\left[\mathcal{A}' \text{ accepts } (\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(k)})\right] = \sum_{\text{perm. } \pi} \frac{1}{n!} \Pr\left[\mathcal{A} \text{ accepts } (\pi(\mathbf{p}^{(1)}), \ldots, \pi(\mathbf{p}^{(k)}))\right],$$

which is at least $2/3$ by the accepting probability of $\mathcal{A}$.

An analogous argument on the failure probability for the case of the distributions $(\mathbf{p}^{(1)}, \ldots, \mathbf{p}^{(k)})$ that should be rejected completes the proof. □

In order to avoid introducing additional randomness in $\mathcal{A}'$, we can try $\mathcal{A}$ on all possible permutations and output the majority vote. This change would not affect the sample complexity, and it can be shown that it maintains correctness.

Definition 3.16 (Fingerprint of a sample) *Let S_1 and S_2 be multisets of at most s samples taken from two black-box distributions over $[n]$, $\mathbf{p}$ and $\mathbf{q}$, respectively. Let the random variable C_{ij}, for $0 \leq i, j \leq s$, denote the number of elements that appear exactly i times in S_1 and exactly j times in S_2. The collection of values that the random variables $\{C_{ij}\}_{0 \leq i,j \leq s}$ take is called the **fingerprint** of the sample.*

For example, let sample sets be $S_1 = \{5, 7, 3, 3, 4\}$ and $S_2 = \{2, 4, 3, 2, 6\}$. Then, $C_{10} = 2$ (elements 5 and 7), $C_{01} = 1$ (element 6), $C_{11} = 1$ (element 4), $C_{02} = 1$ (element 2), $C_{21} = 1$ (element 3), and for remaining i, j's, $C_{ij} = 0$.

Lemma 3.17 *If there exists a symmetric algorithm $\mathcal{A}$ for testing a binary property of distributions defined by a permutation-invariant function, then there exist an*

algorithm for the same task that gets as input only the fingerprint of the sample that $\mathcal{A}$ *takes.*

PROOF: Fix a canonical order for C_{ij}'s in the fingerprint of a sample. Let us define the following transformation on the sample: Relabel the elements such that the elements that appear exactly the same number of times from each distribution (i.e., the ones that contribute to a single C_{ij} in the fingerprint) have consecutive labels and the labels are grouped to conform to the canonical order of C_{ij}'s. Let us call this transformed sample the standard form of the sample. Since the algorithm $\mathcal{A}$ is symmetric and the property is defined by a permutation-invariant function, such a transformation does not affect the output of $\mathcal{A}$. So, we can further assume that we always present the sample to the algorithm in the standard form.

It is clear that given a sample, we can easily write down the fingerprint of the sample. Moreover, given the fingerprint of a sample, we can always construct a sample (S_1, S_2) in the standard form using the following algorithm: (1) Initialize S_1 and S_2 to be empty, and $e = 1$, (2) for every C_{ij} in the canonical order, and for $C_{ij} = k_{ij}$ times, include i and j copies of the element e in S_1 and S_2, respectively, then increment e. This algorithm shows a one-to-one and onto correspondence between all possible sample sets in the standard form and all possible $\{C_{ij}\}_{0 \leq i,j \leq s}$ values.

Consider the algorithm $\mathcal{A}'$ that takes the fingerprint of a sample as input. Next, by using algorithm from above, algorithm $\mathcal{A}'$ constructs the sample in the standard form. Finally, $\mathcal{A}'$ outputs what $\mathcal{A}$ outputs on this sample. □

Remark 3.18 *Note that the definition of the fingerprint from Definition 3.16 can be generalized for a collection of k sample sets from k distributions for any k. An analogous lemma to Lemma 3.17 can be proven for testing algorithms for k-ary*

properties of distributions defined by a permutation-invariant function. We fixed $k = 2$ *for ease of notation and because we will use this specific case later.*

3.1.4 A lower bound on sample complexity of testing closeness

In this section, we give a proof of a lower bound on the sample complexity of testing closeness in L_1 distance as a function of the size, denoted by n, of the domain of the distributions.

Theorem 3.19 *Given any algorithm using only* $o(n^{2/3})$ *samples from two discrete black-box distributions over* $[n]$ *for all sufficiently large* n*, there exist distributions* $\mathbf{p}$ *and* $\mathbf{q}$ *with* L_1 *distance 1 such that the algorithm will be unable to distinguish the case where one distribution is* $\mathbf{p}$ *and the other is* $\mathbf{q}$ *from the case where both distributions are* $\mathbf{p}$*.*

PROOF: By Lemma 3.15, we restrict our attention to symmetric algorithms. Fix a testing algorithm $\mathcal{A}$ that uses $o(n^{2/3})$ samples from each of the input distributions. Next, we define the distributions $\mathbf{p}$ and $\mathbf{q}$ from the theorem statement. Note that these distributions do not depend on $\mathcal{A}$.

Let us assume, without loss of generality, that n is a multiple of four and $n^{2/3}$ is an integer. We define the distributions $\mathbf{p}$ and $\mathbf{q}$ as follows: (1) For $1 \leq i \leq n^{2/3}$, $p_i = q_i = \frac{1}{2n^{2/3}}$. We call these elements the **heavy** elements. (2) For $n/2 < i \leq 3n/4$, $p_i = \frac{2}{n}$ and $q_i = 0$. We call these element the **light** elements of $\mathbf{p}$. (3) For $3n/4 < i \leq n$, $q_i = \frac{2}{n}$ and $p_i = 0$. We call these elements the **light** elements of $\mathbf{q}$. (4) For the remaining i's, $p_i = q_i = 0$.

The L_1 distance of $\mathbf{p}$ and $\mathbf{q}$ is 1. Now, consider the following two cases:

Case 1: The algorithm is given access to two black-box distributions: both of which output samples according to the distribution $\mathbf{p}$.

Case 2: The algorithm is given access to two black-box distributions: the first one outputs samples according to the distribution $\mathbf{p}$ and the second one outputs samples according to the distribution $\mathbf{q}$.

We show that a symmetric algorithm with sample complexity $o(n^{2/3})$ can not distinguish between these two cases. By Lemma 3.15, the theorem follows.

When restricted to the heavy elements, both distributions are identical. The only difference between $\mathbf{p}$ and $\mathbf{q}$ comes from the light elements, and the crux of the proof will be to show that this difference will not change the relevant statistics in a statistically significant way. We do this by showing that the only really relevant statistic is the number of elements that occur exactly once from each distribution. We then show that this statistic has a very similar distribution when generated by Case 1 and Case 2, because the expected number of such elements that are light is much less than the standard deviation of the number of such elements that are heavy.

We would like to have the frequency of each element be independent of the frequencies of the other elements. To achieve this, we assume that algorithm $\mathcal{A}$ first chooses two integers s_1 and s_2 independently from a Poisson distribution with the parameter $\lambda = s = o(n^{2/3})$. The Poisson distribution with the positive parameter λ has the probability mass function $p(k) = \exp(-\lambda)\lambda^k/k!$. Then, after taking s_1 samples from the first distribution and s_2 samples from the second distribution, $\mathcal{A}$ decides whether to accept or reject the distributions. In the following, we show that $\mathcal{A}$ cannot distinguish between Case 1 and Case 2 with success probability at least 2/3. Since both s_1 and s_2 will have values larger than $s/2$ with probability at

least $1 - o(1)$ and we will show an upper bound on the statistical distance of the distributions of two random variables (i.e., the distributions on the samples), it will follow that no symmetric algorithm with sample complexity $s/2$ can.

Let F_i be the random variable corresponding to the number of times the element i appears in the sample from the first distribution. Define G_i analogously for the second distribution. It is well known that F_i is distributed identically to the Poisson distribution with parameter $\lambda = sr$, where r is the probability of element i (cf., Feller ([11], p. 216). Furthermore, it can also be shown that all F_i's are mutually independent. Thus, the total number of samples from the heavy elements and the total number of samples from the light elements are independent.

Recall the definition of the fingerprint of a sample from Section 3.1.3. The random variable C_{ij}, denotes the number of elements that appear exactly i times in the sample from the first distribution and exactly j times in the sample from the second distribution. For the rest of the proof, we shall assume that the algorithm is only given the fingerprint of the sample. The theorem follows by Lemma 3.17.

The proof will proceed by showing that the distributions on the fingerprint when the samples come from Case 1 or Case 2 are indistinguishable. The following lemma shows that with high probability, it is only the heavy elements that contribute to the random variables C_{ij} for $i + j \geq 3$.

Lemma 3.20 *(1) With probability $1 - o(1)$, at most $o(s)$ of the heavy elements appear at least three times in the combined sample from both distributions. (2) With probability $1 - o(1)$, none of the light elements appear at least three times in the combined sample from both distributions.*

PROOF: Fix a heavy element i of probability $\frac{1}{2n^{2/3}}$. Recall that F_i and G_i denote the number of times this element appears from each distribution. The sum of the

probabilities of the samples in which element i appears at most twice is

$$\rho = \exp(-s/n^{2/3})(1 + \frac{s}{n^{2/3}} + \frac{s^2}{2n^{4/3}}).$$

By using the approximation $e^{-x} = 1 - x + x^2/2$, we can show that $1 - \rho = O(s^3/n^2)$. By linearity of expectation, we expect to have $o(s)$ heavy elements that appear at least three times. For the light elements, an analogous argument shows that $o(1)$ light elements appear at least three times. The lemma follows by Markov's inequality. □

Let D_1 and D_2 be the distributions on all possible fingerprints when samples come from Case 1 and Case 2, respectively. The rest of the proof proceeds as follows. We first construct two processes T_1 and T_2 that generate distributions on fingerprints such that T_1 is statistically close to D_1 and T_2 is statistically close to D_2. Then, we prove that the distributions T_1 and T_2 are statistically close. Hence, the theorem follows by the indistinguishability of D_1 and D_2.

Each process has two phases. The first phase is the same in both processes. They randomly generate the frequency counts for each heavy element i using the random variables F_i and G_i defined above. The processes know which elements are heavy and which elements are light, although any distinguishing algorithm does not. This concludes the first phase of the processes.

In the second phase, process T_i determines the frequency counts of each light element according to Case i. If any light element is given a total frequency count of at least three during this step, the second phase of the process is restarted from scratch.

Since the frequency counts for all elements are determined at this point, both process output the fingerprint of the sample they have generated.

Lemma 3.21 *The output of T_1, viewed as a distribution, has L_1 distance $o(1)$ to D_1. The output of T_2, viewed as a distribution, has L_1 distance $o(1)$ to D_2.*

PROOF: The distribution that T_i generates is the distribution D_i conditioned on the event that all light elements appear at most twice in the combined sample. Since this conditioning holds true with probability at least $1 - o(1)$ by Lemma 3.20, $|T_i - D_i| \leq o(1)$. □

Lemma 3.22 $|T_1 - T_2| \leq 1/6$.

PROOF: By the generation process, the L_1 distance between T_1 and T_2 can only arise from the second phase. We show that the second phases of the processes do not generate an L_1 distance larger than $1/6$.

For any variable C_{ij} of the fingerprint, the number of heavy elements that contribute to C_{ij} is independent of the number of light elements that contribute to C_{ij}. Let H be the random variable denoting the number of heavy elements that appear exactly once from each distribution. Let L be the random variable denoting the number of light elements that appear exactly once from each distribution. In Case 1, C_{11} is distributed identically to $H + L$, whereas, in Case 2, C_{11} is distributed identically to H.

Let $\mathcal{C} \stackrel{\text{def}}{=} \{C_{ij}\}_{i,j}$ and $\mathcal{C}^+ \stackrel{\text{def}}{=} \mathcal{C} \setminus \{C_{10}, C_{11}, C_{01}, C_{00}\}$. Since $\sum_{i,j} C_{ij} = n$, without loss of generality, we omit C_{00} in the rest of the discussion. Define $C_{1*} = \sum_j C_{1j}$ and $C_{*1} = \sum_i C_{i1}$. We use the notation $\Pr_{T_i}[\mathcal{C}']$ to denote the probability that T_i generates the random variable $\mathcal{C}'$ (defined on the fingerprint). We will use the fact that for any $\mathcal{C}^+, C_{1*}, C_{*1}$, $\Pr_{T_1}[\mathcal{C}^+, C_{1*}, C_{*1}] = \Pr_{T_2}[\mathcal{C}^+, C_{1*}, C_{*1}]$ in the following calculation. This fact follows from the conditioning that T_i generates on the respective D_i, namely, the condition that it is only the heavy elements that appear

at least three times. Thus, only the heavy elements contribute to the variables C_{ij}, for $i + j \geq 3$, so the distribution on this part of the fingerprint is identical in both cases. The probability that a light element contributes to the random variable C_{20} conditioned on the event that it does not appear more than twice is exactly the probability that it appears twice from the first distribution. Therefore, C_{20} is also identically distributed (conditioned on C_{ij}'s for $i + j \geq 3$) in both cases by the fact that the contribution of the light elements to C_{20} is independent of that of the heavy elements. An analogous argument applies to C_{02}, C_{1*} and C_{*1}. So, we get

$$\begin{aligned}
|T_1 - T_2| &= \sum_{\mathcal{C}} |\Pr_{T_1}[\mathcal{C}] - \Pr_{T_2}[\mathcal{C}]| \\
&= \sum_{\mathcal{C}^+, C_{1*}, C_{*1}} \Pr_{T_1}[\mathcal{C}^+, C_{1*}, C_{*1}] \\
&\qquad \sum_{h,k,l \geq 0} |\Pr_{T_1}[(C_{11}, C_{10}, C_{01}) = (h,k,l) | \mathcal{C}^+, C_{1*}, C_{*1}] \\
&\qquad\qquad - \Pr_{T_2}[(C_{11}, C_{10}, C_{01}) = (h,k,l) | \mathcal{C}^+, C_{1*}, C_{*1}]| \\
&= \sum_{\mathcal{C}^+, C_{1*}, C_{*1}} \Pr_{T_1}[\mathcal{C}^+, C_{1*}, C_{*1}] \\
&\qquad \sum_{h \geq 0} |\Pr_{T_1}[C_{11} = h | \mathcal{C}^+, C_{1*}, C_{*1}] - \Pr_{T_2}[C_{11} = h | \mathcal{C}^+, C_{1*}, C_{*1}]| \\
&= \sum_{h \geq 0} |\Pr[H = h] - \Pr[H + L = h]|
\end{aligned}$$

The third line follows since C_{10} and C_{01} are determined once $\mathcal{C}^+, C_{1*}, C_{*1}, C_{11}$ are determined. In the rest of the proof, we show that the fluctuations in H dominate the magnitude of L.

Let ξ_i be the indicator random variable that takes value 1 when element i appears exactly once from each distribution. Then, $H = \sum_{\text{heavy } i} \xi_i$. By the assumption about the way samples are generated, the ξ_i's are independent. Therefore, H is

distributed identically to the binomial distribution on the sum of $n^{2/3}$ Bernoulli trials with success probability $\Pr[\xi_i = 1] = \exp(-s/n^{2/3})(s^2/4n^{4/3})$. An analogous argument shows that L is distributed identically to the binomial distribution with parameters $n/4$ and $\exp(-4s/n)(4s^2/n^2)$.

As n grows large enough, both H and L can be approximated well by normal distributions. That is,

$$\Pr[H = h] \to \frac{1}{\sqrt{2\pi}\sigma_H} \exp(-(h - \mathrm{E}[H])^2/2\mathrm{Var}[H])$$

as $n \to \infty$. Therefore, by the independence of H and L, $H + L$ is also approximated well by a normal distribution.

Thus, $\Pr[H = h] = \Omega(1/\sigma_H)$ over an interval I_1 of length $\Omega(\sigma_H) = \Omega(s/n^{1/3})$ centered at $\mathrm{E}[H]$. Similarly, $\Pr[H + L = h] = \Omega(1/\sigma_{H+L})$ over an interval I_2 of length $\Omega(\sigma_{H+L})$ centered at $\mathrm{E}[H + L]$. Since $\mathrm{E}[H + L] - \mathrm{E}[H] = \mathrm{E}[L] = O(s^2/n) = o(s/n^{1/3})$, $I_1 \cap I_2$ is an interval of length $\Omega(\sigma_H)$. Therefore,

$$\sum_{h \in I_1 \cap I_2} |\Pr[H = h] - \Pr[H + L = h]| \leq o(1)$$

because for $h \in I_1 \cap I_2$, $|\Pr[H = h] - \Pr[H + L = h]| = o(1/\sigma_H)$. We can conclude that $\sum_h |\Pr[H = h] - \Pr[H + L = h]|$ is less than $1/6$ after accounting for the probability mass of H and $H + L$ outside $I_1 \cap I_2$. □

The theorem follows by Lemma 3.21 and Lemma 3.22. □

3.1.5 An application of closeness test to Markov chains

In [5], we show that closeness tests in L_1 norm can be used to test mixing properties of Markov chains. We show how to test whether iterating a Markov chain for t steps causes it to reach a distribution close to the stationary distribution. We then investigate two notions of being *close* to a rapidly mixing Markov chain that fall

within the framework of property testing, and show how to test that a Markov chain is close to a Markov chain that mixes in t steps by following only $\tilde{O}(tn^{2/3})$ edges, which is sublinear in the size of any reasonable representation of a Markov chain. In the case of Markov chains that come from directed graphs and pass our test, our theorems show the existence of a directed graph that is close to the original one and rapidly mixing.

Goldreich and Ron [16] give a test which they conjecture can be used for testing whether a regular graph is close to being an expander. By close, they mean that by changing a small fraction of the edges one can turn it into an expander. Mixing and expansion are known to be related [25], but our techniques only apply to the mixing properties of random walks on directed graphs, since the notion of closeness we use does not preserve the symmetry of the adjacency matrix. The conductance [25] of a graph is also known to be closely related to expansion and rapid-mixing properties of the graph [25, 18]. Frieze and Kannan [13] show that, given a graph G with n vertices and α, one can approximate the conductance of G to within additive error α in time $O(n2^{\tilde{O}(1/\alpha^2)})$. Their techniques also yield an $O(2^{\mathrm{poly}(1/\epsilon)})$-time test which determines whether an adjacency matrix of a graph can be changed in at most ϵ fraction of the locations to get a graph with high conductance. However, for the purpose of testing whether an n-vertex, m-edge graph is rapid mixing, we would need to approximate its conductance to within $\alpha = O(m/n^2)$; thus only when $m = \Theta(n^2)$ would it run in $O(n)$ time. Our test is more efficient than algorithms whose behavior is mathematically justified at every sparsity level. For the technical exposition and a more detailed discussion of related work, see [5].

3.2 Testing identity of distributions

In this section, we assume that the distributions $\mathbf{p}$ and $\mathbf{q}$ are over $[n]$, where $\mathbf{p}$ is a black-box distribution and $\mathbf{q}$ is explicitly given. The task is to distinguish the case where $\mathbf{p} = \mathbf{q}$ from the case $|\mathbf{p} - \mathbf{q}| > \epsilon$ using as few samples (from $\mathbf{p}$) as possible. We show that this can be done using $\tilde{O}(\sqrt{n}\text{poly}(\epsilon^{-1}))$ samples. Our algorithm (first appeared in [4]) reads the explicit distribution q entirely. By the lower bound we show in Section 3.2.1, this sample complexity optimal up to polylogarithmic factors.

The main technical idea is to use bucketing (Section 2.3) to reduce this problem to that of testing that each of the restrictions of the input distribution is approximately uniform. We first bucket the given distribution $\mathbf{q}$; recall that bucketing gives a partition $(R_0, \ldots, R_k)$ of the domain so that the distribution $\mathbf{q}$ is close to uniform in each of the partitions R_i (Lemma 2.14). For each partition R_i, we sample $\mathbf{p}$ and test if $(\mathbf{p}^{\downarrow R_i})$ is close to uniform on R_i. This uniformity test on the restriction to each R_i is accomplished using a similar argument to that of [16] that was presented in Section 3.1.1.

Now, we give the complete algorithm to test if a black-box distribution $\mathbf{p}$ is close to an explicitly specified distribution $\mathbf{q}$.

Algorithm *TestIdentity*$(\mathbf{p}, \mathbf{q}, n, \epsilon)$

(1) $\mathcal{R} \stackrel{\text{def}}{=} (R_0, \ldots, R_k) = \textit{Bucket}(\mathbf{q}, n, \epsilon/2)$.

(2) Let M be a set of $O(\sqrt{n}\epsilon^{-2} \log n)$ samples from $\mathbf{p}$.

(3) For each partition R_i do

(4) Let $M_i = M \cap R_i$ (preserving repetitions); let $\ell_i = |M_i|$ (counting also repetitions).

(5) If $\mathbf{q}(R_i) \geq \epsilon/k$ then

(6) If $\ell_i < O(\sqrt{n}\epsilon^{-2})$ then FAIL.

(7) Estimate $\|(\mathbf{p}^{\downarrow R_i})\|^2$ using M_i.

(8) If $\|(\mathbf{p}^{\downarrow R_i})\|^2 > (1+\epsilon^2)/|R_i|$ then FAIL.

(9) If $|(\mathbf{p}^{\langle \mathcal{R} \rangle}) - (\mathbf{q}^{\langle \mathcal{R} \rangle})| > \epsilon$ then FAIL.

(10) PASS.

Theorem 3.23 *Algorithm TestIdentity*$(\mathbf{p}, \mathbf{q}, n, \epsilon)$ *is such that: (1) if* $|\mathbf{p} - \mathbf{q}| \leq \frac{\epsilon^2}{4\sqrt{n}\log n}$, *it outputs PASS with constant probability and (2) if* $|\mathbf{p} - \mathbf{q}| > 6\epsilon$, *it outputs FAIL with constant probability. The algorithm uses* $\tilde{O}(\sqrt{n}\text{poly}(\epsilon^{-1}))$ *samples.*

PROOF: Step (9) can be done by using brute force to distinguish between $|(\mathbf{p}^{\langle \mathcal{R} \rangle}) - (\mathbf{q}^{\langle \mathcal{R} \rangle})| > \epsilon$ and $|(\mathbf{p}^{\langle \mathcal{R} \rangle}) - (\mathbf{q}^{\langle \mathcal{R} \rangle})| < \frac{1}{2}\epsilon$. This does not take a significant number of additional samples, as k is logarithmic in n by the definition of bucketing.

Note that by Chernoff bounds, the probability of failing in step (6) can be made sufficiently small, unless there is a large difference between $\mathbf{p}(R_i)$ and $\mathbf{q}(R_i)$ for some i. Suppose that the algorithm outputs PASS. This implies that for each partition R_i for which steps (6)-(8) were performed (which are those for which $\mathbf{q}(R_i) \geq \epsilon/k$), we have $\|(\mathbf{p}^{\downarrow R_i})\|^2 \leq (1+\epsilon^2)/|R_i|$. From Lemma 2.16 we get that for each of these R_i, $|(\mathbf{p}^{\downarrow R_i}) - (\mathbf{q}^{\downarrow R_i})| \leq 2\epsilon$.

We also have that the sum of $\mathbf{q}(R_i)$ over all R_i for which steps (6)-(8) were skipped is at most ϵ. Also, $|(\mathbf{p}^{\langle \mathcal{R} \rangle}) - (\mathbf{q}^{\langle \mathcal{R} \rangle})| \leq \epsilon$ by step (9); so the total difference between $\mathbf{p}$ and $\mathbf{q}$ over these partitions sums up to no more than 3ϵ. Adding this to the 3ϵ difference over the partitions that were not skipped in steps (6)-(8) (given by applying Lemma 2.12 with $|(\mathbf{p}^{\downarrow R_i}) - (\mathbf{q}^{\downarrow R_i})| \leq 2\epsilon$ and $|(\mathbf{p}^{\langle \mathcal{R} \rangle}) - (\mathbf{q}^{\langle \mathcal{R} \rangle})| \leq \epsilon$), we get that $|\mathbf{p} - \mathbf{q}| \leq 6\epsilon$.

On the other hand, suppose $|\mathbf{p} - \mathbf{q}| < \frac{\epsilon^2}{4\sqrt{n}\log n}$. Using Lemma 2.13 for all partitions R_i with $\mathbf{q}(R_i) \geq \epsilon/k$, we have $|(\mathbf{p}^{\downarrow R_i}) - (\mathbf{q}^{\downarrow R_i})| < \epsilon/(2\sqrt{n})$. In terms of

$\|\cdot\|$, this implies $\|(\mathbf{p}^{\downarrow R_i}) - (\mathbf{q}^{\downarrow R_i})\|^2 < \epsilon^2/(4n) < \epsilon^2/(4|R_i|)$. Since from Lemma 2.14, $\|(\mathbf{q}^{\downarrow R_i}) - U_{R_i}\|^2 < \epsilon^2/(4|R_i|)$, then by triangle inequality, $\|(\mathbf{p}^{\downarrow R_i}) - U_{R_i}\|^2 \leq (\|(\mathbf{p}^{\downarrow R_i}) - (\mathbf{q}^{\downarrow R_i})\| + \|(\mathbf{q}^{\downarrow R_i}) - U_{R_i}\|)^2 \leq \epsilon^2/|R_i|$. So by Lemma 2.17, $\|(\mathbf{p}^{\downarrow R_i})\|^2 = \|(\mathbf{p}^{\downarrow R_i}) - U_{R_i}\|^2 + \|U_{R_i}\|^2 \leq (1+\epsilon^2)/|R_i|$. Therefore the algorithm will pass with high probability on all such partitions; it is also not hard to see that the algorithm will pass step (9) as well.

The sample complexity is $\tilde{O}(\sqrt{n}\epsilon^{-2})$ from step (2), which dominates the sample complexity of step (9) (no other samples are taken throughout the algorithm). □

3.2.1 A lower bound on sample complexity of testing identity

We prove a lower bound on sample complexity of testing identity in a special case, namely, testing uniformity of a black-box distribution. The proof hinges on the famous Birthday Problem.

Theorem 3.24 *Given any algorithm using only $o(\sqrt{n})$ samples from a discrete black-box distribution over $[n]$ for all sufficiently large n, there exist distributions $\mathbf{p}$ and $\mathbf{q}$ with L_1 distance 1 such that the algorithm will be unable to distinguish the case where the input distribution is $\mathbf{p}$ from the case where the input distribution is $\mathbf{q}$.*

PROOF: By Lemma 3.15, it suffices to consider only the symmetric algorithms. Let $\mathcal{A}$ be a symmetric algorithm that uses $s = o(\sqrt{n})$ samples. Let distribution $\mathbf{p}$ be the uniform distribution over $[n]$, and let distribution $\mathbf{q}$ be the uniform distribution over $[n/2]$. By the Birthday Problem, it is well know that there exist a constant $c < 1$ such that the probability that we get same element twice before taking $c\sqrt{n}$

samples from a uniform distribution over $[n]$ is at most c^2. Using such an argument, the probability of the existence of a repetition after taking s samples from $\mathbf{q}$ can be upper bounded by

$$\sum_{t=1}^{s} \frac{t-1}{n/2} = \frac{2}{n} \sum_{t=1}^{s} (t-1) = \frac{2}{n} \binom{s}{2} = o(1).$$

A similar argument holds for samples taken from $\mathbf{p}$ as well. Therefore, s samples from $\mathbf{p}$ or $\mathbf{q}$ will consist of s distinct elements with probability at least $1 - o(1)$. Since the algorithm $\mathcal{A}$ is symmetric, we can conclude that $\mathcal{A}$ cannot distinguish $\mathbf{p}$ and $\mathbf{q}$ with probability higher than $\frac{1}{2} + o(1)$. □

Chapter 4

Testing independence of distributions

Consider a joint distribution over the product space of two sets, where a sample from this distribution is a pair. One might want to know whether the two components of a sample from this distribution are correlated. In this chapter, we study the problem of testing independence of a joint distribution over a product space using only samples from the distribution. Namely, we want to test whether the distribution over the first component is independent from the distribution over the second component with no additional assumptions on the structure of the joint distribution.

For the sake of presentation, we abandon the vector notation to represent distributions throughout this chapter and use capital letters to name the distributions and the function notation to denote the probability density functions. The probability density function of a distribution $\mathbf{A}$ is denoted by $\mathbf{A}(\cdot, \cdot)$.

Checking independence of a joint distribution over $[n] \times [m]$is a central question in statistics and there exist many different techniques for attacking it (see [20]). Classical tests such as the χ^2 or Kolmogorov-Smirnoff work well when n and m are

small, but for large n, m these tests require more than $n \cdot m$ samples.

We develop a general algorithm for testing independence problem (first appeared in [4]) with sublinear sample complexity (in the size of $[n] \times [m]$). This is the first sublinear time test which makes no assumptions about the structure of the distribution. Our test uses $\tilde{O}(n^{2/3}m^{1/3}\text{poly}(\epsilon^{-1}))$ samples, assuming without loss of generality that $n \geq m$, and distinguishes between the case that $\mathbf{A} = \mathbf{A}_1 \times \mathbf{A}_2$ and that for all $\mathbf{A}_1, \mathbf{A}_2$, $|\mathbf{A} - \mathbf{A}_1 \times \mathbf{A}_2| \geq \epsilon$.[1] Here, $\mathbf{A}_1$ and $\mathbf{A}_2$ are distributions over $[n]$ and $[m]$, respectively. We also show this bound is tight up to polylogarithmic factors.

4.1 Independence and approximate independence

Let $\mathbf{A}$ be a distribution over $[n] \times [m]$. Then, $\pi_i\mathbf{A}$, for $i \in \{1, 2\}$, denote the marginal distributions obtained by projecting $\mathbf{A}$ into the i-th component. We say that $\mathbf{A}$ is **independent** if, the distributions $\pi_1\mathbf{A}$ and $\pi_2\mathbf{A}$ are independent, equivalently if for all $i \in [n]$ and $j \in [m]$,

$$\mathbf{A}(i, j) = (\pi_1\mathbf{A})(i) \cdot (\pi_2\mathbf{A})(j)$$

or simply $\mathbf{A} = (\pi_1\mathbf{A}) \times (\pi_2\mathbf{A})$.

We say that $\mathbf{A}$ is **ϵ-independent** if there is a distribution $\mathbf{B}$ that is independent and $|\mathbf{A} - \mathbf{B}| \leq \epsilon$. Otherwise, we say $\mathbf{A}$ is *not ϵ-independent* or *ϵ-far from being independent*.

The following claim shows that an ϵ-independent distribution is close to the independent distribution generated by the product of the marginal distributions.

[1] The notation $\mathbf{B} \times \mathbf{C}$ denotes the joint distribution obtained by choosing the first component according to $\mathbf{B}$ and the second component according to $\mathbf{C}$.

Claim 4.1 *Let* $\mathbf{A}, \mathbf{B}$ *be distributions over* $S \times T$*. If* $\mathbf{B}$ *is independent, then*

$$|\mathbf{A} - (\pi_1\mathbf{A}) \times (\pi_2\mathbf{A})| \leq 3|\mathbf{A} - \mathbf{B}|.$$

Claim 4.1 follows from the following lemmas.

Lemma 4.2 ([24]) *Let* $\mathbf{A}_1, \mathbf{B}_1$ *be distributions over* S *and* $\mathbf{A}_2, \mathbf{B}_2$ *be distributions over* T*. Then,* $|\mathbf{A}_1 \times \mathbf{B}_1 - \mathbf{A}_2 \times \mathbf{B}_2| \leq |\mathbf{A}_1 - \mathbf{B}_1| + |\mathbf{A}_2 - \mathbf{B}_2|$.

PROOF: Using the triangle inequality, we get

$$\begin{aligned}
|\mathbf{A}_1 \times \mathbf{A}_2 - \mathbf{B}_1 \times \mathbf{B}_2| &= \sum_{i\in S}\sum_{j\in T} |\mathbf{A}_1(i)\mathbf{A}_2(j) - \mathbf{B}_1(i)\mathbf{B}_2(j)| \\
&\leq \sum_{i\in S}\sum_{j\in T} |\mathbf{A}_1(i)\mathbf{A}_2(j) - \mathbf{A}_1(i)\mathbf{B}_2(j)| + \sum_{i\in S}\sum_{j\in T} |\mathbf{A}_1(i)\mathbf{B}_2(j) - \mathbf{B}_1(i)\mathbf{B}_2(j)| \\
&= \sum_{i\in S}\sum_{j\in T} \mathbf{A}_1(i) \cdot |\mathbf{A}_2(j) - \mathbf{B}_2(j)| + \sum_{i\in S}\sum_{j\in T} \mathbf{B}_2(j) \cdot |\mathbf{A}_1(i) - \mathbf{B}_1(i)| \\
&= \sum_{j\in T} |\mathbf{A}_2(j) - \mathbf{B}_2(j)| + \sum_{i\in S} |\mathbf{A}_1(i) - \mathbf{B}_1(i)| \\
&= |\mathbf{A}_1 - \mathbf{B}_1| + |\mathbf{A}_2 - \mathbf{B}_2|
\end{aligned}$$

□

Lemma 4.3 *Let* $\mathbf{A}, \mathbf{B}$ *be distributions over* $S \times T$*. Then,* $|\pi_1\mathbf{A} - \pi_1\mathbf{B}| \leq |\mathbf{A} - \mathbf{B}|$ *and* $|\pi_2\mathbf{A} - \pi_2\mathbf{B}| \leq |\mathbf{A} - \mathbf{B}|$.

PROOF: By unfolding the definitions of the marginal distributions and the triangle inequality, we get

$$\begin{aligned}
|\pi_1\mathbf{A} - \pi_1\mathbf{B}| &= \sum_{i\in S} |\pi_1\mathbf{A}(i) - \pi_1\mathbf{B}(i)| \\
&= \sum_{i\in S} |\sum_{j\in T} \mathbf{A}(i,j) - \sum_{j\in T} \mathbf{B}(i,j)| \\
&= \sum_{i\in S} |\sum_{j\in T} (\mathbf{A}(i,j) - \mathbf{B}(i,j))| \\
&\leq \sum_{i\in S}\sum_{j\in T} |\mathbf{A}(i,j) - \mathbf{B}(i,j)| \\
&= |\mathbf{A} - \mathbf{B}|.
\end{aligned}$$

An analogous argument for the second part applies. □

PROOF: (of Claim 4.1) Since $\mathbf{B}$ is independent, $\mathbf{B} = (\pi_1\mathbf{B}) \times (\pi_2\mathbf{B})$. Using the triangle inequality, Lemma 4.2 and Lemma 4.3, we get

$$\begin{aligned} |\mathbf{A} - (\pi_1\mathbf{A}) \times (\pi_2\mathbf{A})| &\leq |\mathbf{A} - \mathbf{B}| + |\mathbf{B} - (\pi_1\mathbf{A}) \times (\pi_2\mathbf{A})| \\ &= |\mathbf{A} - \mathbf{B}| + |(\pi_1\mathbf{B}) \times (\pi_2\mathbf{B}) - (\pi_1\mathbf{A}) \times (\pi_2\mathbf{A})| \\ &\leq 3|\mathbf{A} - \mathbf{B}|. \end{aligned}$$

□

Claim 4.1 shows that in order to test whether a joint distribution is independent, one can use the marginal distributions as references. If the distribution is independent, then it is clearly equal to the product of the marginal distributions. And, if it is far from this product, then it is far from any independent distribution.

4.2 A filtering scheme

In Section 2.3, we introduce bucketing as a general tool which decomposes an arbitrary probability distribution into a collection of distributions that are almost uniform. Later, we used this tool in Section 3.2 to reduce the general problem of testing identity to the special case of testing uniformity. Now, we introduce a new tool, a filtering scheme, that will be used in conjunction with bucketing to test independence of joint distributions. These two tools together will provide us a reduction from the general problem of testing independence to the special case of testing independence of a joint distribution that is close to the uniform distribution.

Informally, an $(\mathbf{A}, \mathbf{B})$**-filter** is a (randomized) black-box sampler that can access a black-box distribution $\mathbf{A}$ over $S \times T$ and can simulate a distribution $\mathbf{B}$ over $S \times T$ such that certain 'properties' of $\mathbf{B}$ are related to 'properties' of $\mathbf{A}$. We use the filter

in batch mode, i.e., given an input parameter t, we output t samples according to $\mathbf{B}$. The filter is free to use $\mathbf{A}$ however it wants — it can either sample $\mathbf{A}$ in a single 'preprocessing' stage and use these results or it can dynamically sample $\mathbf{A}$ or do both.

An $(\mathbf{A}, \mathbf{B})$-filter is specified in terms of the relationship between the properties of $\mathbf{B}$ and those of $\mathbf{A}$. For example, we will construct a filter which, under certain conditions, produces a distribution $\mathbf{B}$ that is uniform when $\mathbf{A}$ is independent, and is far from being uniform when $\mathbf{A}$ is far from being independent. This will allow us to distinguish distributions that are independent from those that are far from being independent. The other parameter of interest is the sample complexity of the filter, which is the total number of samples from $\mathbf{A}$ it uses for a given t and $S \times T$.

We first show that there is a filter that takes a distribution $\mathbf{A}$ over $S \times T$ for which the first component is close to uniform, and produces a new distribution which is close to the original one, and for which the first component is uniform. Moreover, this filter preserves independence.

Lemma 4.4 *There exists an $(\mathbf{A}, \mathbf{B})$-filter for distributions over $S \times T$ such that for any t, with high probability, (1) if $\mathbf{A} = (\pi_1\mathbf{A}) \times (\pi_2\mathbf{A})$ then $\mathbf{B} = U_S \times (\pi_2\mathbf{A})$, and (2) if $|\pi_1\mathbf{A} - U_S| \leq \epsilon/4$ then $|\mathbf{B} - \mathbf{A}| \leq \epsilon$. The sample complexity of the filter is $O(\max\{|S|, t\} \log^3 \max\{|S|, t\})$.*

PROOF: First, we describe the construction of the filter. Let t be given and let $\ell = O(\lceil t/|S| \rceil \log |S| \log t)$. The filter maintains a data structure which for every $i \in S$, contains a list L_i of ℓ elements of T. Each list starts out empty and is filled according to the following steps:

(1) Obtain $O(\max\{|S|, t\} \log^3 \max\{|S|, t\})$ samples from $\mathbf{A}$ and for each sample (i, j) from $\mathbf{A}$, add j to L_i if $|L_i| \leq \ell$.

(2) For each $i \in S$, if $|L_i| < \ell$, then discard L_i. In this case, obtain ℓ more samples from $\mathbf{A}$ and for each sample (k, j) from $\mathbf{A}$, add j to L_i.

For $i \in S$, let $\mathbf{B}_i$ be distributed identically to $\pi_2(\mathbf{A}^{\downarrow\{i\}\times T})$ if L_i was not discarded in step (2) and identically to $\pi_2\mathbf{A}$ otherwise. Thus, L_i contains ℓ independent samples of $\mathbf{B}_i$.

Next, we describe the operation of the filter. Upon a sample request, the filter generates a random $i \in_R S$. If $|L_i| > 0$, then the filter picks the first element j in L_i, outputs (i, j), and deletes the first element in L_i. If $|L_i| = 0$, then the filter gets a sample (i', j') from $\mathbf{A}$ and outputs (i, j').

First, notice that with high probability (via a Chernoff bound), no L_i becomes empty in any of the t requests for samples. Also, it is clear that the output of the filter is the distribution defined by generating a uniform $i \in S$ and then simulating the corresponding $\mathbf{B}_i$. The exact distribution of $\mathbf{B}$ may depend on the outcome of the preprocessing stage of the filter, but we show that with high probability $\mathbf{B}$ satisfies the assertions of the lemma.

For the first assertion, note that if $\mathbf{A} = (\pi_1\mathbf{A}) \times (\pi_2\mathbf{A})$, then the second component is independent of the first component. So, $\mathbf{B}_i = \pi_2\mathbf{A}$ for every i (regardless of whether L_i was filled by step (1) or (2)). Thus, $\mathbf{B} = U_S \times (\pi_2\mathbf{A})$.

To show the second assertion, let $I = \{i \mid \pi_1\mathbf{A}(i) \geq 1/(2|S|)\}$. Another application of the Chernoff bound shows that with high probability, for every $i \in I$, $\mathbf{B}_i$ is distributed as $\pi_2((\mathbf{A}^{\downarrow\{i\}\times T}))$ (since L_i would not be discarded in step (2)). Thus, for every $i \in I$, L_i contains ℓ independent samples of $\mathbf{B}_i = \pi_2(\mathbf{A}^{\downarrow\{i\}\times T})$. Also,

since $|\pi_1(\mathbf{A}) - U_S| \leq \epsilon/4$, we have $|S \backslash I| \leq \epsilon|S|/2$. We get

$$\begin{aligned}
|\mathbf{A} - \mathbf{B}| &= \sum_{i \in I} \sum_{j \in T} |\mathbf{A}(i,j) - \mathbf{B}(i,j)| + \sum_{i \in S \backslash I} \sum_{j \in T} |\mathbf{A}(i,j) - \mathbf{B}(i,j)| \\
&\leq \sum_{i \in I} \sum_{j \in T} |\mathbf{A}(i,j) - \mathbf{B}(i,j)| + \sum_{i \in S \backslash I} \sum_{j \in T} (\mathbf{A}(i,j) + \mathbf{B}(i,j)) \\
&= \sum_{i \in I} \sum_{j \in T} \pi_2(\mathbf{A}^{\downarrow \{i\} \times T})(j) \cdot \left| \pi_1 \mathbf{A}(i) - \frac{1}{|S|} \right| + \sum_{i \in S \backslash I} (\pi_1 \mathbf{A}(i) + \pi_1 \mathbf{B}(i)) \\
&\leq \sum_{i \in I} \left| \pi_1 \mathbf{A}(i) - \frac{1}{|S|} \right| + \sum_{i \in S \backslash I} \pi_1 \mathbf{A}(i) + \frac{|S \backslash I|}{|S|} \leq \frac{1}{4}\epsilon + \frac{1}{4}\epsilon + \frac{1}{2}\epsilon = \epsilon
\end{aligned}$$

□

Filters can be composed, i.e., an $(\mathbf{A}, \mathbf{C})$-filter can be combined with a $(\mathbf{C}, \mathbf{B})$-filter to give an $(\mathbf{A}, \mathbf{B})$-filter. If the sample complexity of the $(\mathbf{A}, \mathbf{C})$-filter is given by the function $f(t)$, and that of the $(\mathbf{C}, \mathbf{B})$-filter is given by $g(t)$, then the sample complexity of the combined $(\mathbf{A}, \mathbf{B})$-filter will be given by $h(t) = f(g(t))$.

Corollary 4.5 *There exists an $(\mathbf{A}, \mathbf{B})$-filter for distribution over $S \times T$ such that if $|\pi_1 \mathbf{A} - U_S| \leq \epsilon/25$, and $|\pi_2 \mathbf{A} - U_T| \leq \epsilon/25$, then with high probability, (1) $|\mathbf{B} - \mathbf{A}| \leq (24/25)\epsilon$; (2) if $\mathbf{A} = (\pi_1 \mathbf{A}) \times (\pi_2 \mathbf{A})$ then $\mathbf{B} = U_{S \times T}$; and (3) if $\mathbf{A}$ is not ϵ-independent, then $|\mathbf{B} - U_{S \times T}| \geq (1/25)\epsilon$. The sample complexity of the filter is $O(\max\{|S| + |T|, t\} \log^3 \max\{|S|, |T|, t\})$.*

Proof: We apply the $(\mathbf{A}, \mathbf{C})$-filter from Lemma 4.4 on the first component. Using this filter we obtain a distribution $\mathbf{C}$ (with high probability) such that $|\mathbf{C} - \mathbf{A}| \leq 4\epsilon/25$, $\pi_1 \mathbf{C} = U_S$, and such that $\mathbf{C}$ is independent if $\mathbf{A}$ is independent.

Now, using Lemma 4.3, $|\pi_2 \mathbf{C} - \pi_2 \mathbf{A}| \leq 4\epsilon/25$ and since by our hypothesis, $|\pi_2 \mathbf{A} - U_T| \leq \epsilon/25$, we get $|\pi_2 \mathbf{C} - U_T| \leq \epsilon/5$.

We now construct a $(\mathbf{C}, \mathbf{B})$-filter from Lemma 4.4, only this time switching components and filtering on the second component. Using this filter, we obtain a distribution $\mathbf{B}$ (with high probability) such that $|\mathbf{B} - \mathbf{C}| \leq 20\epsilon/25$ and $\pi_2 \mathbf{B} = U_T$.

Moreover, according to Lemma 4.4 if $\mathbf{A}$ is independent (and thus so are $\mathbf{C}$ and $\mathbf{B}$) then $\pi_1\mathbf{B}$ has the same distribution as $\pi_1\mathbf{C} = U_S$. Since $\pi_1\mathbf{B} = U_S, \pi_2\mathbf{B} = U_T$ and they are independent, we get that $\mathbf{B}$ is uniform on $S \times T$. Clearly, $|\mathbf{B} - \mathbf{A}| \leq |\mathbf{B} - \mathbf{C}| + |\mathbf{C} - \mathbf{A}| \leq (24/25)\epsilon$.

If $\mathbf{A}$ is not ϵ-independent, then $\mathbf{B}$, which is $(24/25)\epsilon$-close to $\mathbf{A}$, is $(1/25)\epsilon$-far from any independent distribution on $S \times T$, in particular $U_{S\times T}$. □

4.3 An algorithm for testing independence

In this section, we give an algorithm for testing independence of a distribution $\mathbf{A}$ over $[n] \times [m]$. Without loss of generality, let $n \geq m$. First of all, we present two different approaches to testing independence. These two methods have different samples complexities and are desirable in different situations.

(1) In the first method, we use the equivalence of testing independence to testing whether $\mathbf{A}$ is close to $\pi_1\mathbf{A} \times \pi_2\mathbf{A}$, which was shown by Claim 4.1. Since it is easy to generate samples of $\pi_1\mathbf{A} \times \pi_2\mathbf{A}$ given samples of $\mathbf{A}$, we can use the closeness test presented in Chapter 3. This immediately gives us a test for independence that uses $\tilde{O}(n^{2/3}m^{2/3})$ samples.

(2) In the second method, we reduce the problem of testing independence to testing independence of many distributions that have the property that they are almost uniform in each of the component. We then reduce the problem of testing independence of such a distribution to testing independence of a distribution that is uniform in each of the component. The first reduction is via bucketing (described in Section 2.3) and the second reduction uses the filtering scheme described in Section 4.2. To finish it off, testing the independence of a distribution that is uniform in each of components is equivalent to testing whether the distribution is uniform

over both components — so we use the same techniques that we used to build the algorithm for testing identity in Chapter 3. We show that for this method, the overall sample complexity is $\tilde{O}(n)$.

Then, we combine these two algorithms in an appropriate manner to exploit the different behavior. In particular, we partition the elements of $[n]$ as 'light' or 'heavy' based on their probability values in $\pi_1\mathbf{A}$.[2] We apply method (1) to the light prefixes and method (2) to the heavy prefixes. Finally, we ensure that the distributions restricted to the heavy and light prefixes are consistent. This asymmetric approach helps us achieve an optimal trade-off in the sample complexities, resulting in the $\tilde{O}(n^{2/3}m^{1/3})$ sample complexity.

Let ϵ be the threshold parameter given to the algorithm such that the the algorithm is expected to distinguish the independent distributions from distributions that are ϵ-far from being independent. Let β be such that $m = n^\beta$ and $0 < \alpha < 1$ be a parameter to be determined later. Let S' denote the set of prefixes with probability mass at least $n^{-\alpha}$; such prefixes are called **heavy**. All the other prefixes are called **light**. Formally, let

$$S' = \{i \in [n] \;\mid\; (\pi_1\mathbf{A})(i) \geq n^{-\alpha}\}.$$

Using $O(n^\alpha \epsilon^{-2} \log n)$ samples, we can estimate $(\pi_1\mathbf{A})(i)$, for $i \in S'$, by $\tilde{\mathbf{A}}_1(i)$ to within an $\epsilon/75$ factor using Theorem 2.8. Let $\tilde{S}$ be the set of all i for which $\tilde{\mathbf{A}}_1(i) \geq n^{-\alpha}/2$. Then $S' \subset \tilde{S}$ and $\tilde{S}$ does not contains any i for which $(\pi_1\mathbf{A})(i) \leq n^{-\alpha}/2$.

Our main idea is to first test that $\mathbf{A}$ is independent conditioned on the set of heavy prefixes (Section 4.3.1) and then to test that $\mathbf{A}$ is independent conditioned on the set of light prefixes (Section 4.3.2). To create these conditionings, we first

[2]We often refer to the first component of a sample from a joint distribution on pairs as the **prefix** of the sample.

distinguish (using $\tilde{O}(\epsilon^{-1})$ samples) between $(\pi_1\mathbf{A})(\tilde{S}) \geq \epsilon$ and $(\pi_1\mathbf{A})(\tilde{S}) \leq \epsilon/2$. If the latter case occurs, then the distribution conditioned on the heavy prefixes cannot contribute more than $\epsilon/2$ to $\mathbf{A}$'s distance from independence. Otherwise, if we are guaranteed that the second case does not occur, we can simulate the distribution for $(\mathbf{A}^{\downarrow\tilde{S}\times[m]})$ easily — we sample from $\mathbf{A}$ until we find a member of $\tilde{S} \times [m]$ which we output; this takes $O(\epsilon^{-1}\log(nm))$ queries with a high enough success probability. We then apply an independence test that works well for heavy prefixes to $(\mathbf{A}^{\downarrow\tilde{S}\times[m]})$.

Next we distinguish between $(\pi_1\mathbf{A})([n]\backslash\tilde{S}) \geq \epsilon$ and $(\pi_1\mathbf{A})([n]\backslash\tilde{S}) \leq \epsilon/2$. Again if the latter occurs, then the distribution conditioned on light elements can contribute at most $\epsilon/2$ to the distance from independence. Otherwise, if the latter does not occur, as before we simulate the distribution $(\mathbf{A}^{\downarrow([n]\backslash\tilde{S})\times[m]})$, and use it with a test that works well for distributions restricted to light prefixes (they will still remain light enough provided that $(\pi_1\mathbf{A})([n]\backslash\tilde{S}) \geq \epsilon/2$).

Finally, we obtain a test for independence (Section 4.3.3) by merging the testing over light and heavy prefixes and using an additional application of Theorem 3.1 to ensure the consistency of the distributions.

4.3.1 The heavy prefixes

We show that using filters, the heavy prefixes can be tested for independence using roughly $\tilde{O}((n^\alpha + m)\text{poly}(\epsilon^{-1}))$ samples. In fact, the following theorem yields a general algorithm for testing independence; it is just that the sample complexity is particularly useful in the heavy prefix case. Note that in this case $|S| = O(n^\alpha)$.

Theorem 4.6 *There is an algorithm that given a black-box distribution* $\mathbf{A}$ *over* $S \times T$*: (1) if* $\mathbf{A}$ *is independent, it outputs PASS with high probability and (2) if* $\mathbf{A}$ *is not* 3ϵ*-independent, it outputs FAIL with high probability. The algorithm uses*

$\tilde{O}((|S|+|T|)\text{poly}(\epsilon^{-1}))$ *samples.*

PROOF: Let $\tilde{\mathbf{A}}_1$ be an explicit distribution which approximates $\pi_1\mathbf{A}$. Consider the following independence test:

Algorithm *TestHeavyIndependence*$(\mathbf{A}, \tilde{\mathbf{A}}_1, \epsilon)$

(1) $\mathcal{S} \stackrel{\text{def}}{=} (S_0, S_1, \ldots, S_k) =$ *Bucket* $(\tilde{\mathbf{A}}_1, S, \epsilon/75)$.

(2) Obtain an approximation $\tilde{\mathbf{A}}_2$ of $\pi_2\mathbf{A}$ within an $\epsilon/75$ factor,
on a $\tilde{T}$ which includes all $j \in [m]$ which have probability at least $(m \log m)^{-1}$

(3) $\mathcal{T} \stackrel{\text{def}}{=} (T_0, T_1, \ldots, T_\ell) =$*Bucket* $(\tilde{\mathbf{A}}_2, \tilde{T}, \epsilon)$; add $T \backslash \tilde{T}$ to T_0.

(4) For $(S_i, T_j), i \in [k], j \in [\ell]$ do

(5) If $\mathbf{A}(S_i \times T_j) \geq \epsilon/(k\ell)$ then

(6) If $(\mathbf{A}^{\downarrow S_i \times T_j})$ is not ϵ-independent, then FAIL.

(7) If $(\mathbf{A}^{\langle \mathcal{S} \times \mathcal{T} \rangle})$ is not $\epsilon/2$-independent, then FAIL.

(8) PASS.

Note that, if needed, $\tilde{\mathbf{A}}_1$ can be obtained using $\tilde{O}|S|\text{poly}(\epsilon^{-1})$ samples. After step (2), S_0 can be ignored (as usual). The independence test in step (7) can be done by brute force, for instance, since the alphabet is only logarithmic in $|S|$ and $|T|$. Also, by bucketing, we know that $|\pi_1\mathbf{A} - U_{S_i}| \leq \epsilon/25, \forall i \in [k]$ and $|\pi_2\mathbf{A} - U_{T_j}| \leq \epsilon/25, \forall j \in [\ell]$. For deciding in step (5) whether to execute step (6), we distinguish between $\mathbf{A}(S_i \times T_j) \geq \epsilon/(k\ell)$ and $\mathbf{A}(S_i \times T_j) \leq \epsilon/(2k\ell)$, by taking $\tilde{O}(k\ell/\epsilon)$ many samples of $\mathbf{A}$ and counting how many of them are in $S_i \times T_j$. Step (6) requires sampling of $(\mathbf{A}^{\downarrow S_i \times T_j})$; this is done by repeatedly sampling $\mathbf{A}$ until a member of $S_i \times T_j$ is found. As we are assured in step (6) that $\mathbf{A}(S_i \times T_j) > \epsilon/(2k\ell)$, it suffices to take $O(\epsilon^{-1} \log^3(nm))$ samples of $\mathbf{A}$ in order to generate a single sample of $(\mathbf{A}^{\downarrow S_i \times T_j})$ (remember that k and ℓ are logarithmic in n and m).

We now present the independence test in step (6) which is used for each pair of buckets from $\mathcal{S}$ and $\mathcal{T}$.

Lemma 4.7 *There is an algorithm that given a black-box distribution* $\mathbf{A}$ *over* $S \times T$ *such that* $|\pi_1\mathbf{A} - U_S| \leq \epsilon/25, |\pi_2\mathbf{A} - U_T| \leq \epsilon/25$*: (1) if* $\mathbf{A}$ *is independent, it outputs PASS with high probability and (2) if* $\mathbf{A}$ *is not* ϵ*-close to* $U_{S\times T}$*, it outputs FAIL with high probability (in particular, only one of these cases can occur for a distribution satisfying the above conditions). The algorithm uses* $\tilde{O}((|S|+|T|)\mathrm{poly}(\epsilon^{-1}))$ *samples.*

Proof: We apply the $(\mathbf{A}, \mathbf{B})$-filter from Corollary 4.5. By its properties, if $\mathbf{A}$ is independent then $\mathbf{B} = U_{S\times T}$, and if $\mathbf{A}$ is not ϵ-close to $U_{S\times T}$, then $|\mathbf{B} - U_{S\times T}| \geq \epsilon/25$ (because $|\mathbf{A} - \mathbf{B}| \leq \frac{24}{25}\epsilon$). We can distinguish between these cases using Theorem 3.11, with $\tilde{O}(\epsilon^{-1}\sqrt{|S \times T|})$ samples from the filter, which in itself takes less than a total of $\tilde{O}(\epsilon^{-4}(|S| + |T|)\log^6(\epsilon^{-1}(|S| + |T|)))$ samples from the distribution. □ Note that in the application of Lemma 4.7, its sampling estimate should be further multiplied by $O(\epsilon^{-1}\log^3(nm))$ to get the total number of samples made from $\mathbf{A}$, because it is applied separately to the restriction of $\mathbf{A}$ to each $S_i \times T_j$.

We now return to the proof of the theorem. If $\mathbf{A}$ is independent, then for all $i \in [k], j \in [\ell]$, the restriction $(\mathbf{A}^{\downarrow S_i \times T_j})$ is independent so steps (4)–(6) pass (remember that Lemma 4.7 ensures that independent distributions pass step (6)). In the above case, also $(\mathbf{A}^{\langle \mathcal{S}\times\mathcal{T}\rangle})$ is independent, so step (7) and thus the entire algorithm passes as well.

Conversely, if for each $i \in [k]$ and $j \in [\ell]$ for which step (6) was performed $|(\mathbf{A}^{\downarrow S_i \times T_j}) - U_{S_i\times T_j}| \leq \epsilon$ (this step will not pass otherwise by Lemma 4.7), and $|(\mathbf{A}^{\langle \mathcal{S}\times\mathcal{T}\rangle}) - D| \leq \frac{1}{2}\epsilon$ where D over $[k] \times [\ell]$ is an independent distribution, then we show that $\mathbf{A}$ is 3ϵ-independent. First note that $\mathbf{A}(T_0) \leq (1-\epsilon)/\log n$. Now, we define a new random variable $\mathbf{B}$ over $S \times T$ which is defined by first generating an

$(i,j) \in [k] \times [\ell]$ according to D, and then generating $(i',j') \in S_i \times T_j$ according to $U_{S_i \times T_j}$. It is easy to see that $\mathbf{B}$ is independent. Finally, by Lemma 2.12, $|\mathbf{A} - \mathbf{B}| \leq (3/2)\epsilon + \epsilon + (1-\epsilon)/\log n \leq 3\epsilon$, where the second term comes for possibly ignoring pairs i,j for which $\mathbf{A}(i,j) < \epsilon/(k\ell)$ and the third term comes from ignoring $\mathbf{A}(T_0)$.

The sample complexity of this algorithm is dominated by the complexity for each pair of buckets going through the test of Lemma 4.7. It brings us to a total sample complexity of $\tilde{O}((|S|+|T|)\text{poly}(\epsilon^{-1}))$ samples.

□

4.3.2 The light prefixes

We show that using the test for L_1 distance between distributions, the light prefixes can be tested for independence using roughly $\tilde{O}((n^{2-2\alpha}m + n^{2/3})\text{poly}(\epsilon^{-1}))$ samples. Formally, we prove:

Theorem 4.8 *There is an algorithm that given a black-box distribution* $\mathbf{A}$ *over* $S \times T$ *with* $\|\pi_1 \mathbf{A}\|_\infty \leq 2\epsilon^{-1}|S|^\alpha$ *such that: (1) if* $\mathbf{A}$ *is independent, it outputs PASS with high probability and (2) if* $\mathbf{A}$ *is not* 3ϵ*-independent, it outputs FAIL with high probability. It uses* $\tilde{O}((|S|^{2-2\alpha}|T| + |S|^{2/3})\text{poly}(\epsilon^{-1}))$ *many samples.*

PROOF: The following is the outline of the algorithm.

Algorithm *TestLightIndependence*$(\mathbf{A}, \epsilon)$

(1) Obtain an approximation $\tilde{\mathbf{A}}_2$ of $\pi_2\mathbf{A}$ within an $\epsilon/75$ factor, on a $\tilde{T}$ which includes all $j \in [m]$ which have probability at least $(m \log m)^{-1}$.

(2) $\mathcal{T} \stackrel{\text{def}}{=} \{T_0, T_1, \ldots, T_\ell\} =$*Bucket* $(\tilde{\mathbf{A}}_2, \tilde{T}, \epsilon)$; add $T \backslash \tilde{T}$ to T_0.

(3) For $j = 1, \ldots, \ell$ do

(4) If $\mathbf{A}(S \times T_j)$ is not small, then

(5) If $|(\mathbf{A}^{\downarrow S \times T_j}) - (\pi_1(\mathbf{A}^{\downarrow S \times T_j})) \times (\pi_2(\mathbf{A}^{\downarrow S \times T_j}))| \geq \epsilon$,

then FAIL.

(6) Let j' be such that $\mathbf{A}(S \times T_{j'}) > \epsilon/(4\ell)$.

(7) For $j = 1, \ldots, \ell$ do

(8) If $\mathbf{A}(S \times T_j)$ is not small, then

(9) If $|(\mathbf{A}^{\downarrow S \times T_{j'}}) - (\mathbf{A}^{\downarrow S \times T_j})| \geq \epsilon$, then FAIL.

(10) PASS.

The decisions in step (4) and step (8) are done in a similar manner to what was done in Theorem 4.6. We distinguish between $\mathbf{A}(S \times T_j) \geq \epsilon/(2\ell)$ and $\mathbf{A}(S \times T_j) \leq \epsilon/(4\ell)$ by taking $\tilde{O}(\ell/\epsilon)$ samples of $\mathbf{A}$. This guarantees that we need to take $O(\text{poly}(\log(nm))\ell/\epsilon)$ samples of $\mathbf{A}$ for every sample of $(\mathbf{A}^{\downarrow S \times T_j})$ required in step (5) and step (9), by re-sampling $\mathbf{A}$ until we obtain a member of the required set (similarly step (6) guarantees this for sampling $(\mathbf{A}^{\downarrow S \times T_{j'}})$).

The projections appearing in step (5) are sampled by sampling the respective distribution and ignoring a coordinate. Obtaining the j' in step (6) can be done for example using a brute-force approximation of $(\mathbf{A}^{\langle \{S\} \times \mathcal{T} \rangle})$.

The test for the distribution difference in step (5) is done by using Theorem 3.10 with parameter ϵ and the distributions $(\mathbf{A}^{\downarrow S \times T_j})$ and $(\pi_1(\mathbf{A}^{\downarrow S \times T_j})) \times (\pi_2(\mathbf{A}^{\downarrow S \times T_j}))$; the bound on the L_∞ norm of the distributions involved will be given below. The test for the difference in step (9) is done similarly, but this time using Theorem 3.1 with parameter ϵ.

Notice that $\|(\mathbf{A}^{\downarrow S \times T_j})\|_\infty \leq 2|S|^{-\alpha}/\epsilon$ for every T_j (because of the bound on $\|\pi_1 \mathbf{A}\|_\infty$), and that $\|\pi_2(\mathbf{A}^{\downarrow S \times T_j})\|_\infty \leq (1 + 3\epsilon)|T_j|^{-1}$.

The total sample complexity for steps (3)–(5) is given by $\log |T|$ times the sample complexity for iteration j. The sample complexity of the latter is given by Theorem 3.10, which is $\tilde{O}((1+3\epsilon)\cdot(|S||T_j|)^2\cdot|S|^{-\alpha}\cdot|S|^{-\alpha}|T_j|^{-1}\cdot\epsilon^{-5})$, times the $\tilde{O}(\ell/\epsilon)$ for sampling from the restrictions to the buckets. This clearly dominates the sample complexity for step (6), and the sample complexity for steps (7)–(9), which is $\tilde{O}(|S|^{2/3}\epsilon^{-5})$ by multiplying the estimate of Theorem 3.1, the sample complexity of the restricted distributions, and the number of iterations. This completes the estimate given in the statement of the theorem.

As for correctness, if $\mathbf{A}$ is independent then it readily follows that the algorithm accepts, while on the other hand it is not hard to see that if the distribution pairs compared in step (5) and step (9) are indeed all ϵ-close, then $\mathbf{A}$ is 3ϵ-independent. □

4.3.3 Putting them together

We now give the algorithm for the general case.

Theorem 4.9 *There is an algorithm that given a distribution* $\mathbf{A}$ *over* $[n]\times[m]$ *and an* $\epsilon > 0$*: (1) if* $\mathbf{A}$ *is independent, it outputs PASS with high probability and (2) if* $\mathbf{A}$ *is not* 7ϵ*-independent, it outputs FAIL with high probability. The algorithm uses* $\tilde{O}(n^{2/3}m^{1/3}\mathrm{poly}(\epsilon^{-1}))$ *samples.*

PROOF: The following is the outline of the algorithm.

Algorithm *TestIndependence*$(\mathbf{A}, n, m, \epsilon)$

(1) Let β be such that $m = n^\beta$, and set $\alpha = (2+\beta)/3$.

(2) Obtain an approximation $\tilde{\mathbf{A}}_1$ of $\pi_1\mathbf{A}$ within an $\epsilon/75$ factor,

on an $\tilde{S}$ which includes all $i \in [n]$ which have probability at least $n^{-\alpha}$

and no $i \in [n]$ which has probability at most $n^{-\alpha}/2$.

(3) If $(\pi_1\mathbf{A})(\tilde{S})$ is not small then

(4) If *TestHeavyIndependence*$((\mathbf{A}^{\downarrow \tilde{S}\times[m]}), (\tilde{\mathbf{A}}_{\mathbf{1}}^{\downarrow \tilde{S}\times[m]}), \epsilon)$ fails then FAIL.

(5) If $(\pi_1\mathbf{A})([n]\backslash\tilde{S})$ is not small then

(6) If *TestLightIndependence*$((\mathbf{A}^{\downarrow([n]\backslash\tilde{S})\times[m]}), \epsilon)$ fails then FAIL.

(7) If both $(\pi_1\mathbf{A})(\tilde{S})$ and $(\pi_1\mathbf{A})([n]\backslash\tilde{S})$ are not small then

(8) If $\pi_2(\mathbf{A}^{\downarrow \tilde{S}\times[m]})$ and $\pi_2(\mathbf{A}^{\downarrow([n]\backslash\tilde{S})\times[m]})$ are not ϵ-close, then FAIL.

(9) PASS.

In the above algorithm, steps (3), (5) and (7) use sampling to distinguish between the cases where the respective quantities are at least ϵ and the cases where they are at most $\epsilon/2$. Step (4) (if required) is done by using Theorem 4.6, and step (6) is done by using Theorem 4.8; by the choice of α in step (1), the number of queries in both is $\tilde{O}(n^{2/3}m^{1/3}\text{poly}(\epsilon^{-1}))$ times the $O(\epsilon^{-1}\log(nm))$ queries required for filtering the restricted distributions (a factor which does not change the above estimate).

For performing step (8) the two distributions are fed into Theorem 3.1, parameterized to guarantee failure if these distributions are more than ϵ-apart; this uses a number of queries that is dominated by the terms in the rest of the algorithm.

It is clear that if $\mathbf{A}$ was independent, then the test will accept with high probability. We now prove that if the test accepts, then $\mathbf{A}$ is 7ϵ-independent.

If steps (4), (6) and (8) were performed and none of the above tests failed, then by a final application of Lemma 2.12, where $\mathcal{R} = \{\tilde{S} \times [m], ([n]\backslash\tilde{S}) \times [m]\}$, we get that our distribution is at least 7ϵ-independent (because step (8) guarantees that the coarsening is not more than ϵ-far from being independent). If steps (4) and (8) were not performed, then $\mathbf{A}(\tilde{S}\times[m]) < \epsilon$, so it contributes no more than ϵ to the farness of $\mathbf{A}$ from being independent, and step (6) is sufficient to guarantee 4ϵ-independence.

Similarly 4ϵ-independence holds if steps (6) and (8) were not performed because of $\mathbf{A}(([n]\backslash\tilde{S}) \times [m])$ has small value. This covers all possible cases and concludes the proof. □

4.4 Lower bound on sample complexity of testing independence

Theorem 4.10 *For any algorithm $\mathcal{A}$ using $o(n^{2/3}m^{1/3})$ samples whenever $n \geq m$, there exist two joint distributions over $[n]\times[m]$ for any sufficiently large $n \geq m$, with one of them being independent and the other not being $(1/6)$-independent, such that $\mathcal{A}$ cannot distinguish between these two joint distributions with probability greater than 2/3.*

PROOF: Fix an algorithm $\mathcal{A}$ using $o(n^{2/3}m^{1/3})$ samples. We first define two joint distributions $\mathbf{A}_0$ and $\mathbf{B}_0$ over $[n] \times [m]$. Let $\beta = \log_n m$ and $\alpha = (2+\beta)/3$.

$$\Pr[\mathbf{A}_0 = (i,j)] = \begin{cases} \frac{1}{2n^\alpha m} & \text{if } 1 \leq i \leq n^\alpha \\ \frac{1}{mn} & n/2 < i \leq n \\ 0 & \text{otherwise} \end{cases}$$

$$\Pr[\mathbf{B}_0 = (i,j)] = \begin{cases} \frac{1}{2n^\alpha m} & \text{if } 1 \leq i \leq n^\alpha \\ \frac{2}{mn} & \text{if } n/2 < i \leq n \quad \text{and} \\ & \quad j \in [1,\ldots,m/2] \\ 0 & \text{otherwise} \end{cases}$$

We now define two joint distributions $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{A}$, $\mathbf{B}$ modify $\mathbf{A}_0$ and $\mathbf{B}_0$ by randomly relabeling each element in $[n]$ and $[m]$. First choose random permutations σ_0 of $[n]$ and $\sigma_1,\ldots,\sigma_n$ of $[m]$. Define $\mathbf{A}$ to be the distribution such

that

$$\Pr\left[\mathbf{A} = (\sigma_0(i), \sigma_i(j))\right] = \Pr\left[\mathbf{A}_0 = (i, j)\right].$$

Likewise define $\mathbf{B}$ to be the distribution such that

$$\Pr\left[\mathbf{B} = (\sigma_0(i), \sigma_i(j))\right] = \Pr\left[\mathbf{B}_0 = (i, j)\right].$$

Note that $\mathbf{A}$ and $\mathbf{B}$ are actually families of distributions (indexed by the permutations). Throughout the rest of the proof, we will refer to $\mathbf{A}$ and $\mathbf{B}$, with an abuse of notation, as individual distributions in these families. Since we fixed the algorithm $\mathcal{A}$, we could choose the permutations $\sigma_0, \ldots, \sigma_n$ to obtain the members of these families that maximizes the error probability of the algorithm $\mathcal{A}$.

The distribution $\mathbf{A}$ is independent whereas the distribution $\mathbf{B}$ is $\frac{1}{6}$-far from independent. This follows from $\mathbf{B}$ being $\frac{1}{2}$-far from $\pi_1\mathbf{B} \times \pi_2\mathbf{B}$ and Proposition 4.1. The distributions $\pi_1\mathbf{A}$ and $\pi_1\mathbf{B}$ are identical, and they give half the weight to a small number, namely n^α, of the elements, and distribute the remaining weight to half of the elements. The distribution $\pi_2\mathbf{A}$ is uniform over its domain independent of the value of $\pi_1\mathbf{A}$. The distribution $\pi_2\mathbf{B}$, however, is uniform over its domain only when $\pi_1\mathbf{B}$ outputs an element with the higher weight, otherwise, conditioned on the event that $\pi_1\mathbf{B}$ takes on a value with the lower probability, $\pi_2\mathbf{B}$ is uniform only on a subset of its domain that is half the size. The choice of σ_i's makes the distribution $\pi_2\mathbf{B}$ uniform on its domain.

Definition 4.11 *For a pair $(i, j) \in [n] \times [m]$, i is the **prefix**. An element $(i, j) \in [n] \times [m]$ such that $\Pr[\mathbf{A}$ (or $\mathbf{B}$) takes on value $(i, j)] = \frac{1}{2n^\alpha m}$ is called a **heavy element**. The prefix i of a heavy element (i, j) is called a **heavy prefix**. Elements and prefixes with non-zero probabilities that are not heavy are called **light**.*

When restricted to the heavy prefixes, both joint distributions are identical. The only difference between **A** and **B** comes from the light prefixes, and the crux of the proof will be to show that this difference will not change the relevant statistics in a statistically significant way. We do this by showing that the only really relevant statistic is the number of prefixes that occur exactly twice and each time with different suffix. We then show that this statistic has a very similar distribution when generated by **A** and **B** because the expected number of such prefixes that are light is much less than the standard deviation of the number of such prefixes that are heavy.

Next, we describe an aggregate representation of the samples that $\mathcal{A}$ takes. We then prove that we can assume without loss of generality that $\mathcal{A}$ is given this representation of the samples as input instead of the samples themselves. Then, we conclude the proof by showing that distributions on the fingerprint when the samples are taken from **A** or **B** are indistinguishable.

Definition 4.12 *Fix a set of samples $S = \{(x_1, y_1), \ldots, (x_s, y_s)\}$ from distribution* **A** *over $[n] \times [m]$. Say the* pattern *of prefix x_i is $\vec{c}$ where c_j is the number of y's such that (x_i, y) appears exactly j times in S. Define the function $d_S(\vec{c})$ to be the number of prefixes x for which the pattern of x is $\vec{c}$. We refer to d_S as the* ***fingerprint*** *of S. We will just use $d(\vec{c})$ when S is clear from context.*

Note that the definition of the fingerprint in this section is different from the one given in Section 3.1.3 due to the difference in the two settings. The next claim shows that the fingerprint of the sample is just as useful as the samples themselves to distinguish between **A** and **B**.

Claim 4.13 *Given algorithm $\mathcal{A}$ which for joint distributions chosen from the family* **A** *or* **B***, correctly distinguishes whether the distribution is independent or ϵ-far from*

independent, there exists algorithm $\mathcal{A}'$ which gets as input only the fingerprint of the generated sample and has the same correctness probability as $\mathcal{A}$.

PROOF: Note that one can view a sample of size s chosen from the distribution $\mathbf{A}$ (respectively $\mathbf{B}$) as first picking s samples from $\mathbf{A}_0$ (respectively, $\mathbf{B}_0$), then picking a set of random permutations of the element labels and outputting the random relabeling of the samples. Thus the randomness used to generate the sample can be divided into two parts: the first set of coins $\phi = (\phi_1, \ldots, \phi_u)$ are the coins used to generate the sample from $\mathbf{A}_0$ ($\mathbf{B}_0$) and the second set of coins $\psi = (\psi_1, \ldots, \psi_v)$ are the coins used to generate the random permutations of the element labels.

The main idea behind the proof is that given the fingerprint of a sample from $\mathbf{A}_0$ (respectively $\mathbf{B}_0$), the algorithm $\mathcal{A}'$ can generate a labeled sample with the same distribution as $\mathbf{A}$ (respectively, $\mathbf{B}$) without knowing which part of the fingerprint is due to heavy or light elements or whether the sample is from $\mathbf{A}$ or $\mathbf{B}$. In particular, given the fingerprint, assign $d(\vec{b})$ distinct labels from $[n]$ to each pattern $\vec{b}$. Suppose that $x_{\vec{b}}$ is assigned to pattern $\vec{b}$. Then create a sample which includes i copies of $(x_{\vec{b}}, y_j)$ for each nonzero b_i and distinct y_j for $1 \leq j \leq b_i$. Then choose random permutations σ_0, $\sigma_1, \ldots, \sigma_n$ of $[n]$ and $[m]$ and use them to relabel the prefixes and suffixes of the sample accordingly.

Thus, $\mathcal{A}'$ generates a sample from the fingerprint and feeds it to $\mathcal{A}$ as input. For each choice of the sample from $\mathbf{A}_0$ according to random coins ϕ, we have that $\Pr_\psi[\mathcal{A}' \text{ correct}] = \Pr_\psi[\mathcal{A} \text{ correct}]$. Therefore, $\Pr_{\phi,\psi}[\mathcal{A}' \text{ correct}] = \Pr_{\phi,\psi}[\mathcal{A} \text{ correct}]$. □

The following lemma shows that it is only the heavy prefixes, which have identical distributions in both $\mathbf{A}$ and $\mathbf{B}$, that contribute to most of the entries in the fingerprint.

Lemma 4.14 *The expected number of light prefixes that occur at least three times in the sample such that at least two of them are the same element is* $o(1)$ *for both* **A** *and* **B**.

PROOF: For a fixed light prefix, the probability that at least three samples will land in this prefix and two of these samples will collide is $o(n^{-1})$. Since there are $n/2$ light prefixes, by the linearity of expectation, the expected number of such light prefixes in the sample is $o(1)$. □

We would like to have the pattern of each prefix be independent of the patterns of the other prefixes. To achieve this we assume that algorithm $\mathcal{A}$ first chooses an integer s_1 from the Poisson distribution with the parameter $\lambda = s = o(n^{2/3}m^{1/3})$. The Poisson distribution with the positive parameter λ has the probability mass function $p(k) = \exp(-\lambda)\lambda^k/k!$. Then, after taking s_1 samples from the input distribution, $\mathcal{A}$ decides whether to accept or reject the distribution. In the following, we show that $\mathcal{A}$ cannot distinguish **A** from **B** with success probability at least 2/3. Since s_1 will have a value larger than $s/2$ with probability at least $1 - o(1)$ and we will show an upper bound on the statistical distance of the distributions of two random variables (i.e., the distributions on the fingerprints), it will follow that no symmetric algorithm with sample complexity $s/2$ can distinguish **A** from **B**.

Let F_{ij} be the random variable that corresponds to the number of times that the element (i, j) appears in the sample. It is well known that F_{ij} is distributed identically to the Poisson distribution with parameter $\lambda = sr_{ij}$, where r_{ij} is the probability of element (i, j) (cf., Feller [11], p. 216). Furthermore, it can also be shown that all F_{ij}'s are mutually independent. The random variable $F_i \stackrel{\text{def}}{=} \sum_j F_{ij}$ is distributed identically to the Poisson distribution with parameter $\lambda = s\sum_j r_{ij}$.

Let $D_{\mathbf{A}}$ and $D_{\mathbf{B}}$ be the distributions on all possible fingerprints when samples are taken from $\mathbf{A}$ and $\mathbf{B}$, respectively. The rest of the proof proceeds as follows. We first construct two processes $P_{\mathbf{A}}$ and $P_{\mathbf{B}}$ that generate distributions on fingerprints such that $P_{\mathbf{A}}$ is statistically close to $D_{\mathbf{A}}$ and $P_{\mathbf{B}}$ is statistically close to $D_{\mathbf{B}}$. Then, we prove that the distributions $P_{\mathbf{A}}$ and $P_{\mathbf{B}}$ are statistically close. Hence, the theorem follows by the indistinguishability of $D_{\mathbf{A}}$ and $D_{\mathbf{B}}$.

Each process has two phases. The first phase is the same in both processes. They randomly generate the prefixes of a set of samples using the random variables F_i defined above. The processes know which prefixes are heavy and which prefixes are light, although any distinguishing algorithm does not. For each heavy prefix, the distribution on the patterns is identical in $\mathbf{A}$ and $\mathbf{B}$ and is determined by choosing samples according to the uniform distribution on elements with that prefix. The processes $P_{\mathbf{A}}$ and $P_{\mathbf{B}}$ use the same distribution to generate the patterns for each heavy prefix. For each each light prefix i that appears k times for $k \neq 2$, both $P_{\mathbf{A}}$ and $P_{\mathbf{B}}$ will determine the pattern of the prefix to be $(k, \vec{0})$. This concludes the first phase of the processes.

In the second phase, $P_{\mathbf{A}}$ and $P_{\mathbf{B}}$ determine the entries of the patterns for the light prefixes that appear exactly twice. These entries are distributed differently in $P_{\mathbf{A}}$ and $P_{\mathbf{B}}$. There are only two patterns to which these remaining prefixes can contribute: $(2, \vec{0})$ and $(0, 1, \vec{0})$. For each light prefix that appears exactly twice, $P_{\mathbf{A}}$ sets the pattern to be $(2, \vec{0})$ with probability $1 - (1/m)$ and $(0, 1, \vec{0})$ otherwise. For such light prefixes, $P_{\mathbf{B}}$ sets the pattern to be $(2, \vec{0})$ with probability $1 - (2/m)$ and $(0, 1, \vec{0})$ otherwise.

Since the patterns for all prefixes are determined at this point, both process output the fingerprint of the sample they have generated.

Lemma 4.15 *The output of $P_{\mathbf{A}}$, viewed as a distribution, has L_1 distance $o(1)$ to $D_{\mathbf{A}}$. The output of $P_{\mathbf{B}}$, viewed as a distribution, has L_1 distance $o(1)$ to $D_{\mathbf{B}}$.*

PROOF: The distribution that $P_{\mathbf{A}}$ generates is the distribution $D_{\mathbf{A}}$ conditioned on the event that all light prefixes has one of the following patterns: $(k, \vec{0})$ for $k \geq 0$ or $(0, 1, \vec{0})$. Since this conditioning holds true with probability at least $1 - o(1)$ by Lemma 4.14, $|P_{\mathbf{A}} - D_{\mathbf{A}}| \leq o(1)$. The same argument applies to $P_{\mathbf{B}}$ and $D_{\mathbf{B}}$. □

Lemma 4.16 *Distributions $P_{\mathbf{A}}$ and $P_{\mathbf{B}}$ have L_1 distance at most $1/6$.*

PROOF: Given the number of times a prefix appears in the sample, the pattern of that prefix is independent of the patterns of all the other prefixes. By the generation process, the L_1 distance between $P_{\mathbf{A}}$ and $P_{\mathbf{B}}$ can only arise from the second phase. We show that the second phases of the processes do not generate an L_1 distance larger than $1/6$.

Let G (respectively, H) be the random variable that corresponds to the values $d(2, \vec{0})$ when the input distribution is $\mathbf{A}$ (respectively, $\mathbf{B}$). Let d' be the part of the fingerprint excluding entries $d(2, \vec{0})$ and $d(0, 1, \vec{0})$. We will use the fact that for any d', $\Pr[P_{\mathbf{A}} \text{ gen. } d'] = \Pr[P_{\mathbf{B}} \text{ gen. } d']$ in the following calculation.

$$
\begin{aligned}
|P_{\mathbf{A}} - P_{\mathbf{B}}| &= \sum_{d} |\Pr[P_{\mathbf{A}} \text{ gen. } d] - \Pr[P_{\mathbf{B}} \text{ gen. } d]| \\
&= \sum_{d'} \Pr[P_{\mathbf{A}} \text{ gen. } d'] \sum_{k \geq 0} \\
&\quad |\Pr\left[P_{\mathbf{A}} \text{ gen. } d(2,\vec{0}) = k | d'\right] - \\
&\quad \Pr\left[P_{\mathbf{B}} \text{ gen. } d(2,\vec{0}) = k | d'\right]| \\
&= \sum_{C \geq 0} \Pr[P_{\mathbf{A}} \text{ gen. } C \text{ prefixes twice}] \sum_{0 \leq k \leq C} \\
&\quad |\Pr\left[P_{\mathbf{A}} \text{ gen. } d(2,\vec{0}) = k | C\right] - \\
&\quad \Pr\left[P_{\mathbf{B}} \text{ gen. } d(2,\vec{0}) = k | C\right]| \\
&= |G - H|
\end{aligned}
$$

Consider the composition of G and H in terms of heavy and light prefixes. In the case of $\mathbf{A}$, let G_h be the number of heavy prefixes that contribute to $d(2,\vec{0})$ and G_l be the number of such light prefixes. Hence, $G = G_h + G_l$. Define H_h, H_l analogously. Then, G_h and H_h are distributed identically. In the rest of the proof, we show that the fluctuations in G_h dominate the magnitude of G_l.

Let ξ_i be the indicator random variable that takes value 1 when prefix i has the pattern $(2,\vec{0})$. Then, $G_h = \sum_{\text{heavy } i} \xi_i$. By the assumption about the way samples are generated, the ξ_i's are independent. Therefore, G_h is distributed identically to the binomial distribution on the sum of n^α Bernoulli trials with success probability $\Pr[\xi_i = 1] = \exp(-s/2n^\alpha)(s^2/8n^{2\alpha})(1 - (1/m))$. An analogous argument shows that G_l is distributed identically to the binomial distribution with parameters $n/2$ and $\exp(-s/n)(s^2/2n^2)(1 - (1/m))$. Similarly, H_l is distributed identically to the binomial distribution with parameters $n/2$ and $\exp(-s/n)(s^2/2n^2)(1 - (2/m))$.

As n and m grow large enough, both G_h and G_l can be approximated well by normal distributions. Therefore, by the independence of G_h and G_l, G is also approximated well by a normal distribution. Similarly, H is approximated well by a normal distribution. That is,

$$\Pr[G = t] \to \frac{1}{\sqrt{2\pi}\sigma_G} \exp(-(t - \mathrm{E}[G])^2/2\mathrm{Var}[G])$$

as $n \to \infty$.

Thus, $\Pr[G = t] = \Omega(1/\sigma_G)$ over an interval I_1 of length $\Omega(\sigma_G)$ centered at $\mathrm{E}[G]$. Similarly, $\Pr[H = t] = \Omega(1/\sigma_H)$ over an interval I_2 of length $\Omega(\sigma_H)$ centered at $\mathrm{E}[H]$. Since $\mathrm{E}[G] - \mathrm{E}[H] = \mathrm{E}[G_l] - \mathrm{E}[H_l] = \exp(-s/n)(s^2/4n)(1/m) = o(\sigma_G)$, $I_1 \cap I_2$ is an interval of length $\Omega(\sigma_{G_h})$. Therefore,

$$\sum_{t \in I_1 \cap I_2} |\Pr[G = t] - \Pr[H = t]| \leq o(1)$$

because for $t \in I_1 \cap I_2$, $|\Pr[G = t] - \Pr[H = t]| = o(1/\sigma_G)$. We can conclude that $\sum_t |\Pr[G = t] - \Pr[H = t]|$ is less than $1/6$ after accounting for the probability mass of G and H outside $I_1 \cap I_2$. □

The theorem follows by Lemma 4.15 and Lemma 4.16.

□

Chapter 5

Approximating the entropy

The (Shannon) entropy is a measure of randomness of a distribution. Formally, for a distribution $\mathbf{p}$ over $[n]$, the entropy of $\mathbf{p}$ is defined as

$$H(\mathbf{p}) \stackrel{\text{def}}{=} -\sum_{i=1}^{n} p_i \log p_i$$

(all the logarithms are base 2). The notion of entropy plays a central role in statistics, physics, information theory, and data compression. For example, knowing the entropy of a random source can shed light on the compressibility of data produced by such a source.

The question of measuring the entropy of a discrete black-box distribution has been considered in both the statistics and physics communities (cf., [17, 27, 21, 26]). None of these works provides a rigorous analysis of the computational efficiency and sample complexity. Furthermore, to the best of our knowledge, the only algorithms which do not require superlinear (in the domain size) sample complexity are those presented in [21, 26]. The algorithms in [21, 26] use estimates of the collision probability to give a reasonable lower bound estimate of the entropy.

A straight-forward algorithm for approximating the entropy of a black-box dis-

tribution takes $\tilde{O}(n)$ samples from the distribution. These samples allow one to estimate the probability of any element that has $\Omega(n^{-1})$ probability mass. The remaining elements have total weight of $o(1)$; thus, they have negligible entropy. It can be shown that the individual probability estimates give a good approximation to the total entropy when plugged into the entropy definition in place of the real values. Note that the sample complexity of this algorithm is superlinear regardless of the approximation ratio desired.

In this chapter, we show that the entropy can be approximated well in sublinear time. In particular, we show that a γ-multiplicative approximation to the entropy can be obtained in time $\tilde{O}(n^{(1+\zeta)/\gamma^2})$, where n is the size of the domain of the distribution and ζ is an arbitrarily small positive constant, provided that the distribution has sufficiently high entropy. We show that one cannot get a multiplicative approximation to the entropy in general. Even for the class of distributions to which our upper bound applies, we show a lower bound of $\Omega(n^{1/(2\gamma^2)})$ samples.

For a set $S \subseteq [n]$, we define $w_{\mathbf{p}}(S) \stackrel{\text{def}}{=} \sum_{i \in S} p_i$. We define the entropy restricted to the set S as

$$H_S(\mathbf{p}) \stackrel{\text{def}}{=} -\sum_{i \in S} p_i \log p_i.$$

Notice that $H_S(\mathbf{p}) + H_{[n]\setminus S}(\mathbf{p}) = H(\mathbf{p})$.

Similar to the previous chapters, we will classify the domain elements based on their probability values according to the input distribution $\mathbf{p}$. Although we still use the heavy–light terminology for this classification, the threshold probability value that determines the boundary of heavy and light elements is determined by the approximation ratio γ. For a distribution $\mathbf{p}$, we define a set of indices that have high probabilities. Formally, we let

$$B_\alpha(\mathbf{p}) \stackrel{\text{def}}{=} \{i \in [n] \mid p_i \geq n^{-\alpha}\}.$$

Given $\gamma > 1$, we say that $\mathcal{A}$ is a γ-**approximation algorithm** for the entropy, if for every input distribution $\mathbf{p}$, $\mathcal{A}$ outputs $\mathcal{A}(\mathbf{p})$ such that $H(\mathbf{p})/\gamma \leq \mathcal{A}(\mathbf{p}) \leq \gamma H(\mathbf{p})$ with probability at least $2/3$.

We present an γ-approximation algorithm for the entropy that takes sublinear number of samples. Moreover, the sample complexity of our algorithm depends on the approximation ratio desired, therefore, the sample complexity reduces as one aims for a weaker approximation ratio. In particular, we prove the following theorem:

Theorem 5.1 *For any $\gamma > 1, 0 < \epsilon_o < 1/2$, there exists an algorithm that can approximate the entropy of a distribution within a multiplicative factor of $(1+2\epsilon_o)\gamma$ with probability at least $2/3$ in $\tilde{O}(n^{\frac{1}{\gamma^2}}/\epsilon_o^2)$ time where n is the size of the domain of the distribution, provided that the entropy of the distribution is at least $\frac{3\gamma}{2\epsilon_o(1-2\epsilon_o)}$.*

Given $\zeta > 0$ and $\gamma' > 1$, one can choose ϵ_o small enough and set $\gamma = \gamma'/(1+2\epsilon_o)$ in Theorem 5.1 to yield a γ'-approximation algorithm which runs in $\tilde{O}(n^{\frac{1+\zeta}{(\gamma')^2}})$ time. Note that choosing ζ small affects both the running time and the set of distributions to which the algorithm can be applied.

To obtain a multiplicative approximation to the entropy of the black-box distribution $\mathbf{p}$, we classify elements in $[n]$ as heavy or light based on whether they belong to $B_\alpha(\mathbf{p})$ for a carefully chosen α. We then approximate the contribution of the entropy of the heavy and light elements separately. We use the approach of the straight-forward algorithm described above for the heavy elements. Then, we estimate of the total probability mass of the light elements. We prove upper and lower bounds on the entropy of the light elements in terms of their total weight and obtain an approximation for their entropy. Finally, we add up the approximations obtained for the entropy of the heavy and the light elements to get an approximation

to the entropy of the distribution.

Section 5.1 shows how to approximate the entropy of the heavy elements, Section 5.2 shows how to approximate the entropy of the light elements, and Section 5.3 combines these approximations to yield Theorem 5.1. In Section 5.4, we present results regarding lower bounds on sample complexity of approximating entropy.

5.1 Approximating the entropy of the heavy elements

The essential idea of estimating the entropy of the heavy elements is to estimate the probability of each of the heavy elements. We show that if we take enough samples, we can estimate the probability of each heavy element from the frequency of samples of that element. Then, we prove that the entropy of a set elements such that we know an estimate of the probability of each member can be approximated using the estimated values in the definition of the entropy.

Lemma 5.2 *For every $0 < \alpha, \epsilon_o \leq 1$ and sufficiently large n, there is an algorithm that uses $O((n^\alpha/\epsilon_o^2) \cdot \log n)$ samples from $\mathbf{p}$ and outputs $\mathbf{q}$ such that with probability at least $1 - n^{-1}$, the following hold for all i:*

1. *if $i \in B_\alpha(\mathbf{p})$, then $|p_i - q_i| \leq \epsilon_o p_i$,*
2. *if $p_i \leq \frac{1-\epsilon_o}{1+\epsilon_o} n^{-\alpha}$, then $q_i \leq (1-\epsilon_o)n^{-\alpha}$.*

PROOF: Let $m = O((n^\alpha/\epsilon_o^2) \cdot \log n)$. Fix element i and let X_j be the indicator variable that indicates j^{th} sample is i. Let $q_i = \sum X_j/m$. By Chernoff bounds, if $p_i \geq n^{-\alpha}$, then

$$\Pr\left[q_i > (1+\epsilon_o)p_i\right] \leq \exp(-\epsilon_o^2 p_i m/3) \leq \exp(-\epsilon_o^2 n^{-\alpha} m/3) \leq n^{-2}.$$

Using a similar argument for the other direction, we can bound the probability that any element i such that $p_i \geq n^{-\alpha}$ is not estimated within $1+\epsilon_o$. Using Chernoff bounds again, we can show that for i such that $p_i < \frac{1-\epsilon_o}{1+\epsilon_o} n^{-\alpha}$,

$$\Pr\left[q_i > (1-\epsilon_o)n^{-\alpha}\right] \leq n^{-2}.$$

Hence, if $i \in B_\alpha(\mathbf{p})$ then $|p_i - q_i| \leq \epsilon_o p_i$. Now, (1) and (2) of the lemma follow from a union bound over all i. □

The following lemma shows that the entropy of elements in $B_\alpha(\mathbf{p})$ can be approximated well using $\mathbf{q}$ (from the statement of Lemma 5.2) instead of $\mathbf{p}$.

Lemma 5.3 *If for each $i \in B$, $|p_i - q_i| \leq \epsilon_o p_i$, then*

$$|H_B(\mathbf{q}) - H_B(\mathbf{p})| \leq \epsilon_o H_B(\mathbf{p}) + 2\epsilon_o w_{\mathbf{p}}(B).$$

PROOF: For $i \in B$, let $q_i = (1+\varepsilon_i)p_i$ such that $|\varepsilon_i| \leq \epsilon_o$. Then,

$$\begin{aligned} H_B(\mathbf{q}) - H_B(\mathbf{p}) &= -\sum (1+\varepsilon_i)p_i \log((1+\varepsilon_i)p_i) \; + \; \sum p_i \log p_i \\ &= -\sum (1+\varepsilon_i)p_i \log p_i - \sum (1+\varepsilon_i)p_i \log(1+\varepsilon_i) + \sum p_i \log p_i \\ &= -\sum \varepsilon_i p_i \log p_i - \sum (1+\varepsilon_i)p_i \log(1+\varepsilon_i). \end{aligned}$$

By the triangle inequality,

$$\begin{aligned} |H_B(\mathbf{q}) - H_B(\mathbf{p})| &\leq |\sum \varepsilon_i p_i \log(1/p_i)| + |\sum (1+\varepsilon_i)p_i \log(1+\varepsilon_i)^{-1}| \\ &\leq \sum |\varepsilon_i| p_i \log p_i + \sum p_i |(1+\varepsilon_i)\log(1+\varepsilon_i)| \\ &\leq \epsilon_o H_B(p) + 2\epsilon_o w_{\mathbf{p}}(B). \end{aligned}$$

The last step above uses the fact that for $|\varepsilon| \leq \epsilon_o \leq 1$, $|(1+\varepsilon)\log(1+\varepsilon)| \leq 2|\varepsilon| \leq 2\epsilon_o$. □

5.2 Approximating the entropy of the light elements

Now, we obtain an approximation of the entropy of the light elements. Suppose set S is such that $S \subseteq [n] \setminus B_\alpha(\mathbf{p})$. Although it may be hard to estimate the probability of any single element in S, the total probability mass of S, $w_{\mathbf{p}}(S)$, can be estimated when it is not too light. Note that if $w_{\mathbf{p}}(S) \leq n^{-\alpha}$, the contribution of entropy from S is below any constant and can be ignored. So, we can assume without loss of generality that $w_{\mathbf{p}}(S) \geq n^{-\alpha}$. In this case, by considering the set S as a single element and using a similar argument to that in the proof of Lemma 5.2, the following holds with probability at least $1 - n^{-2}$: $(1 - \epsilon_o)w_{\mathbf{p}}(S) \leq w_{\mathbf{q}}(S) \leq (1 + \epsilon_o)w_{\mathbf{p}}(S)$.

The following lemma shows upper and lower bounds on the value of the entropy of light elements. A geometric mean of these bounds give an approximation to the entropy of light elements.

Lemma 5.4 $\alpha w_{\mathbf{p}}(S) \log n \leq H_S(\mathbf{p}) \leq w_{\mathbf{p}}(S) \log n + 1/e$.

PROOF: Observe that $H_S(\mathbf{p})$ is a symmetric concave function of $p_1, \ldots, p_n$. To find the maximum value of $H_S(\mathbf{p})$ subject to the constraint that $\sum_{i \in S} p_i = w_{\mathbf{p}}(S)$, we use Lagrange multipliers. Let $u(\mathbf{p}, \lambda) = H_S(\mathbf{p}) + \lambda((\sum_{i \in S} p_i) - w_{\mathbf{p}}(S))$. The maximum is attained when $\partial u / \partial p_i = -\log p_i - (\ln 2)^{-1} + \lambda = 0$ for $i = 1, \ldots, n$ and $\partial u / \partial \lambda = \sum_{i \in S} p_i - w_{\mathbf{p}}(S) = 0$, which yields $p_i = w_{\mathbf{p}}(S)/|S|, \forall i$. This concludes the proof of the upper bound, since $H_S(\mathbf{p}) = w_{\mathbf{p}}(S) \log(|S|/w_{\mathbf{p}}(S)) = w_{\mathbf{p}}(S) \log |S| - w_{\mathbf{p}}(S) \log w_{\mathbf{p}}(S) \leq w_{\mathbf{p}}(S) \log n + 1/e$ for these values of p_i's.

Since $H_S(\mathbf{p})$ is a symmetric concave function it will take its minimum value when the as many as possible of its variables are are at their extreme points, namely,

0 and 1. This follows from the following: for $f(x) \stackrel{\text{def}}{=} -x \log x$, $f(a) + f(b) \leq f(a+\xi) + f(b-\xi)$ when $a < a+\xi < b-\xi < b$ and consequently, the entropy value of light elements could be reduced further when they are not on one of their extreme points. So, $H_S(\mathbf{p})$ will take its minimum value when $n^\alpha w_{\mathbf{p}}(S)$ of p_i's have the value $n^{-\alpha}$, and the rest is 0. In this case, $H_S(\mathbf{p}) = \alpha w_{\mathbf{p}}(S) \log n$. □

5.3 Putting it together

In this section, we describe our approximation algorithm to $H(\mathbf{p})$ and prove Theorem 5.1. Our algorithm uses the results in the previous sections to get approximations to the entropy of the heavy and the light elements. By adding up these approximations, an approximation to the entropy of the distribution is obtained. We refer to the ratio of the number of times an element appears in the sample to the total number of samples as a normalized frequency. Suppose we are seeking a γ-approximation to the entropy.

Algorithm ApproximateEntropy(γ, ϵ_o)

1. Set $\alpha = 1/\gamma^2$.
2. Take $\tilde{O}(n^\alpha/\epsilon_o^2)$ samples from $\mathbf{p}$.
3. Let $\mathbf{q}$ be the normalized frequencies of $[n]$ in the sample
4. Let $B = \{i \mid q_i > (1-\epsilon_o)n^{-\alpha}\}$ ($B_\alpha(\mathbf{p}) \subseteq B$) and $S = [n] \setminus B$.
5. Output $H_B(\mathbf{q}) + \frac{w_{\mathbf{q}}(S) \log n}{\gamma}$.

We now prove Theorem 5.1.

PROOF: (of Theorem 5.1) Using Lemma 5.3 and Lemma 5.4, we have

$$\begin{aligned} H_B(\mathbf{q}) + \frac{w_{\mathbf{q}}(S)\log n}{\gamma} &\leq (1+\epsilon_o)H_B(\mathbf{p}) + 2\epsilon_o + \frac{1+\epsilon_o}{\gamma} w_{\mathbf{p}}(S)\log n \\ &\leq (1+\epsilon_o)(H_B(\mathbf{p}) + \gamma H_S(\mathbf{p})) + 2\epsilon_o \\ &\leq (1+\epsilon_o)\gamma H(\mathbf{p}) + 2\epsilon_o \\ &\leq (1+2\epsilon_o)\gamma H(\mathbf{p}), \end{aligned}$$

if $H(p) \geq 2/\gamma$. Similarly,

$$\begin{aligned} H_B(\mathbf{q}) + \frac{w_{\mathbf{q}}(S)\log n}{\gamma} &\geq (1-\epsilon_o)H_B(\mathbf{p}) - 2\epsilon_o + \frac{1-\epsilon_o}{\gamma} w_{\mathbf{p}}(S)\log n \\ &\geq (1-\epsilon_o)(H_B(\mathbf{p}) + \frac{(H_S(\mathbf{p}) - e^{-1})}{\gamma}) - 2\epsilon_o \\ &= (1-\epsilon_o)(H_B(\mathbf{p}) + H_S(\mathbf{p})/\gamma) - \frac{1-\epsilon_o}{\gamma} e^{-1} - 2\epsilon_o \\ &\geq H(\mathbf{p})/((1+2\epsilon_o)\gamma), \end{aligned}$$

if $H(\mathbf{p}) \geq \frac{3\gamma}{2\epsilon_o(1-2\epsilon_o)} \geq 2/\gamma$. The theorem follows. □

5.4 Lower bounds for approximating the entropy

In this section, we prove two lower bounds on the number of samples needed to approximate the entropy of a distribution within a multiplicative factor of γ. Both of our lower bounds are shown by giving pairs of distributions with entropy ratio γ^2 that are hard to distinguish. The lower bounds follow since an algorithm which approximates the entropy would allow one to distinguish the distributions.

The first thing we show (Lemma 5.5) is that because distributions could have zero entropy, there is no algorithm which can γ-approximate the entropy of *every* distribution. Thus, we restrict our attention to distributions with non-zero entropy;

we show (Theorem 5.6) a lower bound of $\Omega(n^{\frac{1}{2\gamma^2}})$ samples to γ-approximate the entropy for any distribution with entropy at least $(\log n)/\gamma^2$.

Lemma 5.5 *For $\gamma > 1$, there is no algorithm which γ-approximates the entropy of every distribution.*

PROOF: Assume that $\mathcal{A}$ is an algorithm which approximates the entropy of any distribution. For some small constant $0 < c < 1$, let cn^α be an upper bound on the runtime of $\mathcal{A}$ on distributions over $[n]$. Consider the two distributions $\mathbf{p}$ and $\mathbf{q}$ where $\mathbf{p} = (1, 0, \ldots, 0)$ and $\mathbf{q} = (1 - n^{-\alpha}, n^{-\alpha-1}, \ldots, n^{-\alpha-1})$. Any algorithm which uses only cn^α samples is unlikely to distinguish between $\mathbf{p}$ and $\mathbf{q}$. Since the entropy of $\mathbf{p}$ is 0, any algorithm which gives a multiplicative approximation should output 0. On the other hand, any algorithm which approximates the entropy of $\mathbf{q}$ to within a multiplicative factor of γ should output a value which is at least $\frac{1}{\gamma}\alpha n^{-\alpha} \log n > 0$. Thus, any algorithm which γ-approximates the entropy would be able to distinguish between $\mathbf{p}$ and $\mathbf{q}$. □

Theorem 5.6 *Any algorithm which gives a γ-approximation of the entropy for any distribution in $\mathcal{D}_{(\log n)/\gamma^2}$ must use $\Omega(n^{\frac{1}{2\gamma^2}})$ samples.*

PROOF: Consider two distributions $\mathbf{p}$ and $\mathbf{q}$ on n elements where $\mathbf{p}$ is uniform on the set $[n]$ and $\mathbf{q}$ is uniform on a set S which is a randomly chosen subset of $[n]$ of size $n^{\frac{1}{\rho}}$ for some ρ to be determined later. The entropy ratio of $\mathbf{p}, \mathbf{q}$ is ρ. By the analysis of the birthday problem, with probability 1/3, we do not see any repetitions in the sample before we take $\delta n^{\frac{1}{2\rho}}$ samples from either distribution for some constant $\delta < 1$. Hence, $\Omega(n^{\frac{1}{2\rho}})$ samples are needed to distinguish these distributions. The theorem follows by using $\rho = \gamma^2$, which ensures that $H(\mathbf{q})\gamma < H(\mathbf{p})/\gamma$. □

5.5 Some remarks

5.5.1 Entropy estimation via collisions

Several earlier works in statistical physics community [21, 26], suggest the use of the collision probability ($\|\cdot\|_2^2$) to estimate the entropy. In fact, given a bound on the maximum probability, one can show the following lemma (proof omitted), relating the collision probability and the entropy.

Lemma 5.7 *Suppose $\alpha = 1/\gamma$ and $\|p\|_\infty \leq n^{-\alpha}$. Then, the value of $-\log \|\mathbf{p}\|^2$ is a γ-approximation to $H(\mathbf{p})$.*

The following example, however, illustrates the limitations of using collision probability for entropy estimation: Let $\mathbf{p}$ be a distribution such that $p_i = 1/n$ for $i = 1, \ldots n/2$, and $p_i = n^{-\alpha}$ for $i = n/2 + 1, \ldots, n/2 + n^\alpha/2$.

Then $H(\mathbf{p}) = \frac{1+\alpha}{2} \log n$. On the other hand,

$$-\log \|p\|^2 = \log \frac{2n}{n^{1-\alpha} + 1} < 1 + \alpha \log n.$$

Therefore, the ratio

$$\frac{H(\mathbf{p})}{-\log \|p\|^2} \approx \frac{1+\alpha}{2\alpha}.$$

In order for $(1+\alpha)/2\alpha$ to be at most γ, α has to be greater than $1/(2\gamma - 1)$. This implies that in order to use this estimate of the entropy contribution of the light elements along with the estimate of the entropy contribution of the large elements from Section 5.1, we need identify all elements with probability higher than $n^{-\alpha}$. Thus, using this approach we need $\Omega(n^{\frac{1}{2\gamma-1}})$ samples for a γ-approximation. On the other hand, our algorithm yields a better sample complexity of $O(n^{\frac{1+\zeta}{\gamma^2}})$ for arbitrarily small positive ζ.

5.5.2 Uniform distributions over subsets of $[n]$

Consider distributions $\mathcal{E}_k$ which are uniform over some subset $K \subset [n]$ with $|K| = k$. The entropy of this class of distributions is clearly $\log k$. Given a black-box distribution which is promised to be $\mathcal{E}_k$ for some k, the objective is to find k.

Lemma 5.8 *There exists an algorithm that when given black-box access to a distribution $\mathbf{p} \in \mathcal{E}_k$ outputs l such that $k/2 \leq l \leq 2k$ with probability at least $1/5$ using an expected number $O(\sqrt{k})$ of samples.*

PROOF: The algorithm is as follows: Sample until you see some element twice (a collision), say at the t^{th} sample. Output t^2.

In order to prove the lemma, we need to show that both the probability of getting a collision before $\sqrt{k/2}$ samples and the probability not seeing a collision after $\sqrt{2k}$ samples is less than $2/5$.

$$\Pr\left[\text{No collisions after } t \text{ samples}\right] = \prod_{i=1}^{t}(1 - \frac{i-1}{k}).$$

For $t \leq \sqrt{k/2}$,

$$\prod_{i=1}^{t}(1 - \frac{i-1}{k}) \geq 1 - \frac{1}{k}\sum_{i=0}^{t-1} i \geq 1 - \frac{t^2}{k} \geq 1 - \frac{1}{4} = 3/4 > 3/5$$

For $t \geq \sqrt{2k}$,

$$\prod_{i=1}^{t}(1 - \frac{i-1}{k}) \leq (1 - \frac{t}{2k})^t \leq e^{-t^2/2k} \leq e^{-1} < 2/5.$$

□

Chapter 6

Future directions

We study several properties of distributions, and give methods for testing these properties. We focus on the sample complexity of these tasks. In almost all cases, we show lower bounds on the sample complexity that matches the upper bounds we present up to polylogarithmic factors. Thus, the sample complexities of testing closeness, identity, and independence of distributions are characterized tightly. The lower and upper bounds we show for approximating entropy have a polynomially large gap. The sample complexity of approximating entropy remains to be settled.

We use the L_1 norm as a measure of statistical distance. Although this is a common choice, one can imagine other measures being used. For example, the angle between the vectors describing the distributions is one such measure. Our methods do not immediately apply when this measure is used instead.

There are several open problems related to testing independence. One generalization of our setting would be to consider joint distributions on k-tuples. Another generalization to testing independence can be formulated as testing t-wise independence of tuples. Although our independence test can be used to obtain straightforward algorithms of these problems immediately, one might be able to improve

the sample complexity by taking advantage of some carefully chosen trade offs.

Other properties of interest can be formulated. For example, the mutual information is defined on pairs of distributions. It is related to both independence and entropy. Estimating the value of the mutual information of two distributions from samples is a direction that may combine techniques we used in both of these problems.

This line of work can be extended to other properties of distributions. In this work, we make no assumptions about the input distribution. One can imagine that making some assumption about the input distributions could simplify the problem and reduce the sample complexity. Assuming that the input distribution comes from an interesting subclass of all distributions or making assumptions about the method used to generate the distribution would be interesting directions.

We notice that the alteration in the problem setting from testing closeness to testing identity provably reduced the sample complexity. Changing the representation of one of the distributions yielded a simpler problem. One interesting question we can ask is in what other ways the testing algorithm can be provided with some extra help. In a model similar to that of probabilistic proof systems, one can study the testing problems defined on distributions. For example, how much does it help for testing independence if the marginal distributions are given but not trusted? The problem definitely becomes easier, because it reduces to the problem of testing identity.

Bibliography

[1] N. Alon, M. Krivelevich, E. Fischer, and M. Szegedy. Efficient testing of large graphs. In IEEE, editor, *40th Annual Symposium on Foundations of Computer Science: October 17–19, 1999, New York City, New York,*, pages 656–666, 1109 Spring Street, Suite 300, Silver Spring, MD 20910, USA, 1999. IEEE Computer Society Press.

[2] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *JCSS*, 58, 1999.

[3] Ziv Bar-Yossef, Ravi Kumar, and D. Sivakumar. Sampling algorithms: Lower bounds and applications. In *Proceedings of 33th Symposium on Theory of Computing*, Crete, Greece, 6–8 July 2001. ACM.

[4] Tuğkan Batu, Lance Fortnow, Eldar Fischer, Ravi Kumar, Ronitt Rubinfeld, and Patrick White. Testing random variables for independence and identity. In *Proceedings of 42nd FOCS*. IEEE, 2001.

[5] Tuğkan Batu, Lance Fortnow, Ronitt Rubinfeld, Warren D. Smith, and Patrick White. Testing that distributions are close. In *Proceedings of 41th FOCS*. IEEE, 2000.

[6] A. Broder, M. Charikar, A. Frieze, and M. Mitzenmacher. Min-wise independent permutations. *JCSS*, 60, 2000.

[7] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley & Sons, 1991.

[8] N. Cressie and P.B. Morgan. Design considerations for Neyman Pearson and Wald hypothesis testing. *Metrika*, 36(6):317–325, 1989.

[9] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Scientiarum Mathematicarum Hungarica*, 1967.

[10] J. Feigenbaum, S. Kannan, M. Strauss, and M. Viswanathan. An approximate L^1-difference algorithm for massive data streams (extended abstract). In *FOCS 40*, 1999.

[11] William Feller. *An Introduction to Probability Theory and Applications*, volume 1. John Wiley & Sons Publishers, New York, NY, 3rd ed., 1968.

[12] J. Fong and M. Strauss. An approximate L^p-difference algorithm for massive data streams. In *Annual Symposium on Theoretical Aspects of Computer Science*, 2000.

[13] Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *COMBINAT: Combinatorica*, 19, 1999.

[14] Phillip B. Gibbons and Yossi Matias. Synopsis data structures for massive data sets. In *SODA 10*, pages 909–910. ACM-SIAM, 1999.

[15] O. Goldreich and L. Trevisan. Three theorems regarding testing graph properties. Technical Report ECCC-10, Electronic Colloquium on Computational Complexity, January 2001.

[16] Oded Goldreich and Dana Ron. On testing expansion in bounded-degree graphs. Technical Report TR00-020, Electronic Colloquium on Computational Complexity, 2000.

[17] Bernard Harris. The statistical estimation of entropy in the non-parametric case. *Colloquia Mathematica Societatis János Bolyai*, 16:323–355, 1975. Topics in Information Theory.

[18] R. Kannan. Markov chains and polynomial time algorithms. In Shafi Goldwasser, editor, *Proceedings: 35th Annual Symposium on Foundations of Computer Science, November 20–22, 1994, Santa Fe, New Mexico*, pages 656–671, 1109 Spring Street, Suite 300, Silver Spring, MD 20910, USA, 1994. IEEE Computer Society Press.

[19] Sampath Kannan and Andrew Chi-Chih Yao. Program checkers for probability generation. In Javier Leach Albert, Burkhard Monien, and Mario Rodríguez-Artalejo, editors, *ICALP 18*, volume 510 of *Lecture Notes in Computer Science*, pages 163–173, Madrid, Spain, 8–12 July 1991. Springer-Verlag.

[20] E. L. Lehmann. *Testing Statistical Hypotheses*. Wadsworth and Brooks/Cole, Pacific Grove, CA, second edition, 1986. [Formerly New York: Wiley].

[21] S-K. Ma. Calculation of entropy from data of motion. *Journal of Statistical Physics*, 26(2):221–240, 1981.

[22] J. Neyman and E.S. Pearson. On the problem of the most efficient test of statistical hypotheses. *Philos. Trans. Royal Soc. A*, 231:289–337, 1933.

[23] Amit Sahai and Salil Vadhan. A complete promise problem for statistical zero-knowledge. In *Proceedings of the 38th Annual Symposium on the Foundations of Computer Science*, pages 448–457. IEEE, 20–22 October 1997.

[24] Amit Sahai and Salil Vadhan. Manipulating statistical difference. In Panos Pardalos, Sanguthevar Rajasekaran, and José Rolim, editors, *Randomization Methods in Algorithm Design (DIMACS Workshop, December 1997)*, volume 43 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 251–270. American Mathematical Society, 1999.

[25] Alistair Sinclair and Mark Jerrum. Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and Computation*, 82(1):93–133, July 1989.

[26] S.P. Strong, R. Koberle, R. de Ruyter von Steveninck, and W. Bialek. Entropy and information in neural spike trains. *Physical Review Letters*, 80:197–200, 1998.

[27] D. Wolpert and D. R. Wolf. Estimating functions of probability distributions from a finite set of samples. part I: Bayes estimators and the shannon entropy. *Physical Review E*, 52(6):6841–6854, 1995.

[28] Kenji Yamanishi. Probably almost discriminative learning. *Machine Learning*, 18(1):23–50, 1995.