Estimating the Unseen: An n/log(n)-sample Estimator for Entropy and Support Size, Shown Optimal via New CLTs*

Gregory Valiant†
UC Berkeley
Berkeley, California
gregory.valiant@gmail.com

Paul Valiant‡
UC Berkeley
Berkeley, California
pvaliant@gmail.com

ABSTRACT

We introduce a new approach to characterizing the unobserved portion of a distribution, which provides sublinear–sample estimators achieving arbitrarily small additive constant error for a class of properties that includes entropy and distribution support size. Additionally, we show new matching lower bounds. Together, this settles the longstanding question of the sample complexities of these estimation problems, up to constant factors.

Our algorithm estimates these properties up to an arbitrarily small additive constant, using $O(n/\log n)$ samples, where n is a bound on the support size, or in the case of estimating the support size, $1/n$ is a lower bound on the probability of any element of the domain. Previously, no explicit sublinear–sample algorithms for either of these problems were known. Our algorithm is also computationally extremely efficient, running in time linear in the number of samples used.

In the second half of the paper, we provide a matching lower bound of $\Omega(n/\log n)$ samples for estimating entropy or distribution support size to within an additive constant. The previous lower-bounds on these sample complexities were $n/2^{O(\sqrt{\log n})}$.

To show our lower bound, we prove two new and natural multivariate central limit theorems (CLTs); the first uses Stein's method to relate the sum of independent distributions to the multivariate Gaussian of corresponding mean and covariance, under the earthmover distance metric (also known as the Wasserstein metric). We leverage this central limit theorem to prove a stronger but more specific central limit theorem for "generalized multinomial" distributions—a large class of discrete distributions, parameterized by matrices, that represents sums of independent binomial or multinomial distributions, and describes many distributions encountered in computer science. Convergence here is in the strong sense of statistical distance, which immediately implies that any algorithm with input drawn from a generalized multinomial distribution behaves essentially as if the input were drawn from a discretized Gaussian with the same mean and covariance. Such tools in the multivariate setting are rare, and we hope this new tool will be of use to the community.

Categories and Subject Descriptors: F.2 [Analysis of Algorithms and Problem Complexity]: Miscellaneous

General Terms: Algorithms, Theory.

1. INTRODUCTION

Given samples from an unknown discrete distribution, what can we infer about the distribution? The empirical distribution of the samples roughly captures the portion of the distribution which we have observed, but *what can we say about the unobserved portion of the distribution*? Answers to this question are, at least implicitly, central to many estimation problems fundamental to statistics. Despite much research from both the statistics and computer science communities (originating, coincidentally, in independent work of Fisher [20], and Turing [21]—arguably the founding fathers of modern statistics and computer science), this question is still poorly understood. For the two important problems of estimating the support size, and estimating the entropy of a distribution, basic questions, such as the sample complexity of these tasks, have not been resolved. And this is not solely a theoretical question: in contrast to many tasks for which existing algorithms or heuristics perform well in practice (in some cases despite poor worst-case performance), for these two problems, there seems to be no approach that is fully embraced by practitioners [13]. Despite this, much of the recent theoretical work on these problems analyzes properties of existing heuristics. A new, practical algorithm for these tasks may, potentially, have widespread immediate applications in the many fields for which these problems arise, including Database Management, Biology, Ecology, Genetics, Linguistics, Neuroscience, and Physics (see the discussion and extensive bibliographies in [12, 33]).

We introduce a new approach to characterizing the unobserved portion of a distribution, which provides sublinear–sample additive estimators for a class of properties that includes entropy and distribution support size. Together with our new lower bounds, this settles the longstanding open question of the sample complexities of these estimation prob-

*Preliminary full versions, [39] and [40], available at: `http://www.eccc.uni-trier.de/report/2010/179` and `http://www.eccc.uni-trier.de/report/2010/180`

†Supported by an NSF graduate research fellowship.

‡Supported by an NSF postdoctoral research fellowship.

STOC'11, June 6–8, 2011, San Jose, California, USA.

lems (up to constant factors). Our algorithm estimates these properties up to an arbitrarily small additive constant, using $O(n/\log n)$ samples. Our lower bounds show that no algorithm on $o(n/\log n)$ samples can achieve this. Here, n is a bound on the support size, and is a natural parameterization of the difficulty of estimating entropy.[1] Previously, no explicit sublinear-sample algorithms for either of these problems were known.[2] The previous lower–bounds on these sample complexities were $n/2^{\Theta(\sqrt{\log n})}$, given by Valiant in [41], and a prior slightly weaker bound of $n/2^{\Theta(\sqrt{\log n}\cdot\log\log n)}$ for support size given by Raskhodnikova *et al.* [35]. Finally, we note that our algorithm runs in time *linear* in the number of samples used.

The algorithm we exhibit estimates any statistical property which is independent of the labeling of the elements ("symmetric") and sufficiently smooth. Rather than directly trying to estimate a specific property of the distribution, we instead take the *canonical* approach and return to the original question "*what can we infer about the true distribution*" given a sublinear number of samples? Our algorithm returns a distribution that is, with high probability, "close" in some sense to the true distribution. Specifically, we return a distribution D with the property that if we had taken our samples from the hypothetical D instead of from the unknown true distribution, then with high probability the number of support elements occurring once, twice, etc. in this sample will closely match the corresponding parameters of the actual sample. How does one find such a distribution? Via linear programming, the computer scientist's battle-axe—bringing this powerful tool to bear on these problems opens up results that withstood previous approaches to constructing such estimators. Given the distribution D returned by our algorithm, to obtain an estimate for some property, we may simply evaluate the property on D. Unsurprisingly, this yields a very good estimate; surprisingly, one can actually prove this.

1.1 Central Limit Theorems and Lower Bounds

The matching lower bounds hinge upon our two new multivariate central limit theorems; the first relates the sum of independent distributions to the multivariate Gaussian of corresponding mean and covariance, under the earthmover distance matric (also known as the Wasserstein metric). Our proof of this central limit theorem is via Stein's method. We leverage this central limit theorem to prove a stronger but more specific central limit theorem for "generalized multinomial" distributions—a large class of discrete distributions, parameterized by matrices, that generalize binomial and multinomial distributions and describe many distributions encountered in computer science (for example, [18, 19, 37, 41]).

Despite the increasing understanding of the various settings for which central limit theorems apply, most of the attention has been on univariate formulations. And as one might expect, the number of useful formulations of the central limit theorem seems to grow with the dimension; it is, perhaps, not surprising that the particularly natural and useful versions we prove here seem absent from the statistics literature [16].

The connection between our central limit theorem for generalized multinomial distributions, and estimating symmetric properties of distributions, such as entropy and support size, is that generalized multinomial distributions capture the distribution over vectors $(m_1, m_2, \ldots)$, where m_i is the number of domain elements for which we see i representatives in a sample. Our central limit theorem allows us to cleanly reason about the statistical distance between these distributions of summary statistics. Specifically, this will allow us to argue that there are pairs of very different distributions D, D'—different in terms of entropy, or support size, for example—such that there is small statistical distance between the distribution of what we will see given k samples from D and the distribution of what we will see given k samples from D'; thus we can conclude that *no* algorithm can distinguish a set of k samples from D from a set of k samples from D' with high probability, which, in particular, implies that no estimator for entropy, when given k samples from D, can accurately return $H(D)$, rather than $H(D')$.

Finally, we note that the construction and analysis of the distributions D, D' requires considerable effort, and involve two polynomial constructions that may be of independent interest, one involving Laguerre polynomials, one involving Hermite polynomials.

1.2 Historical Background

The problem of estimating an unknown discrete distribution from "too few" samples has a very rich history of study in both statistics and computer science, with early contributions from both R.A Fisher, and Alan Turing. In the early 1940's, R. A. Fisher was approached by a naturalist, Corbet, who had just returned from two years of collecting butterflies in the Malay peninsula. Corbet presented Fisher with data on his butterfly collections—specifically, he indicated the number of species for which he had only seen a single specimen (118 species), the number of species for which he had two specimens (74 species), three specimens (44 species), and so on. Corbet hoped that from this data, the great statistician Fisher would be able to deduce some properties of the true distribution of butterflies in Malay, and in particular, he wanted an estimate of the number of new species he might discover if he were to return to the Malay jungle for another 2 years. Using basic properties of the Poisson distribution, Fisher provided a partial answer to these questions in [20].

At roughly the same time, at the height of WWII, Alan Turing and I.J. Good were working on a similar problem in the rather different context of the pivotal British war–effort to analyze the statistics of the German Enigma Machine ciphers. After the war, the results of their work, the *Good-Turing frequency estimation* scheme were published [21]. In addition to many practical applications of the Good-Turing estimates, there has been considerable recent work from the computer science community analyzing variants of these estimation schemes [29, 31, 32, 42, 43]. While the high–level goals of these estimators are related to our own, the analysis typically fixes a distribution and considers the behavior as the number of samples taken approaches infinity, and thus is somewhat orthogonal to the questions considered here.

The specific problem of estimating the support size of an unknown distribution (also referred to as the problem of es-

[1]For the problem of estimating the distribution support size, it is typically assumed that all elements in the support occur with probability at least $1/n$, since without such a lower bound it is impossible to estimate support size.

[2]See [33] for a nonconstructive proof of the existence of a $o(n)$-sample entropy estimator.

timating the number of species in a population, or the "distinct elements problem") has a very long history of study and arises in many contexts (see [12] for several hundred references). Because arbitrarily many species can lie in an arbitrarily small amount of probability mass, analysis of the sample complexity of this problem is generally parameterized in terms of n, where elements of the distribution are restricted to have probability mass at least $1/n$. Tight multiplicative bounds of $\Omega(n/\alpha^2)$ for approximating the entropy to a multiplicative factor of α are given in [4, 15] though they are somewhat unsatisfying as the worst-case instance is distinguishing a distribution with support size *one* from a distribution of support size α^2. The first strong lower bounds for *additively* approximating the support size were given in [35], showing that for any constant $\delta > 0$, any estimator that obtains additive error at most $(1/2-\delta)n$ with probability at least $2/3$ requires at least $n/2^{\Theta(\sqrt{\log n}\cdot\log\log n)}$ samples. To the best of our knowledge, there were no improvements upon the trivial $\Omega(n)$ upper bound for this problem.

For the problem of entropy estimation, there has been recent work from both the computer science and statistics communities. Batu *et al.* [6, 7, 8], Guha *et al.* [23], and Valiant [41] considered the problem of multiplicatively estimating the entropy; in all these works, the estimation algorithm has the following basic form: given a set of samples, discard the species that occur infrequently and return the entropy of the empirical distribution of the frequently–occurring elements, adjusted by some function of the amount of missing probability mass. In particular, no attempt is made to understand the portion of the true distribution consisting of infrequently occurring elements—the "unseen", or little–seen, portion. In a different direction, Paninski gave a simple though non–constructive proof of the existence of a sublinear sample estimator for additively approximating the entropy to within a constant; the proof is via a direct application of the Stone-Weierstrass theorem to the set of Poisson functions [33, 34]. The best previous lower bounds were $n/2^{\Theta(\sqrt{\log n})}$, given in [41].

Additionally, there has been much work on estimating the support size (and the general problem of estimating frequency moments) and estimating the entropy in the setting of *streaming*, in which one has access to very little memory and can perform only a single pass over the data [2, 3, 10, 14, 25, 26, 27, 44].

Teleologically, perhaps the work most similar to our own is Orlitsky *et al.*'s investigation into what they term "pattern maximum likelihood" [1, 30]. Their work is prompted by the following natural question: given a set of samples, what distribution maximizes the likelihood of seeing the observed species frequencies, that is, the number of species observed once, twice, etc? (What Orlitsky *et al.* term the *pattern* of a sample, we call the *fingerprint*, as in Definition 3.) While it seems unclear how to prove that such a likelihood maximizing distribution would, necessarily, have similar property values to the true distribution, at least intuitively one might hope that this is true. From a computational standpoint, while Orlitsky *et al.* show that such likelihood maximizing distributions can be found in some specific settings, the problem of finding or approximating such distributions in the general setting seems daunting.

1.2.1 Stein's Method

Since Stein's seminal paper [38], presented in 1970, describing an alternative proof approach—what became known as "Stein's method"— for proving Berry-Esseen-style central limit theorems, there has been a blossoming realization of its applicability to different settings. In particular, there have been several successful applications of Stein's method in multivariate settings [17, 22, 36].

To prove our first central limit theorem, we closely follow the treatment for the multivariate limit theorem given by Götze in [22] (see also [11] for an exposition). The distinction between our first central limit theorem (which is in terms of earthmover distance), and that of Götze, lies in the distance metric. Götze's result shows convergence in terms of the discrepancy between the probabilities of any *convex set*. Applying this result, intuitively, seems to require decomposing some high-dimensional set into small convex pieces, which, unfortunately, tends to weaken the result by exponential factors. It is perhaps for this reason that, despite much enthusiasm for Götze's result, there is a surprising absence of applications in the literature, beyond small constant dimension.

2. DEFINITIONS AND EXAMPLES

We state the key definitions that will be used throughout, and provide some illustrative examples.

DEFINITION 1. *A* distribution *on* $[n] = \{1, \ldots, n\}$ *is a function* $p : [n] \to [0, 1]$ *satisfying* $\sum_i p(i) = 1$. *Let* $\mathcal{D}^n$ *denote the set of distributions over domain* $[n]$.

Throughout this paper, we will use n to denote the size of the domain of our distribution, and k to denote the number of samples from it that we have access to.

We now define the notion of a *symmetric property*. Informally, symmetric properties are those that are invariant to renaming the domain elements.

DEFINITION 2. *A* property *of a distribution is a function* $\pi : \mathcal{D}^n \to \mathbb{R}$. *Additionally, a property is* symmetric *if, for all distributions* D, *and all permutations* σ, $\pi(D) = \pi(D \circ \sigma)$.

DEFINITION 3. *Given a sequence of samples* $X = (x_1, \ldots, x_k)$, *the associated* fingerprint, *denoted* $\mathcal{F}_X$, *is the "histogram of the histogram" of the samples. Formally,* $\mathcal{F}_X$ *is the vector whose* i^{th} *component,* $\mathcal{F}_X(i)$ *is the number of elements in the domain that occur exactly* $i \geq 1$ *times in sample* X. *In cases where the sample* X *is unambiguous, we omit the subscript.*

Throughout, we will be dealing exclusively with symmetric properties. For such properties, the fingerprint of a sample contains all the useful information about the sample: for any estimator that uses the actual samples, there is an estimator of equal performance that takes as input only the fingerprint of the samples (see [6, 9], for an easy proof). Note that in some of the literature the fingerprint is alternately termed the *pattern*, *histogram*, or *summary statistics* of the sample.

Throughout, we will be representing sets of samples via their fingerprint, and analogously, we will be representing distributions by their *histogram*.

DEFINITION 4. *The* histogram *of a distribution* p *is a mapping* $h : (0, 1] \to \mathbb{N} \cup \{0\}$, *where* $h(x) = |\{i : p(i) = x\}|$. *Additionally, we allow* generalized histograms, *which do not necessarily take integral values.*

Any symmetric property is clearly a function of the histogram of the distribution. Since $h(x)$ denotes the number of elements that have probability x, it follows that $\sum_{x:h(x)\neq 0} h(x)$ equals the support size of the distribution. The probability mass at probability x is $x \cdot h(x)$, thus $\sum_{x:h(x)\neq 0} x \cdot h(x) = 1$, for any histogram that corresponds to a distribution.

We now define what it means for two distributions to be "close"; because the values of symmetric properties depend only upon the histograms of the distributions, we must be slightly careful in defining this distance metric so as to ensure that it will be well–behaved with respect to the properties we are considering. In particular, "close" distributions will have similar values of entropy and support size.

DEFINITION 5. *For two distributions p_1, p_2 with respective histograms h_1, h_2, we define the* relative earthmover distance *between them, $R(p_1, p_2) := R(h_1, h_2)$, as the minimum over all schemes of moving the probability mass of the first histogram to yield the second histogram, of the cost of moving that mass, where the per-unit cost of moving mass from probability x to y is $|\log(x/y)|$.*

Note that statistical distance is upper bounded by relative earthmover distance.

The structure of the distribution of fingerprints intimately involves the Poisson distribution. Throughout, we use $Poi(\lambda)$ to denote the Poisson distribution with expectation λ, and for a nonnegative integer j, $poi(\lambda, j) := \frac{\lambda^j e^{-\lambda}}{j!}$, denotes the probability that a random variable distributed according to $Poi(\lambda)$ takes value j.

We provide two clarifying examples of the above definitions:

EXAMPLE 1. *Consider a sequence of fish species, found as samples from a certain lake, $X = (trout, salmon, trout, cod, cod, whale, trout, eel, salmon)$. We have $\mathcal{F}_X = (2, 2, 1)$, indicating that two species occurred exactly once (whale and eel), two species occurred exactly twice (salmon and cod), and one species occurred exactly three times (trout).*

Suppose that the true distribution of fish is the following:

$$Pr(trout) = 1/2, \quad Pr(salmon) = 1/4,$$
$$Pr(cod) = Pr(whale) = Pr(eel) = Pr(shark) = 1/16.$$

The associated histogram *of this distribution is $h : \mathbb{R}^+ \to \mathbb{Z}$ defined by $h(1/16) = 4$, $h(1/4) = 1$, $h(1/2) = 1$, and for all $x \notin \{1/16, 1/4, 1/2\}$, $h(x) = 0$. If we now consider a second distribution over $\{a, b, c\}$ defined by the probabilities $Pr(a) = 1/2$, $Pr(b) = 1/4$, $Pr(c) = 1/4$, and let h' be its associated histogram, then the relative earthmover distance $R(h, h') = \frac{1}{4}|\log \frac{1/4}{1/16}|$, since we must take all the mass that lies at probability $1/16$ and move it to probability $1/4$ in order to turn the first distribution into one that yields a histogram identical to h'.*

EXAMPLE 2. *Consider the uniform distribution on $[n]$, which has histogram h such that $h(\frac{1}{n}) = n$, and $h(x) = 0$ for $x \neq \frac{1}{n}$. Let $k \leftarrow Poi(5n)$ be a Poisson-distributed random number, and let X be the result of drawing k independent samples from the distribution. The number of occurrences of each element of $[n]$ will be independent, and distributed according to $Poi(5)$. Note that $\mathcal{F}_X(i)$ and $\mathcal{F}_X(j)$ are* not *independent (since, for example, if $\mathcal{F}_X(i) = n$ then it must be the case that $\mathcal{F}_X(j) = 0$, for $i \neq j$). A fingerprint of a typical trial will look roughly like $\mathcal{F}(i) \approx n \cdot poi(5, i)$, for all $i \geq 1$.*

Throughout, we will restrict our attention to properties that satisfy a weak notion of continuity, defined via the relative earthmover distance.

DEFINITION 6. *A symmetric distribution property π is (ϵ, δ)-*continuous *if for all distributions D_1, D_2 with respective histograms h_1, h_2 satisfying $R(h_1, h_2) \leq \delta$ it follows that $|\pi(D_1) - \pi(D_2)| \leq \epsilon$.*

We note that both entropy and support size are easily seen to be continuous with respect to the relative earthmover distance.

FACT 1. *For a distribution $p \in \mathcal{D}^n$, and $\delta > 0$*

- *The entropy, $H(p) := -\sum_i p(i) \cdot \log p(i)$ is (δ, δ)-continuous, with respect to the relative earthmover distance.*
- *The support size $S(p) := |\{i : p(i) > 0\}|$ is $(n\delta, \delta)$-continuous, with respect to the relative earthmover distance, over the set of distributions which have no probabilities in the interval $(0, \frac{1}{n})$.*

2.1 Property Testers

A property tester takes as input k independent samples from a distribution, and is considered good if it correctly classifies the distribution with probability at least $\frac{2}{3}$.

In this paper, we consider the very related notion of a "Poissonized" tester, which, for a distribution p receives input constructed in the following way:

- Draw $k' \leftarrow Poi(k)$.
- Return k' samples from p.

The reason why Poissonized testers are substantially easier to analyze, is the fundamental fact, illustrated in Example 2, that the numbers of samples drawn from each element of the support of p will be *independent* of each other, and, specifically, distributed as independent (univariate) Poisson processes.

Further, we note that these two notions of testing—"regular" testing, and Poissonized testing—have sample complexities within a constant factor of each other, since one can simulate each with the other, with high probability (via tail bounds). The criteria that testers succeed with probability $\frac{2}{3}$ is arbitrary, and, indeed, may be amplified exponentially by repeating the tester and returning the majority answer.

3. MAIN RESULTS

We introduce a novel approach to creating estimators for symmetric distribution properties. We hope (and believe) that variants of our proposed estimator will prove useful in practice.

Our main technical result is a canonical estimator for relative-earthmover continuous properties. We stress that our estimator is truly canonical in that it is agnostic to the choice of property that one is trying to estimate. In particular, the estimator works by first constructing a distribution completely independently of the property in question, and then simply returning the evaluation of the property on

this distribution. Even if the property in question is computationally intractable to evaluate, the first stage of our estimator still runs in time linear in the number of samples, returning a distribution capturing the value of the property.

THEOREM 1. *For sufficiently large n, and any constant $c > 1$, given $c\frac{n}{\log n}$ independent samples from $D \in \mathcal{D}^n$, with probability at least $1 - o(\frac{1}{poly(n)})$ over the random samples, our algorithm returns a distribution D', representable as an $O(c\frac{n}{\log n})$-length vector, such that the relative-earthmover distance between D and D' satisfies*

$$R(D, D') \leq O\left(\frac{1}{\sqrt{c}}\right).$$

Furthermore, our algorithm runs in time $O(c\frac{n}{\log n})$.

For entropy and support size, Theorem 1 together with Fact 1 yields:

COROLLARY 1. *There exists a constant c such that for any positive $\epsilon < 1$ and sufficiently large n, given $\frac{c}{\epsilon^2}\frac{n}{\log n}$ independent samples from $D \in \mathcal{D}^n$, in time $O(\frac{c}{\epsilon^2}\frac{n}{\log n})$ our estimator will output a pair of real numbers (h, s) such that with probability $1 - o(\frac{1}{poly(n)})$*

- *h is within ϵ of the entropy of D, and*
- *s is within $n\epsilon$ of the support size of D, provided none of the probabilities in D lie in $(0, \frac{1}{n})$.*

Our lower bound is the following:

THEOREM 2. *For any positive constant $\phi < \frac{1}{4}$, there exists a pair of distributions p^+, p^- that are $O(\phi|\log\phi|)$-close in the relative earthmover distance, respectively, to the uniform distributions on n and $\frac{n}{2}$ elements, but which are indistinguishable to $k = \frac{\phi}{32} \cdot \frac{n}{\log n}$-sample testers.*

That is, estimating entropy to any constant error less than $\frac{\log 2}{2}$ requires $\Theta(\frac{n}{\log n})$ samples, as does estimating support size to any constant error less than $\frac{n}{4}$.

Further, by choosing a positive $\epsilon < 1$ and then constructing the distributions $p^+_\epsilon, p^-_\epsilon$ that, with probability ϵ draw samples from p^+, p^- respectively and otherwise return another symbol, $\perp$, we note that the entropy gap between p^+_ϵ and p^-_ϵ is an ϵ fraction of what it was originally, and further that distinguishing them clearly requires a factor $\frac{1}{\epsilon}$ more samples. That is,

COROLLARY 2. *For large enough n and small enough ϵ, the sample complexity of estimating entropy to within ϵ grows as $\Omega(\frac{n}{\epsilon \log n})$.*

We note that while the positive results of Theorem 1 match these lower bounds in their dependence on n, there is a gap in the dependence on the desired accuracy, ϵ, with a $\frac{1}{\epsilon}$ dependence in the lower bounds and a $\frac{1}{\epsilon^2}$ dependence in the upper bound. Phrased differently, for an optimal entropy estimator, as the number of samples increases, *does the error decay linearly, or with the square root of the number of samples?*

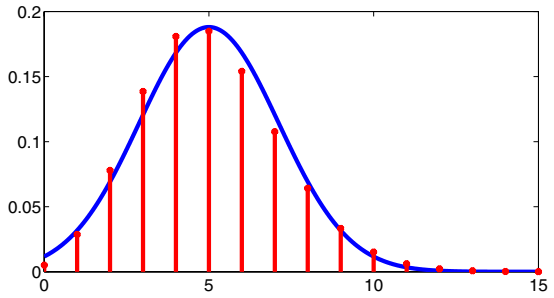

Figure 1: The binomial distribution with $p = 0.1$ and 50 samples (red bars), compared with the Gaussian distribution of matching mean and variance (blue curve). Theorem 3, implies that the earthmover distance between these distributions is at most $0.9(2.7 + 0.83 \log 50)$.

3.1 Two Multivariate Central Limit Theorems

The proof of our lower bound hinges on our multivariate central limit theorems. We suspect that these fundamental limit theorems will have many applications beyond property testing. Our first central limit theorem, proved directly via Stein's method, applies to the very general setting of sums of bounded independent random variables, and is in terms of the *earthmover distance*, also referred to as the *Wasserstein* distance metric.

DEFINITION 7. *Given two distributions A, B in $\mathbb{R}^k$, then, letting* $\text{Lip}(\mathbb{R}^k, 1)$ *denote the set of functions $h : \mathbb{R}^k \to \mathbb{R}$ with Lipschitz constant 1, that is, where for any $x, y \in \mathbb{R}^k$ we have $|h(x) - h(y)| \leq ||x - y||$, then the* earthmover distance *between A and B is defined as*

$$d_W(A, B) = \sup_{h \in \text{Lip}(\mathbb{R}^k, 1)} E[h(A)] - E[h(B)].$$

THEOREM 3. *Given n independent distributions $\{Z_i\}$ of mean 0 in $\mathbb{R}^k$ and a bound β such $||Z_i|| < \beta$ for any i and any sample, then the earthmover distance between $\sum_{i=1}^n Z_i$ and the normal distribution of corresponding mean (0) and covariance is at most $\beta k(2.7 + 0.83 \log n)$.*

Figure 3.1 provides a simple illustration of Theorem 3, in the univariate setting ($k = 1$); of course, in the univariate setting, such central limit theorems are standard (see [5]).

We note the parameters of Theorem 3: as samples from more and more distributions are added in, the performance of the approximation only gets very mildly worse, increasing with the logarithm of the number of samples, n. In fact, we strongly suspect that, in analogy with univariate results, there should be no dependence on n in the theorem. The linear dependence on k, the dimension, is more fundamental; it is not hard to show that this dependence must be of order at least $\sqrt{k}$, so one might conjecture a tight form of the theorem's bound to be $\Theta(\beta\sqrt{k})$.

We note that it is somewhat more standard for central limit theorems of this type to be stated in terms of third moments, instead of a bound β on each random variable. In the full version of this paper we do indeed proceed in this way, and Theorem 3 is derived as a clean special case, sufficient for our applications.

3.1.1 A CLT for Statistical Distance

To provide algorithmic lower-bounds, we must work with a much more stringent distance metric than earthmover distance. In our second central limit theorem, we work with *statistical distance* (sometimes referred to as total variation distance, D_{TV}, or L_1 distance). Fundamentally, if distributions A and B have statistical distance 0.1, then *any* algorithm taking an input drawn from A must behave identically at least 90% of the time to the algorithm run on an input drawn from B.

We first note that the conditions of Theorem 3 are not strong enough to imply any sort of statistical distance bound: the discrete distribution illustrated in Figure 3.1 has (maximal) statistical distance 1 from its Gaussian approximation. However, the intuition for our second central limit theorem is the observation that the statistical distance between the two distributions of Figure 3.1 is in fact very small if we first round the Gaussian distribution to be supported on the lattice points. We now define the class of distributions to which our limit theorem will apply.

DEFINITION 8. *The* generalized multinomial distribution *parameterized by a nonnegative matrix ρ each of whose rows sum to at most 1, is denoted M^ρ, and is defined by the following random process: for each row $\rho(i,\cdot)$ of matrix ρ, interpret it as a probability distribution over the columns of ρ—including, if $\sum_{j=1}^{k}\rho(i,j) < 1$, an "invisible" column 0—and draw a column index from this distribution; return a row vector recording the total number of samples falling into each column (the histogram of the samples).*

The "invisible" column is used for the same reason that the binomial distribution is taken to be a univariate distribution; while one could consider it a bivariate distribution, counting heads and tails separately, it is convenient to consider tails "invisible", as they are implied by the number of heads.

DEFINITION 9. *The k-dimensional* discretized *Gaussian distribution, with mean μ and covariance matrix Σ, denoted $\mathcal{N}^{disc}(\mu,\Sigma)$, is the distribution with support $\mathbb{Z}^k$ obtained by picking a sample according to the Gaussian $\mathcal{N}(\mu,\Sigma)$, then rounding each coordinate to the nearest integer.*

Our second central limit theorem, that we leverage for this paper's lower bounds, is the following:

THEOREM 4. *Given a generalized multinomial distribution M^ρ, with k dimensions and n rows, let μ denote its mean and Σ denote its covariance matrix, then*

$$D_{tv}\left(M^\rho, \mathcal{N}^{disc}(\mu,\Sigma)\right) \leq \frac{k^{4/3}}{\sigma^{1/3}} \cdot 2.2 \cdot (3.1 + 0.83\log n)^{2/3},$$

where σ^2 is the minimum eigenvalue of Σ.

We overview some of the key ideas of the proof. Note that even among distributions over the lattice points, bounds on the earthmoving distance do not necessarily translate into bounds on statistical distance—consider a distribution supported on the even integers, versus one supported only on the odd integers, or some much worse high-dimensional analogue. However, one elementary and completely general way to convert earthmover distance bounds, such as those of Theorem 3, into statistical distance bounds is to convolve the distributions by a smooth distribution that is "wide enough".

Thus the statistical distance between *convolved* versions of these distributions is small. We must, however, "deconvolve" to achieve the desired result. Deconvolution, in general, is very poorly behaved and can blow up badly. The saving grace in our setting is the fact that any multinomial distribution is in fact *unimodal* in each coordinate direction. (Intuitively, at least for the one-dimensional case, unimodality is what prevents one distribution from being supported on, say, only the even integers.) Specifically, we prove a "deconvolution lemma" that has good bounds when the result of deconvolution is unimodal.

While binomial distributions are trivially unimodal, the analysis rapidly gets complicated. The general result for the univariate case is known as *Newton's inequalities*. The multivariate case, which we rely on in our proof of Theorem 4, was proven only recently in a 2008 work of Gurvits—see Fact 1.10:2 of [24].

4. THE ESTIMATOR

In this section we describe the algorithmic portion of this work: how to estimate entropy or support size—or indeed any relative-earthmover continuous property—in $k = O(\frac{n}{\log n})$ samples and time.

Given the fingerprint $\mathcal{F}$ derived from a set of $Poi(k)$ samples, we will construct a linear program that has no objective function, and whose feasible polytope corresponds, roughly, to a set of "plausible" histograms. Each of these plausible histograms, h' has the property that with reasonable probability, the fingerprint derived from taking $Poi(k)$ samples from h' will be quite similar to $\mathcal{F}$. Intuitively, such a histogram is a natural guess for the true histogram. We prove that this intuition is, in fact, correct, in the sense that for well-behaved properties such as entropy and support size, with high probability, h' and the true distribution will have similar property values. See Figure 4 for several examples of fingerprints and their corresponding histograms.

Before proceeding, it will be helpful to gain some intuition for the distribution over fingerprints yielded by taking $Poi(k)$ samples from some histogram h. Since the number of occurrences of different domain elements are independent (because we are taking $k' \leftarrow Poi(k)$ samples, instead of exactly k), the probability that some domain element ℓ occurs exactly i times is simply $poi(kx_\ell, i)$, where x_ℓ is the probability of ℓ. Thus the random variable $\mathcal{F}(i)$ can be expressed as the sum of n independent boolean random variables, and

$$E[\mathcal{F}(i)] = \sum_{x:h(x)\neq 0} h(x)poi(kx,i),$$

and because of independence, $\mathcal{F}(i)$ must be tightly concentrated about this expectation.

At this point we can also see why the desired task of finding a plausible histogram h' can be represented as a linear program: the expected fingerprint entries are *linear* functions of the histogram values $h(x)$, with coefficients $poi(kx,i)$. For the sake of clarity we outline a linear program that illustrates the intuition behind our estimator. Given input $\mathcal{F}$, the linear program will return a histogram h with the property that the expected fingerprint of taking $Poi(k)$ samples closely matches the actual fingerprint entries $\mathcal{F}$.

Intuition. *Discretize the histogram support; choose* $0 < x_1 < \ldots < x_m < 1$. *The LP variables are* $v = v_1, \ldots, v_m$, *where* v_i *is thought of as* $h(x_i)$. *Given a set of* k *samples having fingerprint* $\mathcal{F}$*:*
Find $v_1, \ldots, v_m \geq 0$ *satisfying:*

1. $\sum_{i=1}^{m} x_i v_i = 1$ *(total probability is 1)*
2. $\forall j$, $\sum_{i=1}^{m} v_i \cdot poi(x_i k, j) \in [\mathcal{F}(j) - k^{.6}, \mathcal{F}(j) + k^{.6}]$. *(expected fingerprints are close to* $\mathcal{F}$*)*

There is one slight complication: assume that there is some element ℓ that occurs very frequently—say with probability $1/2$, thus $h(1/2) = 1$. Thus $E[\mathcal{F}(k/2)] \approx 1/\sqrt{k}$, though $\mathcal{F}(k/2)$ will not be concentrated about this expectation since $\mathcal{F}(k/2)$ will either be zero, or one. The number of times ℓ occurs in the sample will be tightly clustered about its expectation of $k/2$, however this type of concentration is quite different from having $\mathcal{F}(i)$ tightly concentrated about its expectation, as is the case in the infrequently occurring portion of the fingerprints.

To capitalize on these two types of concentration—the concentration about $E[\mathcal{F}(i)]$ for small i, and concentration in the number of occurrences of frequently occurring elements—we deal with these two regimes separately. For the frequently-occurring elements, say elements whose probabilities are at least k^{-1+a}, for some small constant $a \in (0, 1)$, we can simply let the returned histogram, h', agree with the empirical distribution, namely setting $h'(j/k) = \mathcal{F}(j)$. For the portion of h' below probability k^{-1+a}, we would like the fingerprint expectations for samples from h' to roughly agree with the observed fingerprints $\mathcal{F}(j)$ in this regime (roughly, for $j \leq k \cdot k^{-1+a} = k^a$).

To avoid the issues which may arise near the threshold between the "low probability" and "high probability" regimes , we choose the location of this threshold so as to have relatively little probability mass in the nearby region.

Given a k-sample fingerprint $\mathcal{F}$, choose $c \in [1, 2]$ such that the total "mass" in $\mathcal{F}$ between frequencies ck^a and $ck^a + 4k^{.6a}$ is at most $4k^{-.4a}$. Namely, $\sum_{j=\lceil ck^a \rceil}^{\lceil ck^a + 4k^{.6a} \rceil} j\mathcal{F}(j) \leq 4k^{1-.4a}$. Note that such a choice of c can be found, for otherwise the total number of samples accounted for by fingerprint entries in the interval $[k^a, 2k^a]$ would exceed k.

We now formally define our linear program. Let $a = 1/50$.

LINEAR PROGRAM 1.
Given a k-sample fingerprint $\mathcal{F}$;
Let $A := ck^{-1+a}$, $B := 4k^{-1+.6a}$, and $\gamma := k^{-3/2}$, the LP variables $v_x \geq 0$ for all $x \leq A + B/2$ in the set $X := \{\gamma, 2^2\gamma, 3^2\gamma, 4^2\gamma, \ldots, A + B/2\}$.

Find v_x satisfying:

1. $\sum_{x \in X : x \geq A} x v_x \leq 16k^{-.4a}$
2. $\sum_{x \in X} x v_x + \sum_{j \geq k(A+B)} \frac{j}{k} \mathcal{F}(j) = 1$
3. **For all integers** $i \leq k(A + B/4)$,

$$\sum_{x \in X} v_x poi(kx, i) \in \left[\mathcal{F}(i) - 4k^{.6+a}, \mathcal{F}(i) + 4k^{.6+a}\right].$$

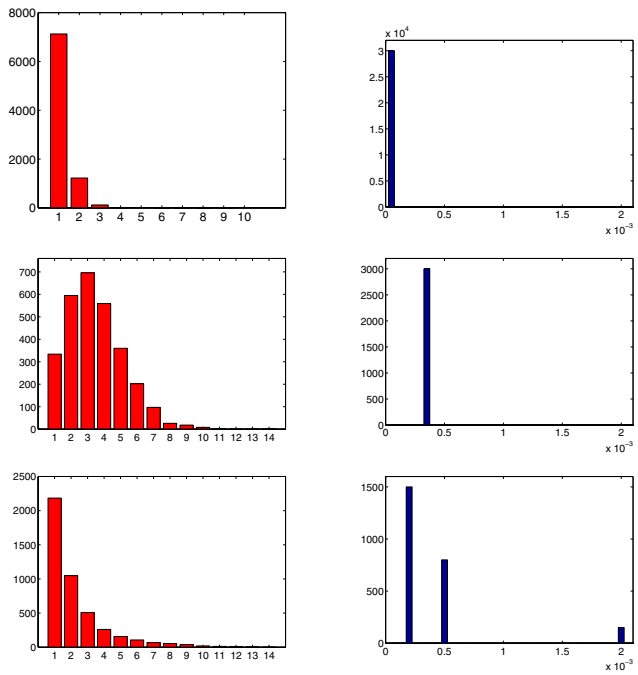

Figure 2: Three fingerprints (left column) derived from 10,000 samples, together with the corresponding histograms (right column) from which the samples were drawn. Intuitively, our estimator is solving the inversion problem: given a fingerprint, it finds a histogram from which the samples could, plausibly, have been taken.

We consider a solution of this linear program to be the low-probability portion of a generalized histogram. In words, the first condition guarantees that there is relatively little probability mass near the "threshhold" probability $A \approx k^{-1+a}$. The second condition guarantees that if we adjoin the empirical distribution from $\mathcal{F}$ above the threshhold probability to the linear program solution, the total probability mass will be 1. The third condition guarantees that if we let Y be a set of $Poi(k)$ samples from the distribution corresponding to this "histogram", for each positive integer $i \leq k(A + B/4) \approx k^a$, $E[\mathcal{F}_Y(i)] \approx \mathcal{F}(i)$, up to a slight margin of error.

We remark that we carefully chose the set X of probabilities for which we solve. If we instead take the set X to be a very fine mesh—for example $\{\frac{1}{k^2}, \frac{2}{k^2}, \ldots, 1\}$—several of the proofs would simplify, but then the computation time to solve the resulting linear program would be $O(k^7)$. We instead opt to take a coarse quadratically-spaced mesh so as to minimize the number of variables for which we solve. Perhaps coincidentally, while our approach seems to require at least $k^{1/4}$ variables in the LP, we use $|X| = \Theta(k^{\frac{1}{4}+a}) \leq k^{1/3.5}$ variables, and thus the LP can be solved in time linear in k, the number of samples [28].

Given a solution to the linear program v, the definition below extends v to yield the histogram h^v, which we refer to as the *histogram associated to the solution* v. Roughly, to obtain h^v, we start with v and first adjoin the empirical distribution for probabilities above $A + B$, then round each value down to the nearest integer. Finally, to compensate for the decrease in mass resulting from the rounding, we scale the support by a factor of $1 + \epsilon$ (while keeping the values of

the histogram fixed) thereby increasing the total mass in the histogram by a factor of $(1+\epsilon)$, where ϵ is chosen so as to make the total probability mass equal 1 after the rounding. We formalize this process below:

DEFINITION 10. *Let $X = \{\gamma, 2^2\gamma, 3^2\gamma, 4^2\gamma, \ldots, A+B/2\}$ be the set of probabilities for which the linear program solves. Given a k-fingerprint $\mathcal{F}$ and a solution v to the associated linear program, the corresponding histogram h^v is derived from v according to the following process in which generalized histogram h' is constructed, then rounded to create h^v.*

1. *set $h'(*) = 0$ and $h^v(*) = 0$.*
2. *for all $x \in X$, let $h'(x) := v_x$.*
3. *for all integers $j \geq k(A+B)$, let $h'(j/k) := \mathcal{F}(j)$.*
4. *for all probabilities $x : h'(x) \neq 0$, set $h^v((1+\epsilon)x) := \lfloor h'(x) \rfloor$, where $\epsilon := \frac{\sum_{x \in X} x(v_x - \lfloor v_x \rfloor)}{1 - \sum_{x \in X} x(v_x - \lfloor v_x \rfloor)}$.*

Note that the recovered histogram h^v is, in fact a histogram, since $h^v : (0,1] \to \mathbb{Z}$, and, because of the last step, $\sum_{y:h^v(y)\neq 0} y h^v(y) = 1$.

ALGORITHM: 1. THE ESTIMATOR
Given a set of k samples having fingerprint $\mathcal{F}$:

- Construct the linear program of Definition 1 corresponding to $\mathcal{F}$.
- Find a solution v to the the linear program. If no solution exists, output FAIL.
- Output histogram h^v associated to solution v, as defined in Definition 10.

The correctness of our estimator is captured in the following proposition, which implies Theorem 1:

PROPOSITION 1. *For a constant $\delta \in (0,1]$, consider a sample consisting of k independent samples from a distribution h of support size at most $\delta k \log k$. With probability at least $1 - e^{-k^{.04}}$, the linear program of Definition 1 has a solution and furthermore, for any solution to the linear program, v, the associated histogram h^v constructed from v in Definition 10 satisfies*

$$R(h, h^v) = O(\sqrt{\delta} \cdot \max\{1, |\log \delta|\}).$$

The proof of Theorem 1 has two parts. In the first part, we show that, with the claimed probability, the linear program has a feasible point v, whose associated histogram h^v is close in relative earthmover distance to the true distribution, h. This part is straightforward—intuitively, as long as our grid of probabilities X is sufficiently fine, the true histogram, rounded so as to have support X, should be a feasible point.

In the second part of the proof we argue that for any two solutions to the linear program v, w, their associated histograms are close in relative earthmover distance. This is the part of the proof where the benefit of our linear program becomes apparent, and the $n/logn$ sample complexity is exposed. Unsurprisingly, nearly all of the technical challenge of our positive results lie in this portion of the proof. Unfortunately, in this extended abstract, we do not have space to even set up the intuition behind this argument, though we note that it hinges upon a Chebyshev polynomial construction, which may be of independent interest.

5. LOWER BOUNDS FOR PROPERTY ESTIMATION

In this section we outline the derivation of our lower bounds, which rests crucially on the central limit theorem for generalized multinomial distributions, Theorem 4.

We provide an explicit construction via Laguerre polynomials of two distributions, p^+, p^- that are close—in the relative earthmover metric—to uniform distributions respectively on n and $\frac{n}{2}$ elements, for $n = \Theta(k \log k)$. The crucial and elusive property of the pair p^+, p^- which we will explain over the course of this section is that the result of drawing $Poi(k)$ samples from p^+ will be information-theoretically indistinguishable, with high probability, from sampling instead from p^-.

Perhaps the most naive way to try to distinguish samples from p^+ versus from p^- is via their fingerprint expectations. So the first step to constructing indistinguishable distributions is to ensure that the corresponding vectors of fingerprint expectations are approximately equal. As we show, this is essentially the *only* step, though proving that the construction is this "easy" requires considerable work.

5.1 Fourier Analysis, Hermite Polynomials, and "Fattening"

As introduced in Section 1.1, generalized multinomial distributions capture the distribution of fingerprints induced by drawing $Poi(k)$ samples from a given distribution. And thus the final step of the proof that p^+ and p^- are indistinguishable in $Poi(k)$ samples will be to apply the central limit theorem for generalized multinomial distributions (Theorem 4) to the distributions of fingerprints of p^+, p^- respectively, approximating each as a discretized Gaussian. This will be sufficient provided a) the Gaussians are sufficiently similar, and b) the statistical distance bound when Theorem 4 is applied is suitably small. We analyze each part in turn.

5.1.1 Similar Expectations Induce Similar Covariances

For two Gaussians to be statistically close, three things should hold: the Gaussians have similar expectations; the Gaussians have similar covariances; and the minimum eigenvalue of the covariance matrix must be large. This third condition we defer to Section 5.1.2. In this section we argue the intuition for the somewhat surprising fact that, in the present setting, similar expectations induce similar covariances.

Recall the effect of a single element of probability x on the distribution of fingerprints: for each integer $i > 0$, with probability $poi(xk, i)$, that element will occur i times in the sample and hence end up incrementing the ith fingerprint entry by 1. Thus the contribution of this element to the expectation of the ith fingerprint entry equals $poi(xk, i)$.

Similarly, since covariance adds for sums of independent distributions, we may compute the contribution of an element of probability x to the (i,j)th entry of the fingerprint covariance matrix, which we compute here for the case $i \neq j$. The covariance of random variables X, Y is $E[XY] - E[X]E[Y]$; since in our case X represents the event that the distribution element is sampled i times, and Y represents the event that it is sampled j times, $E[XY] = 0$ as they can never both occur. Thus the contribution to the

covariance is just

$$poi(xk,i)poi(xk,j) = \frac{(xk)^{i+j}}{e^{2xk}i!j!} = \frac{\binom{i+j}{i}}{2^{(i+j)}} poi(2xk, i+j).$$

Our claim that similar expectations imply similar covariances may now be rephrased as: each "skinny poisson" function $poi(2xk, \ell)$ can be approximated as a linear combination of "regular poisson" functions $\sum_i \alpha_{i,\ell} poi(xk, i)$, with small coefficients. Specifically, the coefficients $\alpha_{i,\ell}$ allow one to approximate the fingerprint covariances as a linear function of the fingerprint expectations; if one matches, then so does the other. In fact, one can approximate such a "skinny poisson" to within ϵ as a sum of regular poissons using coefficients of total magnitude (roughly) no more than $\frac{1}{\epsilon}$, indeed, for intuitively the same reasons that the analogous claim holds true for Gaussians. As opposed to the relatively simple case of Gaussians, proving our claim is perhaps the most technical part of this paper, making heavy use of Hermite polynomials in Fourier space.

5.1.2 CLT Performance

If we apply Theorem 4 to the distribution of the first m fingerprint entries, and the covariance matrix of the distribution of these fingerprint entries has minimum eigenvalue σ^2, then the resulting bound on the statistical distance is $\frac{m^{4/3}}{\sigma^{1/3}}$ times logarithmic factors. Since σ^2 is never going to exceed order of k, we clearly cannot use $m = k$. That is, we must apply the central limit theorem to only a small subset of the fingerprints. Additionally, we must ensure that σ^2 is big for this portion—intuitively that the distribution of these fingerprints is "fat in every direction".

Set $m = \log k$. We assume that p^+ and p^- are constructed so as to be supported on probabilities at most $\frac{\log k}{8k}$, and have similar fingerprint expectations and covariances. This bound of $\frac{\log k}{8k}$ ensures that we will almost never see any element of p^+ or p^- more than $\log k$ times; that is, the portion of the fingerprint below m "captures the whole story". However, if we were to try to apply the central limit theorem at this stage, the bound would be horrendous because the variance in the higher fingerprints (say the mth), is tiny. Thus we "fatten" the distributions of fingerprints by smearing a small $(1/polylog(k))$ amount of the probability mass in p^+ and p^- uniformly among probabilities, up to m/k. Because we fatten p^+ and p^- *identically*, their fingerprint expectations and covariances still closely match. Given the fattened pair of distributions, we can now obtain satisfactory bounds from our central limit theorem. To complete the argument, we make use of the natural coupling of the portions of the fingerprints above m, stemming from the identical fattened portions of the distributions p^+, p^-.

Thus the Hermite polynomial argument guarantees matching covariances; "fattening" in conjunction with our central limit theorem for generalized multinomial distributions guarantees all the rest. What remains is to construct p^+, p^- with matching fingerprint expectations.

5.2 The Laguerre Construction

We will construct the pair of histograms, p^+, p^- explicitly, via Laguerre polynomials. We begin by letting p^+, p^- be the uniform distributions over support n and $n/2$, respectively. We then modify p^+, p^- by transferring some of the probability mass to make elements with higher probabilities, so as to ensure that the fingerprint expectations of $Poi(k)$ samples from p^+ and p^- roughly agree.

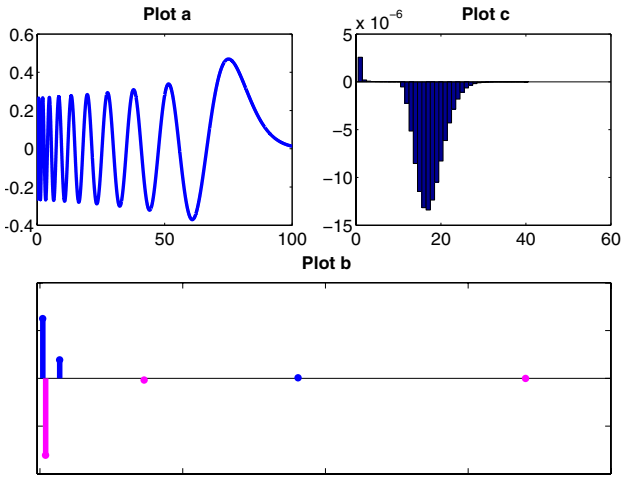

Figure 3: a) The 10th Laguerre polynomial, multiplied by $e^{-x/2}x^{1/4}$, illustrating that it behaves as $a \cdot e^{x/2}x^{-1/4} \cdot \sin(b \cdot \sqrt{x})$ for much of the relevant range.
b) $f(x)$, representing histograms $p^+(x), p^-(x)$ respectively above and below the x-axis.
c) The discrepancy between the first 40 fingerprint expectations of p^+, p^-; the first 10 expected fingerprint entries almost exactly match, while the discrepancy in higher fingerprint expectations is larger, though still bounded by $2 \cdot 10^{-5}$.

The condition that the expected ith fingerprint entries of p^+ and p^- agree is simply that $\sum_{x:p^+(x)\neq 0} p^+(x) poi(kx, i) = \sum_{x:p^-(x)\neq 0} p^-(x) poi(kx, i)$. Equivalently, define the function $f(x) : [0,1] \to \mathbb{R}$ by $f(x) = p^+(x) - p^-(x)$. The condition that p^+ and p^- have the same expected first j fingerprints can be expressed as $\sum_{x:f(x)\neq 0} f(x) poi(kx, i) = 0$, for all integers $i \leq j$. Since $poi(kx, i) := \frac{e^{-kx}k^i x^i}{i!}$, this condition is equivalent to the function $g(x) := f(x)e^{-kx}$ being *orthogonal* to polynomials of degree at most j. The following easily verified fact outlines an approach to creating such a function.

FACT 2. *Given a polynomial P of degree $j+2$ whose roots $\{x_i\}$ are real and distinct, letting P' be the derivative of P, then for any $\ell \leq j$ we have $\sum_{i=1}^{j+2} \frac{x_i^\ell}{P'(x_i)} = 0$.*

To construct $f(x)$, choose a polynomial $P(x) = (x - 1/n)(x - 2/n)\prod_{i=1}^{j}(x - r_i)$, for some set of j distinct values r_i, with $2/n < r_i < 1$, then let $g(x)$ be the function that is supported at the roots of P, and takes value $1/P'(x)$ for the $j+2$ values of x for which $P(x) = 0$. To obtain $f(x)$, simply set $f(x) = g(x)e^{kx}$.

If we interpret the positive portion of $f(x)$ as p^+ and the negative portion as p^-, we will, by Fact 2, have two histograms whose first j fingerprint expectations agree. Additionally, p^+ will have some probability mass at probability $1/n$, and p^- will have some probability mass at $2/n$.

The tricky part, however, is in picking the r_i so as to ensure that *most* of the probability mass of h_1 is on probability $1/n$, and most of the mass of h_2 is on probability $2/n$. If this is not the case, then p^+ and p^- will not be close

(in relative–earthmover distance) to the uniform distributions over n and $n/2$ elements, respectively and thus may have similar entropies, or support sizes, failing us as a lower bound. Further complicating this task, is that whatever weight is at $x > 2/n$ in $g(x)$, ends up being multiplied by e^{kx}. To offset this exponential increase, we should carefully choose the polynomial P so that the inverses of its derivatives, $1/P'(x)$, decay exponentially when evaluated at roots x of P. Such polynomials are hard to come by; fortunately, the Laguerre polynomials have precisely this property.

6. REFERENCES

[1] J. Acharya, A. Orlitsky, and S. Pan. The maximum likelihood probability of unique-singleton, ternary, and length-7 patterns. In *IEEE Symp. on Information Theory*, 2009.

[2] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *J. Comput. System Sci.*, 58:137–147, 1999.

[3] Z. Bar-Yossef, T. Jayram, R. Kumar, D. Sivakumar, and L. Trevisan. Counting distinct elements in a data stream. In *Proc. 6th Workshop on Rand. and Approx. Techniques*.

[4] Z. Bar-Yossef, R. Kumar, and D. Sivakumar. Sampling algorithms: Lower bounds and applications. In *STOC, 2001*.

[5] A. Barbour and L. Chen. *An Introduction to Stein's Method*. Singapore University Press, 2005.

[6] T. Batu. Testing properties of distributions. *Ph.D. thesis, Cornell University, 2001*.

[7] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. In *STOC, 2002*.

[8] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. The complexity of approximating the entropy. *SIAM Journal on Computing*, 2005.

[9] T. Batu, L. Fortnow, R. Rubinfeld, W. Smith, and P. White. Testing that distributions are close. In *FOCS, 2000*.

[10] K. Beyer, P. J. Haas, B. Reinwald, Y. Sismanis, and R. Gemulla. On synopses for distinct-value estimation under multiset operations. In *ACM SIGMOD Int. Conf. on Management of Data, 2007*.

[11] R. Bhattacharya and S. Holmes. An exposition of Götze's estimation of the rate of convergence in the multivariate central limit theorem. *Stanford Department of Statistics Technical Report*, 2010-02, March 2010.

[12] J. Bunge. Bibliography of references on the problem of estimating support size, available at http://www.stat.cornell.edu/~bunge/bibliography.html.

[13] J. Bunge and M. Fitzpatrick. Estimating the number of species: A review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.

[14] A. Chakrabarti, G. Cormode, and A. McGregor. A near-optimal algorithm for computing the entropy of a stream. In *SODA, 2007*.

[15] M. Charikar, S. Chaudhuri, R. Motwani, and V. Narasayya. Towards estimation error guarantees for distinct values. In *PODS, 2000*.

[16] S. Chatterjee. Personal communication. May 2010.

[17] S. Chatterjee and E. Meckes. Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Am. J. Probab. Math. Stat.*, 4:257–283, 2008.

[18] C. Daskalakis and C. H. Papadimitriou. Computing equilibria in anonymous games. In *FOCS*, 2007.

[19] C. Daskalakis and C. H. Papadimitriou. Discretized multinomial distributions and Nash equilibria in anonymous games. In *FOCS*, 2008.

[20] R. Fisher, A. Corbet, and C. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of the British Ecological Society*, 1943.

[21] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika, 40(16):237–264*, 1953.

[22] F. Götze. On the rate of convergence in the multivariate CLT. *Annals of Probability*, 19(2):724–739, 1991.

[23] S. Guha, A. McGregor, and S. Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *SODA, 2006*.

[24] L. Gurvits. On Newton(like) inequalities for multivariate homogeneous polynomials. http://www.optimization-online.org/DB_FILE/2008/06/1998.pdf, 2008.

[25] N. Harvey, J. Nelson, and K. Onak. Sketching and streaming entropy via approximation theory. In *FOCS, 2008*.

[26] P. Indyk and D. Woodruff. Tight lower bounds for the distinct elements problem. In *FOCS, 2003*.

[27] D. Kane, J. Nelson, and D. Woodruff. An optimal algorithm for the distinct elements problem. In *PODS, 2010*.

[28] N. Karmarkar. A new polynomial time algorithm for linear programming. *Combinatorica*, 4(4):373–395, 1984.

[29] D. A. McAllester and R. Schapire. On the convergence rate of Good-Turing estimators. In *COLT*, 2000.

[30] A. Orlitsky, N. Santhanam, K.Viswanathan, and J. Zhang. On modeling profiles instead of values. *Uncertainity in Artificial Intelligence*, 2004.

[31] A. Orlitsky, N. Santhanam, and J. Zhang. Always Good Turing: Asymptotically optimal probability estimation. In *FOCS, 2003*.

[32] A. Orlitsky, N. Santhanam, and J. Zhang. Always Good Turing: Asymptotically optimal probability estimation. *Science*, 302(5644):427–431, October 2003.

[33] L. Paninski. Estimation of entropy and mutual information. *Neural Comput.*, 15(6):1191–1253, 2003.

[34] L. Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE Trans. on Information Theory*, 50(9):2200–2203, 2004.

[35] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM J. Comput.*, 39(3):813–842, 2009.

[36] G. Reinert and A. Röllin. Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. *Annals of Probability*, 37(6):2150–2173, 2009.

[37] B. Roos. On the rate of multivariate Poisson convergence. *Journal of Multivariate Analysis, 69:120-134, 1999*.

[38] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. *Proc. Sixth Berkeley Symp. on Mathematical Statistics and Probability*, 2, 1972.

[39] G. Valiant and P. Valiant. A CLT and tight lower bounds for estimating entropy. http://www.eccc.uni-trier.de/report/2010/179/, 2010.

[40] G. Valiant and P. Valiant. Estimating the unseen: a sublinear-sample canonical estimator of distributions. http://www.eccc.uni-trier.de/report/2010/180/, 2010.

[41] P. Valiant. Testing symmetric properties of distributions. In *STOC, 2008*.

[42] A. Wagner, P. Viswanath, and S. Kulkami. Strong consistency of the Good-Turing estimator. In *IEEE Symp. on Information Theory*, 2006.

[43] A. Wagner, P. Viswanath, and S. Kulkami. A better Good-Turing estimator for sequence probabilities. In *IEEE Symp. on Information Theory*, 2007.

[44] D. Woodruff. The average-case complexity of counting distinct elements. In *The 12th Int. Conf. on Database Theory, 2009*.