Electronic Colloquium on Computational Complexity, Report No. 179 (2010)

A CLT and tight lower bounds for estimating entropy

Gregory Valiant Paul Valiant

November 17, 2010

Abstract

We prove 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). 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. This central limit theorem relates a generalized multinomial distribution to a multivariate Gaussian distribution, discretized by rounding to the nearest lattice points. In contrast to the metric of our first central limit theorem, this bound is in terms 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.

In the second part of the paper, we employ this central limit theorem to establish a lower bound of $\Omega(\frac{n}{\log n})$ on the sample complexity of additively estimating the entropy or support size of a distribution (where $1/n$ is a lower bound on the probability of any element in the domain). Together with the canonical estimator constructed in the companion paper [33], this settles the longstanding open question of the sample complexities of these estimation problems, up to constant factors. In particular, for any constant c, there is a pair of distributions D, D' each of whose elements occurs with probability at least $1/n$, and whose entropies satisfy $H(D) - H(D') > c$, such that no algorithm on $o(\frac{1}{c|\log c|}\frac{n}{\log n})$ samples can distinguish D from D' with probability greater than $2/3$, and analogously for the problem of estimating the support size. The previous lower-bounds on these sample complexities were $n/2^{\Theta(\sqrt{\log n})}$, for constant c, from [34]. We explicitly exhibit such a pair of distributions via a Laguerre polynomial construction that may be of independent interest.

ISSN 1433-8092

1 Introduction

Perhaps the chief triumph of modern statistics is the central limit theorem. During the past century, our understanding of the various settings for which central limit theorems apply has expanded immensely. Nevertheless, 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. So perhaps it is not surprising that the particularly natural and useful versions we prove here seem absent from the statistics literature[13].

We prove 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, [16, 17, 34, 30]).

We then apply this central limit theorem to the problem of lower bounding the sample complexity of additively estimating the entropy, or support size, of a distribution. These two estimation problems have a long history of study in statistics and computer science, and have practical applications across many fields, including Biology, Ecology, Genetics, Linguistics, Neuroscience, and Physics (see, for example, the list of hundreds of references in [10], and the discussion in [26]). Despite much work, the sample complexity of these estimation problems was still largely open. One explanation for the weakness in the lower bounds was the lack of a characterization for the distribution over sets of sample. Our central limit theorem for generalized multinomial distributions provides precisely this characterization.

We leverage this central limit theorem—albeit through considerable additional effort involving two polynomial constructions that may be of independent interest, one involving Laguerre polynomials, one involving Hermite polynomials—to establish a lower bound of $\Omega(\frac{n}{\log n})$ on the sample complexity of additively estimating the entropy, or support size, of a distribution (where n is a bound on the support size[1]). Together with the canonical estimator constructed in the companion paper [33], this settles the longstanding open question of the sample complexities of these estimation problems (up to constant factors). In particular, we show that for any constant c, there is a pair of distributions D, D' with $H(D) - H(D') > c$, such that no algorithm on $o(\frac{1}{c|\log c|}\frac{n}{\log n})$ samples can distinguish D from D' with probability greater than 2/3, and analogously for the problem of estimating the support size. Additionally, D and D' have supports of size at most n, and each element in their domains occurs with probability at least $\frac{1}{n}$. The previous lower-bounds on these sample complexities, for constant c, were $n/2^{\Theta(\sqrt{\log n})}$, given by Valiant in [34], 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.* [28].

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

[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 a lower bound on this probability it is impossible to estimate support size.

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 $H(D')$.

1.1 Related Work

Since Stein's seminal paper [31], presented in 1970, in which he described 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. There have been several successful applications of Stein's method in multivariate settings[19, 14, 29]. We closely follow the treatment for the multivariate limit theorem given by Götze in [19] (see also [9] 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.

The problem of estimating an unknown discrete distribution from few samples has a very rich history of study in both Statistics and Computer Science. The specific problem of estimating the support size of an unknown distribution (also referred to as the problem of estimating the number of species in a population) has a very long history of study and arises in many contexts (see [10] 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 this problem to a multiplicative factor of α are given in [3, 12] though they are somewhat unsatisfying as the worst-case instance consists of 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 [28], 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. Prior to the upper bound of $O(\frac{n}{\log n})$ matching our lower bound, shown in the companion paper [33], 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.* [5, 6, 15], Guha *et al.* [20], and Valiant [34] considered the problem of multiplicatively estimating the entropy. In [26, 27], Paninski proved, non-constructively, the existence of a sublinear sample estimator for additively approximating the entropy. The best previous lower bound of $n/2^{\Theta(\sqrt{\log n})}$ is given in [34].

There has also been considerable interest and work in the related problems of estimating these properties in the *streaming* model in which one has access to very little memory and can perform only a single pass over the data [1, 2, 8, 11, 22, 23, 24, 35].

1.2 Outline

The paper is divided into two parts. In Section 2 we introduce our two central limit theorems, which are presented in full in Appendices A and B. These appendices are self-contained, and include all necessary definitions.

In Section 3 we state and prove our lower bounds for the sample complexity of estimating entropy and support size. We now state some definitions and examples used in the body of the paper.

1.3 Definitions and Examples

We state the key definitions, and provide some illustrative examples.

Definition 1. *A* distribution *on $[n] = \{1, \ldots, n\}$ is a function $p : [n] \rightarrow [0, 1]$ such that $\sum_i p(i) = 1$. 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. (Specifically, it will prove helpful to consider a set of samples whose size is distributed according to a Poisson process of expectation k, as discussed in Section 1.3.2.) For a distribution $D \in \mathcal{D}^m$, we denote the entropy $H(D) := -\sum_i p(i) \log p(i)$, and the support size $S(D) := |\{i : p(i) > 0\}|$.

Entropy and support size are *symmetric* properties, in that their value is invariant to relabeling the domain: for any distribution D, and any permutation σ, $H(D) = H(D \circ \sigma)$, and similarly for the support size.

Definition 2. *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.*

Note that for the two properties in question, 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 [5, 7] for an easy proof for general symmetric properties). Note that in some of the literature the fingerprint is alternately termed the *pattern*, *histogram*, or *summary statistics* of the sample.

Analogous to the fingerprint of a set of samples, is what we call the *histogram of the distribution*, which captures the number of domain elements that occur with each frequency. Any symmetric property is clearly a function of only the histogram of the distribution.

Definition 3. *The* histogram *of a distribution p is a mapping $h : (0, 1] \rightarrow \mathbb{Z}$, where $h(x) = |\{i : p(i) = x\}|$.*

For clarity of exposition, we often relax the above definition to allow histograms $h : (0, 1] \rightarrow \mathbb{R}$, that do not take integral values. For the range of parameters that we use, the rounding issues that arise are insignificant.

We now define what it means for two distributions to be "close"; because the values of the properties in question 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.

Definition 4. *We define the* relative earthmover distance *between two histograms of distributions, $R(h_1, h_2)$, 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 the statistical distance is upper bounded by relative earthmover distance. The following easy fact shows that entropy and support size are well-behaved with respect to the relative earthmover distance:

Fact 5. *For any pair of distribution h, h' with $R(h,h') \leq \delta$, we have $|H(h) - H(h')| \leq \delta$, and $|S(h) - S(h')| \leq n\delta$, where $\frac{1}{n} \leq \min\{x : h(x) \neq 0, \text{ or } h'(x) \neq 0\}$.*

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. Additionally, for integers $i \geq 0$, we refer to the function $poi(x, i)$, viewed as a function of the variable x, as the jth *Poisson function.*

1.3.1 Examples

We now provide two clarifying examples of the above definitions:

Example 6. *Consider a sequence of fish species, found as samples from a certain lake $X = (a, b, a, c, c, d, a, e, b)$, where each letter denotes a distinct fish species. We have $\mathcal{F}_X = (2, 2, 1)$, indicating that two species occurred exactly once (species d and e), two species occurred exactly twice (species b and c), and one species occurred exactly three times (species a).*

Suppose that the true distribution of fish is the following:

$$Pr(a) = 1/2, \quad Pr(b) = 1/4, \quad Pr(c) = Pr(d) = Pr(e) = 1/12.$$

The associated histogram *of this distribution is $h : \mathbb{R}^+ \to \mathbb{Z}$ defined by $h(1/12) = 3$, $h(1/4) = 1$, $h(1/2) = 1$, and for all $x \notin \{1/12, 1/4, 1/2\}$, $h(x) = 0$. If we now consider a second distribution over $\{j, k, \ell\}$ defined by the probabilities $Pr(j) = 1/2, \quad Pr(k) = 1/4, \quad Pr(\ell) = 1/4$, and let h' be its associated histogram, then the relative earthmover distance $R(h, h') = \frac{1}{4}|\log \frac{1/4}{1/12}|$, since we must take all the mass that lies on frequency $1/12$ and move it to frequency $1/4$ in order to turn the first distribution into one that yields a histogram identical to h'.*

Example 7. *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, 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 $\mathcal{F}(i) \approx n \cdot poi(5, i)$.*

1.3.2 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 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 7, that the number 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.

1.3.3 Generalized Multinomial Distributions

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.

2 Two Multivariate Central Limit Theorems

We introduce our two main central limit theorems here, though see Appendix A for the full presentation.

Our aim is to prove a central limit theorem that approximates the discrete generalized multinomial distribution in the statistical distance metric. The main tool is Stein's method, which is uniquely well-suited to the task of comparing distributions to Gaussians.

We note, however, that *prima facie* the statistical distance between a multinomial and a Gaussian is 1. This is simply because the multinomial is a discrete distribution, and thus is not close in a distributional sense, to any smooth distribution. We must thus conduct the analysis in two parts.

2.1 A Multivariate CLT for Earthmover Distance (See Appendix A)

Stein's method is in some sense very dependent on *smoothness* of the target distribution, in our case, a multivariate Gaussian. It represents the Gaussian as the distribution induced by a certain random walk—that is, the Gaussian is the *stationary distribution* of the random walk. It compares the given distribution S to the Gaussian by then examining how the random walk would affect S.

The infinitesimal nature of a random walk makes earthmover distance particularly well-suited for analysis by this method. In short, in this part, we show that the generalized multinomial distribution is well-approximated *in the earthmover distance sense* by a Gaussian. In the next part, we leverage several convexity properties of the multinomial and Gaussian distributions to show that this in fact suffices to show that, when rounded to the nearest lattice points, the Gaussian distribution actually approximates the multinomial in the stronger statistical distance sense.

Definition. *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 2. *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 $\beta k(2.7 + 0.83 \log n)$.*

2.2 A CLT for Generalized Multinomial Distributions (See Appendix B)

In this section we leverage the central limit theorem of Theorem 2 to show our second central limit theorem that bounds the *statistical distance*, denoted by D_{tv} between generalized multinomial distributions and (discretized) Gaussian distributions. While Theorem 2 certainly applies to generalized multinomial distributions, the goal of this section is to derive a bound in terms of the rather more stringent statistical distance. The main hurdle is relating the "smooth" nature of the Gaussian distribution and earthmover distance metric to the "discrete" setting imposed by a statistical distance comparison with the discrete generalized multinomial distribution.

The analysis to compare a Gaussian to a generalized multinomial distribution proceeds in two steps. Given the earthmover distance bound provided by Theorem 2, we first smooth both sides via convolution with a suitably high-variance distribution to convert this bound into a statistical distance bound, albeit not between the original two distributions but between convolved versions of them. The second step is via a "deconvolution" lemma that relies on the unimodality in each coordinate of generalized multinomial distributions.

The central limit theorem that we leverage in the rest of the paper to prove property testing lower bounds is the following:

Definition. *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.*

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.

3 Lower Bounds for Property Estimation

In this section we use the central limit theorem for generalized multinomial distributions, Theorem 4, to show our lower bounds for property testing.

We provide an explicit construction via Laguerre polynomials of two distributions that are close, in the relative earthmover metric, to uniform distributions respectively on n and $\frac{n}{2}$ elements, for $n = \Theta(k \log k)$. This pair of distributions is constructed to have the additional property that their fingerprint expectations are very close. As we aim to approximate the distributions of fingerprints by Gaussians, which are parameterized by their mean and covariance matrices, we must argue that the covariance matrices corresponding to these two distributions are also very close. We prove this via a general result that applies to all distributions, and not just the constructed pair; the proof appears in Appendix C and relies heavily on Hermite polynomials.

Applying the central limit theorem requires one additional construction. Because the convergence bound in Theorem 4 is in terms of the smallest eigenvalue of the covariance matrix, in order to obtain a satisfactory bound, we "fatten" each distribution so that it has sufficient variance in every direction. Such a "fattening" changes both distributions correspondingly, and has only a small effect on the distributions under the relative earthmover metric.

Our theorem, which we prove over the course of this section is the following:

Theorem 1. *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.*

Specifically, for any constant $\epsilon > 0$, there exists a pair of distributions of support at most n and for which each domain element occurs with probability at least $1/n$, satisfying:

1. *$|H(p^+) - H(p^-)| \geq \epsilon$*
2. *$|S(p^+) - S(p^-)| \geq n\epsilon$, where $S(D) := |\{x : \Pr_D[x] > 0\}|$*
3. *No algorithm can distinguish a set of $O\left(\frac{n}{\epsilon|\log \epsilon| \log n}\right)$ samples from p^+ versus p^- with probability $2/3$.*

We will construct the p^+, p^- of the theorem explicitly, via Laguerre polynomials. We now state the properties of these polynomials that we will use.

Let $L_j(x)$ denote the jth Laguerre polynomial, defined as $L_j(x) = \frac{e^x}{j!}\frac{d^j}{dx^j}\left(e^{-x}x^j\right)$.

Fact 9. *For each integer $j \geq 0$,*

1. *For $x \in [0, \frac{1}{j}]$, $L_j(x) \in [1 - jx, 1]$;*
2. *L_j has j real roots, all lying in $[\frac{1}{j}, 4j]$;*
3. *Letting x_i denote the ith root of L_j, for $i \in \{1, \ldots, j\}$, we have $x_i \geq \frac{i^2}{3j}$;*
4. *For $i < j/2$, $|\frac{dL_j(x)}{dx}(x_i)| \geq \frac{e^{x_i/2} j^{1/4}}{2x_i^{3/4}}$ and for any i, $|\frac{dL_j(x)}{dx}(x_i)| \geq \frac{e^{x_i/2}}{\sqrt{\pi} x_i^{3/4}}$.*

Proof. Since L_j is a polynomial of degree j with j positive real roots, none of the inflection points lie below the smallest root. Since $L_j(0) = 1$, $L_j'(0) = -j$, and $L_j''(0) > 0$, we have that $L_j(x) \geq 1 - jx$ for x less than or equal to the smallest root of L_j. Thus the smallest root of L_j must be at most $\frac{1}{j}$, and $L_j(x) \geq 1 - jx$ for $x \leq \frac{1}{j}$. The fact that the largest root is at most $4j$ follows from [32], Theorem 6.32. The third fact appears in [32], p. 129, and the fourth fact follows from p. 100. □

Definition 10. *Given real number $\phi \in (0, \frac{1}{4})$ and letting $j = \log k$, consider the degree $j+2$ polynomial $M_{j,\phi}(x) \triangleq -(x - \phi\frac{1}{j})(x - 2\phi\frac{1}{j})L_j(x)$. Let $v(x)$ be the function that takes value $1/M'_{j,\phi}(x)$ for every x where $M_{j,\phi}(x) = 0$, and is 0 otherwise, where M' is the derivative of M. Define the distributions $p^+_{j,\phi}, p^-_{j,\phi}$ such that for each x where $v(x) > 0$, the distribution $p^+_{j,\phi}$ contains $v(x)e^{x/32}$ probability mass at probability $\frac{1}{32k}x$, and for each x where $v(x) < 0$ the distribution $p^-_{j,\phi}$ contains $|v(x)|e^{x/32}$ probability mass at probability $\frac{1}{32k}x$, where each distribution is then normalized to have total probability mass 1.*

We note that since each element in the support of either $p^+_{\log k,\phi}$ or $p^-_{\log k,\phi}$ is defined to have probability at least $\frac{\phi}{32k\log k}$, both distributions have support at most $\frac{32}{\phi}k\log k$, which we take as n, in the context of both the entropy and the support size problems.

Lemma 11. *Distributions $p^+_{\log k,\phi}$ and $p^-_{\log k,\phi}$ are $O(\phi|\log\phi|)$-close, respectively, in the relative earthmover distance to the uniform distributions on $\frac{32}{\phi}k\log k$ and $\frac{16}{\phi}k\log k$ elements.*

Proof. Letting $j = \log k$, consider the values of $\frac{d}{dx}M_{j,\phi}(x)$ at its zeros. We first consider the two zeros at $\frac{\phi}{j}$ and $2\frac{\phi}{j}$. Note that $-\frac{d}{dx}(x - \phi\frac{1}{j})(x - 2\phi\frac{1}{j}) = -2x + 3\phi\frac{1}{j}$, having values $\pm\phi\frac{1}{j}$ respectively at these two points. By the product rule for differentiation, $\frac{d}{dx}M_{j,\phi}(x)$ at these points is thus respectively $\leq \phi\frac{1}{j}$ and $\geq -\phi\frac{1}{j}$, by the first part of Fact 9.

Let x_i denote the ith zero of L_j. We note that since by definition, $\phi < \frac{1}{4}$, and from Fact 9, each $x_i \geq \frac{1}{j}$, we have $(x_i - \phi\frac{1}{j})(x_i - 2\phi\frac{1}{j}) \geq \frac{3}{8}x_i^2$. At each x_i, we may thus bound $|\frac{d}{dx}M_{j,\phi}(x)| =$

$|(x-\phi\frac{1}{j})(x-2\phi\frac{1}{j})\frac{d}{dx}L_j(x)| \geq \frac{3}{8}x^2\frac{e^{x/2}j^{1/4}}{2x^{3/4}}$ for $i \leq j/2$ and by $\frac{3}{8}x^2\frac{e^{x/2}}{\sqrt{\pi}x^{3/4}}$ otherwise, which we will denote as $\frac{3}{8}e^{x/2}x^{5/4}\left(\frac{j^{1/4}}{2}[i > j/2] + \frac{1}{\sqrt{\pi}}[i \geq j/2]\right)$.

Consider the *unnormalized* versions of $p^+_{j,\phi}, p^-_{j,\phi}$, that is, containing probability mass $|1/\frac{d}{dx}M_{j,\phi}(x)|e^{x/32}$ at each probability $\frac{1}{32k}x$ where $\frac{d}{dx}M_{j,\phi}(x)$ is positive or negative respectively (without scaling so as to make total probability mass be 1). Let c_1, c_2 respectively be the constants that $p^+_{j,\phi}, p^-_{j,\phi}$ respectively must be multiplied by to normalize them. Recall from above that $|\frac{d}{dx}M_{j,\phi}(x)| \leq \phi\frac{1}{j}$ for the point $x = \phi\frac{1}{j}$ in the support of $p^+_{j,\phi}$ and the point $x = 2\phi\frac{1}{j}$ in the support of $p^-_{j,\phi}$, which implies that the probability mass at each of these points is at least $e^{2\phi\frac{1}{j}/32}\frac{j}{\phi} \geq \frac{j}{\phi}$. From these point masses alone we conclude $c_1, c_2 \leq \frac{\phi}{j}$.

We now consider the earthmover cost of moving all the weight of the unnormalized version of $p^+_{j,\phi}$ to $x = \phi\frac{1}{j}$ or all the weight of the unnormalized version of $p^-_{j,\phi}$ to $x = 2\phi\frac{1}{j}$, which we will then multiply by c_1, c_2 respectively. Note that the per-unit-weight relative earthmover cost of moving weight from an x_i to either $x = \phi\frac{1}{j}$ or $x = 2\phi\frac{1}{j}$ is at most $\log|\phi| + \log(jx_i)$. As we have bounded the weight at x_i (for either $p^+_{j,\phi}$ or $p^-_{j,\phi}$) as $\frac{8}{3}e^{-15x_i/32}x_i^{-5/4}\left(\frac{2}{j^{1/4}}[i < \frac{j}{2}] + \sqrt{\pi}[i \geq \frac{j}{2}]\right)$, and since, from Fact 9, $x_i \geq \frac{i^2}{3j}$, we may thus bound the relative earthmover distance by substituting this into the preceding expression, multiplying by the cost $|\log\phi| + \log(jx_i)$ and our bound $c_1, c_2 \leq \frac{\phi}{j}$, and summing over i:

$$\sum_{i=1}^{j}\frac{\phi}{j}\left(|\log\phi| + 2\log i\right)\frac{8}{3}e^{-\frac{5i^2}{32j}}\left(\frac{i^2}{3j}\right)^{-5/4}\left(\frac{2}{j^{1/4}}[i < j/2] + \sqrt{\pi}[i \geq j/2]\right) = O(\phi|\log\phi|)$$

as desired. □

We note the following general fact that we will use to bound the discrepancy in the fingerprint expectations of $p^+_{j,\phi}$ and $p^-_{j,\phi}$.

Fact 12. *Given a polynomial P of degree j whose roots $\{x_i\}$ are real and distinct, letting P' be the derivative of P, then for any $\ell \leq j-2$ we have $\sum_{i=1}^{j}\frac{x_i^\ell}{P'(x_i)} = 0$.*

Proof. We assume, without loss of generality, that P is monic.

To prove this, consider the general prescription for constructing a degree $j-1$ polynomial through j given points (x_i, y_i): $f(x) = \sum_{i=1}^{j} y_i\left(\prod_{m\neq i}(x - x_m)\right)\Big/\left(\prod_{m\neq i}(x_i - x_m)\right)$. We note that the coefficient of x^{j-1} in this polynomial is $f(x) = \sum_{i=1}^{j} y_i\left(\prod_{m\neq i}(x_i - x_m)\right)^{-1}$, where for each i, the expression $\left(\prod_{m\neq i}(x_i - x_m)\right)^{-1}$ is exactly $1/P'(x_i)$. Thus since polynomial interpolation is unique, $\sum_{i=1}^{j}\frac{x_i^\ell}{P'(x_i)}$ computes the x^{j-1} coefficient in the polynomial x^ℓ, which, for $\ell \leq j-2$ equals 0, as desired. □

Fact 13. *(From [18].) For $\lambda > 0$, and an integer $n \geq \lambda$,*

$$\sum_{i=n}^{\infty} poi(\lambda, i) \leq \frac{poi(\lambda, n)}{1 - \lambda/(n+1)}.$$

Lemma 14. *For any i, the ith fingerprint expectations for distributions $p^+_{j,\phi}, p^-_{j,\phi}$ are equal to within $o(1)$.*

Proof. Recall that the expected contribution of an element of probability x to the ith fingerprint entry equals $poi(xk, i)$.

Consider, as in the proof of Lemma 11, the unnormalized versions of $p^+_{j,\phi}, p^-_{j,\phi}$, that is, containing weight $|1/\frac{d}{dx}M_{j,\phi}(x)|e^{x/32}$ at each probability $\frac{1}{32k}x$ where $\frac{d}{dx}M_{j,\phi}(x)$ is positive or negative respectively (without scaling so as to make total probability mass be 1), and let c_1, c_2 respectively be the constants that $p^+_{j,\phi}, p^-_{j,\phi}$ respectively must be multiplied by to normalize them.

Fact 12 directly implies that for any $i \leq j$, the ith fingerprint expectations for (unnormalized) $p^+_{j,\phi}$ and $p^-_{j,\phi}$ are identical.

Consider the fingerprint expectations for $i > j = \log k$. We note that $\sum_{i=0}^{\infty} i \cdot poi(xk, i) = xk$, and thus the sum over all i of i times the ith fingerprint expectations is exactly xk times the probability mass of the unnormalized distribution we started with. We thus relate c_1 and c_2 in this way.

By construction, $p^+_{j,\phi}$ and $p^-_{j,\phi}$ consist of elements with probability at most $\frac{\log k}{8k}$. Thus, for $x \leq \frac{\log k}{8k}$, we bound $\sum_{i=1+\log k}^{\infty} i \cdot poi(xk, i)$. We note that $i \cdot poi(xk, i) = xk \cdot poi(xk, i-1)$, yielding, from Fact 13 a bound $xk \sum_{i=\log k}^{\infty} poi(\frac{1}{8}\log k, i) \leq \frac{7}{8}xk \cdot poi(\frac{1}{8}\log k, \log k)$. We compare this to $poi(\log k, \log k) \leq 1$ by noting that, in general, $\frac{poi(y/8,y)}{poi(y,y)} = \frac{e^{7y/8}}{8^y} \leq 3.3^{-y}$, yielding a bound of $\frac{7}{8}xk \cdot 3.3^{-\log k} = o(x)$. That is, for an element of the distribution with probability x, its total contribution to the expected fingerprint entries with index greater than $\log k$ is $o(x)$; summing over all x yields $o(1)$ for the sum of these fingerprint entries.

As noted above, the sum over *all* fingerprint entries equals k. Thus, this method of computing the total probability mass of (unnormalized) $p^+_{j,\phi}$ and $p^-_{j,\phi}$ has a relative contribution of only $o(\frac{1}{k})$ from the $i > \log k$ portion; since the fingerprint expectations are identical for $i \leq \log k$, we conclude that the normalizing constants c_1, c_2 are within a factor $1 \pm o(\frac{1}{k})$ of each other.

We conclude that since the unnormalized distributions had identical expected fingerprints for $i \leq \log k$, the normalized distributions have fingerprints differing by at most a factor of $1 \pm o(\frac{1}{k})$, as desired. Further, the above argument implies that for any (normalized) distribution consisting of elements of probabilities at most $\frac{1}{8}\log k$, the expected total fingerprint entries above $\log k$ is $o(1)$, yielding that the corresponding expectations for $p^+_{j,\phi}$ and $p^-_{j,\phi}$ match to within this bound. □

Our overall aim here is to mold $p^+_{j,\phi}$ and $p^-_{j,\phi}$ into distributions with the property that the distributions of their respective fingerprints are “close”, respectively, to two very similar multivariate Gaussian distributions. As fingerprints are integer-valued vectors, while Gaussian distributions are continuous, we instead consider Gaussian distributions *rounded to the nearest lattice point.* Discreteness is still an obstacle, however, and the central limit theorem we put forward thus yields better bounds as the variance of the distributions in each direction increases. With this motivation in mind, we introduce the next construction which will modify $p^+_{j,\phi}$ and $p^-_{j,\phi}$ very little in the relative earthmover metric, while making the distributions of their histograms suitably “fat”.

Definition 15. *Define the* fattening *operator F that, given a histogram p, constructs a new histogram p^F as follows:*

- *Provisionally set $p^F(x) = \left(1 - \frac{(\log k)-1}{2\log^2 k}\right) p(x)$ for each x;*
- *For each integer $i \in \{1, \ldots, \log k\}$, increment $p^F(\frac{i}{k}) \leftarrow p^F(\frac{i}{k}) + \frac{k}{\log^3 k}$*

We note that, given a distribution, fattening returns a distribution. Further, for the sake of distribution support size lower bounds, we note that no elements are added below probability $\frac{1}{k}$, so that $p^{F+}_{j,\phi}$ and $p^{F-}_{j,\phi}$ retain the bound of $\frac{\phi}{32k\log k}$ on the probability of each domain element. Finally, we note that the bounds of Lemma 14 may only improve under fattening, as identical modifications are made to each distribution.

Claim 16. *The relative earthmover distances between the fattened and original version of $p^+_{j,\phi}$ and $p^-_{j,\phi}$ respectively are both $O(\frac{|\log\phi|+\log\log k}{\log k})$.*

Proof. We note that all the probabilities of $p^+_{j,\phi}$ and $p^-_{j,\phi}$ are between $\frac{\phi}{32k\log k}$ and $\frac{\log k}{k}$, incurring a per-unit-mass relative earthmover cost of at most $O(|\log\phi| + \log\log k)$. Since Definition 15 introduces less than $\frac{1}{\log k}$ new probability mass, shrinking the original histogram to make room, we can thus "move earth" from the original distribution to the modified distribution at cost the product of these two terms, namely $O(\frac{|\log\phi|+\log\log k}{\log k})$. □

We next show that for any fattened distribution, the variance of the distribution of the fingerprint is large in any direction. Specifically, for any unit vector $v \in \mathbb{R}^{\log k}$, we find an integer i such that elements of probability $\frac{i}{k}$—such as those added in Definition 15—have high-variance fingerprints along the direction v. Instead of proving this result only for $p^{F+}_{j,\phi}$ and $p^{F-}_{j,\phi}$, we prove it more generally, so that we may more easily invoke our central limit theorem.

Lemma 17. *For any vector $v \in \mathbb{R}^{\log k}$ of length 1, there exists an integer $i \in \{1, \ldots, \log k\}$ such that, drawing $\ell \leftarrow Poi(i)$ conditioned on $\ell \leq \log k$, the variance of $v(\ell)$ is at least $\frac{1}{6\log^{9/2} k}$, where we take $v(0) = 0$.*

Proof. We note the crucial stipulation that $v(0) = 0$, for otherwise, a uniform vector would have zero variance.

Given a unit vector v, there exists $i \in \{1, \ldots, \log k\}$ such that $|v(i) - v(i-1)| \geq \frac{1}{\log^2 k}$, since otherwise (since $v(0) = 0$) we would have $|v(i)| \leq \frac{i}{\log^2 k}$, implying $\sum_{i=1}^{\log k} v(i)^2 < 1$. Consider such an i.

Since in general, $i! \leq \frac{i^i}{e^i} 3\sqrt{i}$, we have that $poi(i, i-1) = poi(i,i) = \frac{i^i e^{-i}}{i!} \geq \frac{1}{3\sqrt{i}}$, which implies that, just the two possibilities $Poi(i) = i$ or $Poi(i) = i-1$ alone are enough to induce variance in $v(\ell)$ of the product of our bound on their total probability mass, $\frac{2}{3\sqrt{i}} \geq \frac{2}{3\log^{1/2} k}$ and the square of half $|v(i) - v(i-1)| \geq \frac{1}{\log^2 k}$, yielding $\frac{1}{6\log^{9/2} k}$. □

As a final ingredient before we may assemble the pieces of our main result, we show how to compare the *variances* of the respective distributions of fingerprints of the distributions $p^{F+}_{\log k,\phi}$ and $p^{F-}_{\log k,\phi}$. Lemma 14 has already shown that the fingerprint *expectations* are very close. One might suspect that analyzing the variances would require entirely different bounds, but as it turns out, "close fingerprint expectations imply close fingerprint variances".

To analyze this claim, we note that, for a histogram h, the ith fingerprint expectation is $\sum_{x:h(x)\neq 0} h(x)\cdot poi(xk, i)$. Since, for random variables X, Y, their covariance equals $E[XY] - E[X]E[Y]$, and covariance sums for independent distributions, we have that the covariance of the ith and jth fingerprint entries, for $i \neq j$, equals $-\sum_{x:h(x)\neq 0} h(x) poi(xk, i) poi(xk, j)$. We simplify this product,

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

to reveal a scaled version of a "squashed" version of the usual Poisson—that is, with $2xk$ instead of xk as the argument. The variance of the ith fingerprint entry may similarly be computed as $\sum_{x:h(x)\neq 0} h(x)\cdot \left(poi(xk,i) - poi(xk,i)^2\right)$, where, similarly, $poi(xk,i)^2 = 2^{-2i}\binom{2i}{i} poi(2xk, 2i)$.

The point of the next result, proved in Appendix C, is that one may express "squashed" Poisson functions $poi(2xk, i)$ as linear combinations of Poisson functions $poi(xk, j)$; thus, since the expectations relative to (regular) Poisson functions $poi(xk, j)$ match for $p^{F+}_{\log k,\phi}$ and $p^{F-}_{\log k,\phi}$, the same will hold true (though with greater error) for the expectations relative to the "squashed" Poisson functions $poi(2xk, i)$, and hence the variances and covariances will also approximately match.

Lemma 18. *For any $\epsilon > 0$ and integer $i \geq 0$, one may approximate $poi(2x, i)$ as a linear combination $\sum_{j=0}^{\infty} \alpha(j) poi(x, j)$ such that*

1. *For all $x \geq 0$, $|poi(2x, i) - \sum_{j=0}^{\infty} \alpha(j) poi(x, j)| \leq \epsilon$; and*
2. *$\sum_{j=0}^{\infty} |\alpha(j)| \leq \frac{1}{\epsilon} \cdot 200 \max\{\sqrt[4]{i}, 24 \log^{3/2} \frac{1}{\epsilon}\}$.*

We are thus equipped to bound the statistical distance between the distributions of fingerprints, which implies the indistinguishability of $p^{F+}_{j,\phi}$ and $p^{F-}_{j,\phi}$ to k-sample property testers.

Proposition 19. *For a positive constant $\phi < 1/4$, the statistical distance between the distribution of $Poi(k)$-sample fingerprints from $p^{F+}_{\log k,\phi}$ and $p^{F-}_{\log k,\phi}$ goes to 0 as k goes to infinity.*

Proof. We note that since both $p^{+}_{\log k,\phi}$ and $p^{-}_{\log k,\phi}$ consist of elements with probabilities at most $\frac{1}{8} \log k$, tail bounds (see the proof of Lemma 14 for the calculations) show that the probability that *any* such element occurs more than $\log k$ times is $o(1)$. We thus assume for the rest of this proof that this has not occurred.

Consider, for either fattened distribution, $p^{F+}_{\log k,\phi}$ or $p^{F-}_{\log k,\phi}$, the portion of the fingerprint above $\log k$, which we denote $\mathcal{F}_{>\log k}$. Since by assumption, only the "fattened" portion of either distribution contributes to $\mathcal{F}_{>\log k}$, and since these portions are identical, we have that the probability of a given $\mathcal{F}_{>\log k}$ occurring from $p^{F+}_{\log k,\phi}$ equals its probability of occurring from $p^{F-}_{\log k,\phi}$. We complete the proof by comparing, for each $\mathcal{F}_{>\log k}$, the conditional distributions of the fingerprints at or below $\log k$ *conditioned on the value $\mathcal{F}_{>\log k}$ and which elements of the distribution contributed to $\mathcal{F}_{>\log k}$*.

Note that the fattening process introduces $\frac{k}{\log^3 k}$ elements to the distribution at probability $\frac{i}{k}$ for each $i \in \{1, \ldots, \log k\}$. Since the number of occurrences of one of these elements is distributed as $Poi(i)$, for $i \leq \log k$, in expectation no more than half of these elements will be sampled more than $\log k$ times. Since the number of times each element is sampled is independent (as we are taking a Poisson-distributed number of samples), Chernoff bounds imply that the number of elements sampled more than $\log k$ times will be at most $\frac{3}{4} \frac{k}{\log^3 k}$ with probability $1 - e^{k^{\Theta(1)}}$, for each i. By a union bound over $i \leq \log k$, with probability at least $1 - o(1)$, conditioning on which elements contribute to $\mathcal{F}_{>\log k}$ will leave at least $\frac{1}{4} \frac{k}{\log^3 k}$ elements at each probability $\frac{i}{k}$ that are not fixed in the conditional distributions.

By Lemma 17, for each unit vector $v \in \mathbb{R}^{\log k}$, there is an index i such that each element of probability $\frac{i}{k}$ contributes $\frac{1}{6 \log^{9/2} k}$ to the (conditional) fingerprint variance in the direction of v. As the previous paragraph showed that there are at least $\frac{1}{4} \frac{k}{\log^3 k}$ elements with this property that are disjoint from the elements comprising $\mathcal{F}_{>\log k}$. Thus the fingerprint variance is at least $\sigma^2 := \frac{k}{24 \log^{15/2} k}$, in any direction v.

We thus apply our central limit theorem, Theorem 4, to the distributions of fingerprints of each of $p^{F+}_{\log k,\phi}$ and $p^{F-}_{\log k,\phi}$, conditioned on $\mathcal{F}_{>\log k}$. We note that each such distribution is a generalized multinomial distribution (see Definition 8) with $\log k$ columns and at most $n = \frac{32}{\phi} k \log k$ rows. We invoke the central limit theorem, to conclude that each such distribution may be approximated by the multivariate Gaussian distribution of the same mean and covariance, rounded to the nearest lattice points, to within statistical distance $\frac{\log^{4/3} k}{\sigma^{1/3}} \cdot 2.2 \cdot (3.1 + 0.83 \log n)^{2/3}$, which is $o(1)$ since the k in the numerator of $\sigma^2 = \frac{k}{24 \log^{15/2} k}$ dominates the logarithmic terms.

For a given $\mathcal{F}_{>\log k}$, let μ^+, μ^- denote respectively the vectors of conditional fingerprint expectations, for $p^{F+}_{\log k,\phi}$ and $p^{F-}_{\log k,\phi}$ respectively; let Σ^+, Σ^- denote respectively the corresponding covariance matrices. As we have just shown that the conditional distributions of fingerprints are statistically close to the multivariate Gaussian distributions $\mathcal{N}(\mu^+, \Sigma^+)$, $\mathcal{N}(\mu^-, \Sigma^-)$, respectively, each rounded to the nearest lattice point, it remains to compare the statistical distance of these distributions. We note immediately that rounding to the nearest lattice point can only decrease the statistical distance. We thus must bound $D_{tv}(\mathcal{N}(\mu^+, \Sigma^+), \mathcal{N}(\mu^-, \Sigma^-))$, which we will do with Proposition 30 once we have analyzed the disparities between the means and covariances.

Lemma 14 showed that the fingerprint expectations of $p^+_{\log k,\phi}$ and $p^-_{\log k,\phi}$ match to within $o(1)$. Fattening can only improve this, and since the conditioning applies only to the identical fattened region, it remains true that $|\mu^+(i) - \mu^-(i)| = o(1)$ for each i.

As we noted in the discussion preceding this result, approximating Poisson functions $poi(2xk, i)$ as linear combinations of Poisson functions $poi(xk, j)$ means that we can approximate each entry of the covariance matrix Σ by a linear combination of entries of the expectation vector μ. We thus invoke Lemma 18 for $\epsilon = \frac{1}{\sqrt{k}}$ to see that, indeed, there exist constants $\alpha_i(j)$ with $\sum_{j=0}^{\infty} |\alpha_i(j)| \leq \sqrt{k} \cdot 200 \max\{\sqrt[4]{i}, 24 \log^{3/2} \sqrt{k}\} = O(\sqrt{k} \log^{3/2} k)$ such that we may approximate entries $\Sigma(\ell, m)$ via coefficients $\alpha_{\ell+m}(j)$, where the error contributed by each domain element is at most ϵ. As there are at most $n = \frac{32}{\phi} k \log k$ domain elements, this approximation error is at most $\frac{32}{\phi}\sqrt{k} \log k$. Thus by the triangle inequality, the discrepancy $|\Sigma^+(\ell, m) - \Sigma^-(\ell, m)|$ for each element of the covariance matrix is bounded by twice this, plus the discrepancy due to $|\alpha_i(j)|$ times the difference $|\mu^+(i) - \mu^-(i)|$. We combine the bounds we have just derived to yield

$$|\Sigma^+(\ell, m) - \Sigma^-(\ell, m)| = O(\frac{\sqrt{k}}{\phi} \log^{3/2} k).$$

The two Gaussians $\mathcal{N}(\mu^+, \Sigma^+)$ and $\mathcal{N}(\mu^-, \Sigma^-)$ thus have means within $o(1)$, covariance matrices within $O(\frac{\sqrt{k}}{\phi} \log^{3/2} k)$, and variances at least $\sigma^2 = \frac{k}{24 \log^{15/2} k}$ in each direction—which thus lower-bounds the magnitude of the smallest eigenvalues of Σ^+, Σ^- respectively. For any positive constant ϕ, as k gets large, Proposition 30 implies that $D_{tv}(\mathcal{N}(\mu^+, \Sigma^+), \mathcal{N}(\mu^-, \Sigma^-)) = o(1)$, as claimed. □

Theorem 1. *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.*

Specifically, for any constant $\epsilon > 0$, there exists a pair of distributions of support at most n and for which each domain element occurs with probability at least $1/n$, satisfying:

1. *$|H(p^+) - H(p^-)| \geq \epsilon$*
2. *$|S(p^+) - S(p^-)| \geq n\epsilon$, where $S(D) := |\{x : \Pr_D[x] > 0\}|$*
3. *No algorithm can distinguish a set of $O\left(\frac{n}{\epsilon|\log \epsilon| \log n}\right)$ samples from p^+ versus p^- with probability 2/3.*

Proof. Let k be such that $n = \frac{32}{\phi} k \log k$. Construct $p^+ = p^{F+}_{\log k,\phi}$ and $p^- = p^{F-}_{\log k,\phi}$ according to Definition 10 followed by Definition 15. Then Lemma 11 and Claim 16 imply that p^+, p^- that are $O(\phi|\log \phi|)$-close in the relative earthmover metric, respectively, to the uniform distributions on n and $\frac{n}{2}$ elements. Proposition 19 shows that the distribution of inputs seen by a property tester in the two cases are statistically close to within $o(1)$. Thus no tester can distinguish them with probability 2/3.

For the second part of the theorem we let ϕ be such that $\phi|\log \phi| = \epsilon$. Claims 1 and 2 follow from the relative earthmover continuity of the entropy and support size functions, respectively. □

References

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

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

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

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

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

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

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

[8] 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.*

[9] 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.

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

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

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

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

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

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

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

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

[18] P. W. Glynn. Upper bounds on Poisson tail probabilities. *Operations Research Letters*, 6(1):914, 1987.

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

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

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

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

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

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

[25] L. Kantorovich and G. Rubinstein. On a functional space and certain extremal problems. *Vestnik Leningrad Univ. Math. 13:52-59*, 1958.

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

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

[28] 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.

[29] 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.

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

[31] 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.

[32] G. Szegö. Orthogonal polynomials, 4th edition. *American Mathematical Society, Colloquium Publications, 23. Providence, RI*, 1975.

[33] G. Valiant and P. Valiant. Estimating the unseen: a sublinear-sample canonical estimator of distributions. *Available at: http://www.cs.berkeley.edu/~gvaliant/papers/unseenVV.pdf*, 2010.

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

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

A A Multivariate Central Limit Theorem for Earthmover Distance

While the central limit theorem is foundational to modern statistics, 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. So perhaps it is not surprising that the particularly natural version we prove in this section seems absent from the statistics literature.

The main result of this section is a general central limit theorem for sums of independent random variables in high dimension. As with the Berry-Esseen bound, and the classic multivariate central limit theorem of Götze[19], our bound is in terms of what may be considered the third moments of the distribution, under a suitable change of basis. We note that our bounds have an extra logarithmic term and suspect this could be removed with a tighter analysis. The results of this section apply for both discrete and continuous distributions; we leverage these results in the next section for the discrete case.

The Berry-Esseen theorem bounds convergence to the Gaussian in terms of the maximum discrepancy between their respective cumulative distribution functions. Multiplying by two, this metric may be seen as a stand-in for the following: the maximum, over all intervals in $\mathbb{R}$, of the discrepancy between the probabilities of that interval under the two distributions. Götze's result can be thought of as generalizing this notion in the natural way to higher dimensions: convergence is shown relative to the discrepancy between the probabilities of any *convex set* ([19], and see [9] for discussion). 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.

For our purposes, and, we suspect, many others, convergence with respect to a more versatile distance metric is desired. The bound in our central limit theorem is in terms of (Euclidean) earthmover distance. We leverage this result to show, in Appendix B, a central limit theorem for generalized multinomial distributions in terms of *statistical distance*, the metric of choice for obtaining results in property testing.

Given a distribution S_n that is the sum of samples from n independent distributions $X_1, \ldots, X_n$ in $\mathbb{R}^k$, we aim to bound the earthmover distance of S_n from the Gaussian G of corresponding mean and covariance. We aim to bound the earthmover distance (also known as the *Wasserstein distance*) between S_n and G, which we will denote as $d_W(S_n, G)$. Intuitively, this distance $d_W(A, B)$ is defined as "the minimum, over all schemes of moving the probability mass of A to make B, of the cost of moving this mass, where the per-unit cost of moving mass from point x to point y is simply the (Euclidian) distance between x and y." It is often easier to define and work with the *dual* formulation of earthmover distance (this is the Kantorovich-Rubinstein theorem, [25], but may be intuitively seen as exactly what one would expect from linear programming duality):

Definition 20. *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*

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

It will be convenient for us to assume that our test functions, h, in addition to being Lipschitz continuous, are also differentiable. We note that even restricting the test functions to be *smooth* does not affect the above definition, as, for any Lipschitz-continuous function h, letting h_ϵ be the convolution of h with a Gaussian of radius ϵ for any $\epsilon > 0$, we note that h_ϵ is smooth, and $|h(x) - h_\epsilon(x)| \leq \epsilon\sqrt{k}$; thus for any random variables A, $\lim_{\epsilon \to 0} E[h_\epsilon(A)] = E[h(A)]$, and the earthmover distance definition remains unaltered.

The main central limit theorem of this section is:

Theorem 2. *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 $\beta k(2.7 + 0.83 \log n)$.*

We prove this as a consequence of the following theorem, which is somewhat tighter though more unwieldy. As it turns out, if the variance of $\sum_{i=1}^n Z_i$ is much larger in a certain direction than in others, then the earthmover bound is more forgiving of samples from Z_i that are large in that direction.

We prove this theorem using an adaptation of the celebrated Stein's method (see [4] for an introduction) as implemented for the multivariate case in [19]. (See also [9].)

Theorem 3. *Given n independent distributions $\{Z_i\}$ in $\mathbb{R}^k$, each having mean 0, and having total covariance equal to $k \times k$ matrix Σ, let T be the Cholesky factorization of Σ—that is, a $k \times k$ matrix such that $TT^\intercal = \Sigma$, making $T^{-1}\sum_{i=1}^n Z_i$ have covariance equal to the $k \times k$ identity matrix. Then the earthmover distance between $\sum_{i=1}^n Z_i$ and the normal distribution of mean 0 and covariance Σ is at most*

$$\sum_{i=1}^n 1.16 E\left[||Z_i|| \cdot ||T^{-1}Z_i||\right] \cdot E\left[||T^{-1}Z_i|| \log\left(1 + \frac{2.76}{||T^{-1}Z_i||}\right)\right] \\ + 0.49 E\left[||Z_i|| \cdot ||T^{-1}Z_i||^2 \cdot \log\left(1 + \frac{9.41}{||T^{-1}Z_i||}\right)\right]. \tag{1}$$

Proof of Theorem 2. We prove this from Theorem 3. In Equation 1 we note that both the first and second term have exactly one factor of $||Z_i||$, which we may upper-bound by β. Further, since the function $f(x) = x \log(1 + \frac{1}{x})$ is increasing for positive x, the rearrangement inequality implies that the first term is bounded by the corresponding expression with all parts put inside a single expectation. Thus Equation 1 is bounded by

$$\beta \sum_{i=1}^n E\left[||T^{-1}Z_i||^2 \left(1.16 \log\left(1 + \frac{2.76}{||T^{-1}Z_i||}\right) + 0.49 \log\left(1 + \frac{9.41}{||T^{-1}Z_i||}\right)\right)\right] \tag{2}$$

Define a new distribution Y such that for every vector x, $Pr[Y = x] = \frac{1}{c}||x|| \sum_{i=1}^n Pr[T^{-1}Z_i = x]$, where $c = \sum_{i=1}^n E[||T^{-1}Z_i||]$ is chosen so that Y is a valid distribution (that is, having total probability mass 1). (If the Z_i are continuous random variables, we define the distribution Y correspondingly.) We note that, letting $g(x) = x \cdot (1.16 \log(1 + \frac{2.76}{x}) + 0.49 \log(1 + \frac{9.41}{x}))$, we have that Equation 2 equals $\beta c \cdot E[g(||Y||)]$. The concavity of f implies the concavity of g, which implies by Jensen's inequality that $E[g(||Y||)] \leq g(E[||Y||])$. We have that $E[||Y||] = \frac{1}{c}\sum_{i=1}^n E[||T^{-1}Z_i||^2] = E\left[||T^{-1}\sum_{i=1}^n Z_i||\right] = \frac{k}{c}$, since covariance adds for independent distributions, and T is the matrix that transforms $\sum_{i=1}^n Z_i$ to have covariance the identity matrix.

Thus the earthmover distance is bounded by $\beta k(1.16 \log(1 + \frac{2.76c}{k}) + 0.49 \log(1 + \frac{9.41c}{k}))$. As this is an increasing function of c, it remains to bound c. We can crudely bound c by defining the distribution W that uniformly picks $i \in \{1, \ldots, n\}$ and then draws a sample from $T^{-1}Z_i$; we note that $c = n \cdot E[||W||]$. We bound c by observing that $E[||W||^2] = \frac{k}{n}$, from which, by the convexity of the squaring function and Jensen's inequality, we have that $c = nE[||W||] \leq n\sqrt{E[||W||^2]} = \sqrt{nk} \leq k\sqrt{n}$. Thus the earthmover distance is bounded by $\beta k(1.16 \log(1 + 2.76\sqrt{n}) + 0.49 \log(1 + 9.41\sqrt{n}))$, which, for $n \geq 1$ is easily seen to be less than the desired bound of $\beta k(2.7 + 0.83 \log n)$. □

A.1 A CLT via Stein's Method

We now prove Theorem 3 via Stein's method.

Proof of Theorem 3. We let $X_i = T^{-1}Z_i$ and work with X_i instead of Z_i throughout. While the earthmover distance in the original basis is defined via the supremum over differentiable "test functions" in $\mathrm{Lip}(\mathbb{R}^k, 1)$, when we work with X_i, the test functions instead range over $T \circ \mathrm{Lip}(\mathbb{R}^k, 1)$, that is, for $\ell \in \mathrm{Lip}(\mathbb{R}^k, 1)$, we take $h(x) = \ell(Tx)$.

The heart of Stein's method consists of constructing a simple transformation $h \to f_h$ that takes test functions $h \in T \circ \mathrm{Lip}(\mathbb{R}^k, 1)$ and transforms them to appropriate functions f_h such that for any distribution S_n, we have

$$E[h(S_n)] - E[h(\Phi)] = E[S_n \cdot \nabla f_h(S_n) - \triangle f_h(S_n)], \tag{3}$$

where $\triangle f_h$ represents the Laplacian of f_h and ∇f_h the gradient of f_h. When one takes Taylor expansions of each of the two terms on the right hand side, one can arrange to have a pair of terms that have second-order dependence on S_n cancel, leaving only third-order terms remaining, which is what will yield the third-order dependence in the theorem. We cite [9] for the result that Equation 3 is satisfied when, letting $\phi_r(x) \triangleq (2\pi r^2)^{-k/2} e^{-\frac{||x||}{2r^2}}$ be the k-dimensional Gaussian of mean 0 and radius r, we define

$$f_h(x) \triangleq \int_0^\infty (h * \phi_{\sqrt{1-e^{-2s}}})(e^{-s}x) - E[h(\Phi)]\, ds, \tag{4}$$

where we consider $h * \phi_0 = h$.

We take $S_n = \sum_{i=1}^n X_i$, and let S_{-i} denote $S_n - X_i$, that is, the sum of samples from all but one of the distributions; by definition S_{-i} is independent of X_i. We use the first-order expansion $f(x+y) = f(x) + \int_0^1 y \cdot \nabla f(x+ty)\, dt$, where $y \cdot \nabla f(x+ty)$ is simply the directional derivative of f in the direction y evaluated at $x + ty$. In coordinates, this is

$$f(x+y) = f(x) + \int_0^1 \sum_{a=1}^k y(a) D_a f(x+ty)\, dt,$$

where we use D_a to denote the partial derivative in the ath coordinate. Similarly, the second-order expansion is

$$f(x+y) = f(x) + y \cdot \nabla f(x) + \int_0^1 (1-t) \sum_{a,b=1}^k y(a)y(b) D_{ab} f(x+ty)\, dt,$$

where as above, $\sum_{a,b=1}^k y(a)y(b)D_{ab}f(x+ty)$ is just the "directional second derivative" of f, in the direction y, evaluated at $x + ty$. Thus we may expand $S_n \cdot \nabla f(S_n) = \sum_{i=1}^n X_i \cdot \nabla f(S_{-i} + X_i) = \sum_{i=1}^n \sum_{a=1}^k X_i(a) D_a f(S_{-i} + X_i)$ to second order as

$$\sum_{i=1}^n \sum_{a=1}^k X_i(a) \left(D_a f(S_{-i}) + \left(\sum_{b=1}^k X_i(b) D_{ab} f(S_{-i}) \right) + \left(\int_0^1 (1-t) \sum_{b,c=1}^k X_i(b) X_i(c) D_{abc} f(S_{-i} + t \cdot X_i)\, dt \right) \right). \tag{5}$$

We note that since X_i has mean 0 and is independent of S_{-i}, the first term has expectation 0. We now aim to cancel the expectation of the second term against an expansion of $\triangle f(S_n)$. Note that the expected value of the factor $X_i(a)X_i(b)$ in the second term is just the (a, b)th component of the covariance matrix of X_i, which we write as $\mathrm{Cov}(X_i)(a, b)$. Since by assumption, the sum over i of the covariance matrices $\mathrm{Cov}(X_i)$ equals the identity matrix, we may rewrite $\triangle f(S_n) = \sum_{(a=b)=1}^{k} D_{ab}f(S_n) = \sum_{i=1}^{n}\sum_{a,b=1}^{k}\mathrm{Cov}(X_i)(a,b)D_{ab}f(S_n)$. Expanding the ith term of this to first order centered at S_{-i}, for each i, yields

$$\sum_{i=1}^{n}\sum_{a,b=1}^{k}\mathrm{Cov}(X_i)(a,b)\left(D_{ab}f(S_{-i}) + \int_0^1 \sum_{c=1}^{k} X_i(c)D_{abc}f(S_{-i}+t\cdot X_i)\,dt\right), \tag{6}$$

where the expectation of the first term above is seen to be exactly the expectation of the second term of Equation 5, and thus the difference between the expectations of Equations 5 and 6, which for $f = f_h$ equals $E[h(S_n)] - E[h(\Phi)]$ by construction, will consist only of the last, third-order terms from each expression.

Let ζ_i denote the expectation of the last term of Equation 5 for the corresponding i, and η_i denote the expectation of the last term of Equation 6 for the corresponding i. By the above, $d_W(S_n,\Phi)$ is thus bounded by the supremum over $h \in T \circ \mathrm{Lip}(\mathbb{R}^k, 1)$ of $\sum_{i=1}^{n}|\zeta_i| + |\eta_i|$. We thus turn to bounding ζ_i, η_i. We assume throughout that $X_i \neq 0$, as, when $X_i = 0$ the corresponding terms of Equations 5 and 6 are trivially seen to be 0.

Defining $g_s(x) = h(e^{-s}x)$, we note that we may reexpress the first term in the definition of f_h as $(h * \phi_{\sqrt{1-e^{-2s}}})(e^{-s}x) = (g_s * \phi_{\sqrt{e^{2s}-1}})(x)$. Letting $\widetilde{X}_i$ denote an independent sample from the distribution X_i, we note that we may replace $\mathrm{Cov}(X_i)(a,b)$ in Equation 6 by $E[\widetilde{X}_i(a)\widetilde{X}_i(b)]$, thus yielding that η_i equals the expectation of

$$\int_0^\infty\int_0^1\sum_{a,b,c=1}^{k}\widetilde{X}_i(a)\widetilde{X}_i(b)X_i(c)D_{abc}(g_s * \phi_{\sqrt{e^{2s}-1}})(S_i + t\cdot X_i)\,dt\,ds,$$

where we note that the final term $E[h(\Phi)]$ of Equation 4 is constant, and hence its third derivative does not contribute to η_i, and is thus omitted in the above equation.

We note that the expression $\sum_{a,b,c=1}^{k}\widetilde{X}_i(a)\widetilde{X}_i(b)X_i(c)D_{abc}$ is just a third directional derivative, with two differentiations in the direction of the vector $\widetilde{X}_i$ and one in the direction X_i, which we may denote as $D_{\widetilde{X}_i}D_{\widetilde{X}_i}D_{X_i}$. Since convolution commutes with differentiation, η_i thus equals the expectation of

$$\begin{aligned}&\int_0^\infty\int_0^1 (D_{\widetilde{X}_i}g_s * D_{\widetilde{X}_i}D_{X_i}\phi_{\sqrt{e^{2s}-1}})(S_i + t\cdot X_i)\,dt\,ds\\ &\qquad = \int_0^\infty\int_0^1\int_{\mathbb{R}^k} D_{\widetilde{X}_i}g_s(x)D_{\widetilde{X}_i}D_{X_i}\phi_{\sqrt{e^{2s}-1}}(S_i + t\cdot X_i - x)\,dx\,dt\,ds\\ &\qquad = \int_0^\infty\int_{\mathbb{R}^k} D_{\widetilde{X}_i}g_s(x)\int_0^1 D_{\widetilde{X}_i}D_{X_i}\phi_{\sqrt{e^{2s}-1}}(S_i + t\cdot X_i - x)\,dt\,dx\,ds\end{aligned}$$

Because h, by definition, is the composition of matrix T with a differentiable function of Lipschitz constant 1, g_s is the composition of T with a function of Lipschitz constant e^{-s} and thus we can bound the absolute value of this last expression by

$$\int_0^\infty ||T\widetilde{X}_i||e^{-s}\int_{\mathbb{R}^k}\left|\int_0^1 D_{\widetilde{X}_i}D_{X_i}\phi_{\sqrt{e^{2s}-1}}(t\cdot X_i + x)\,dt\right|\,dx\,ds, \tag{7}$$

where we have made the substitution $S_i - x \to x$. We bound the integral over $\mathbb{R}^k$ in two ways. First, since a univariate Gaussian of variance r^2 is unimodal, the integral of the absolute value of its derivative is simply twice its maximum, namely $2\cdot\frac{1}{\sqrt{2\pi r^2}}$. Since ϕ_r can be expressed as the product of k

univariate Gaussians along orthogonal basis directions, each of variance r^2, and having integral 1, we have that $\int_{\mathbb{R}^k} |D_{\widetilde{X}_i}\phi_{\sqrt{e^{2s}-1}}|\,dx = \frac{2||\widetilde{X}_i||}{\sqrt{2\pi(e^{2s}-1)}}$, just the corresponding univariate expression in the basis direction $\frac{\widetilde{X}_i}{||\widetilde{X}_i||}$. Since integration is the inverse of differentiation, we have that $\int_0^1 D_{\widetilde{X}_i}D_{X_i}\phi_{\sqrt{e^{2s}-1}}(t\cdot X_i+x)\,dt = D_{\widetilde{X}_i}\phi_{\sqrt{e^{2s}-1}}(X_i+x) - D_{\widetilde{X}_i}\phi_{\sqrt{e^{2s}-1}}(x)$, and by the triangle inequality we may thus bound the $\mathbb{R}^k$ integral of Equation 7 as twice what we just computed: $\frac{4||\widetilde{X}_i||}{\sqrt{2\pi(e^{2s}-1)}}$.

For large s, however, this bound is not effective, and in this case we instead take

$$\begin{aligned}\int_{\mathbb{R}^k}\left|\int_0^1 D_{\widetilde{X}_i}D_{X_i}\phi_{\sqrt{e^{2s}-1}}(t\cdot X_i+x)\,dt\right|\,dx &\le \int_{\mathbb{R}^k}\int_0^1\left|D_{\widetilde{X}_i}D_{X_i}\phi_{\sqrt{e^{2s}-1}}(t\cdot X_i+x)\right|\,dt\,dx\\ &= \int_{\mathbb{R}^k}\left|D_{\widetilde{X}_i}D_{X_i}\phi_{\sqrt{e^{2s}-1}}(x)\right|\,dx\end{aligned}$$

Letting $y_i = \frac{X_i}{||X_i||}$ denote the unit vector in the X_i direction, and z_i denote an orthogonal unit vector such that, for real numbers u, v we have $\widetilde{X}_i = u\cdot y_i + v\cdot z_i$, we thus have $D_{\widetilde{X}_i}D_{X_i} = ||X_i||(u\cdot D^2_{y_i} + v\cdot D_{z_i}D_{y_i})$, and by the triangle inequality we may bound

$$\int_{\mathbb{R}^k}\left|D_{\widetilde{X}_i}D_{X_i}\phi_{\sqrt{e^{2s}-1}}(x)\right|\,dx \le ||X_i||\int_{\mathbb{R}^k}\left|u\cdot D^2_{y_i}\phi_{\sqrt{e^{2s}-1}}(x)\right| + \left|v\cdot D_{y_i}D_{z_i}\phi_{\sqrt{e^{2s}-1}}(x)\right|\,dx, \qquad (8)$$

where we may now leverage the orthogonality of y_i and z_i.

As above, we note that since the Gaussian can be expressed as the product of one-dimensional Gaussians along any orthogonal basis, and since y_i and z_i are orthogonal unit vectors, we have that $\int_{\mathbb{R}^k}|D_{y_i}D_{z_i}\phi_{\sqrt{e^{2s}-1}}(x)|\,dx = \left(\frac{2}{\sqrt{2\pi(e^{2s}-1)}}\right)^2 = \frac{2}{\pi(e^{2s}-1)}$, just the square of the univariate case we computed above. Similarly, $\int_{\mathbb{R}^k}|D^2_{y_i}\phi_{\sqrt{e^{2s}-1}}(x)|\,dx$ equals the corresponding expression for a univariate Gaussian, the integral of the absolute value of its second derivative, which by definition is the total variation of its first derivative. As the derivative of a univariate Gaussian of variance r^2 takes maximum and minimum values at $\pm r$, at which locations it has values respectively $\mp\frac{e^{-1/2}}{r^2\sqrt{2\pi}}$, and has no other local optima, its total variation is just four times this, which, for $r^2 = e^{2s}-1$ gives us $\int_{\mathbb{R}^k}|D^2_{y_i}\phi_{\sqrt{e^{2s}-1}}(x)|\,ds = \frac{4e^{-1/2}}{(e^{2s}-1)\sqrt{2\pi}}$.

Thus, since $|u|^2+|v|^2 = ||\widetilde{X}_i||^2$, we bound Equation 8 as $\frac{||X_i||}{e^{2s}-1}$ times $|u|\frac{4e^{-1/2}}{\sqrt{2\pi}}+|v|\frac{2}{\pi}$. We bound this last expression by the Cauchy-Schwarz inequality as $||\widetilde{X}_i||\sqrt{\left(\frac{4e^{-1/2}}{\sqrt{2\pi}}\right)^2+\left(\frac{2}{\pi}\right)^2} = ||\widetilde{X}_i||\frac{2}{\pi}\sqrt{1+2\pi e^{-1}}$. Equation 8 is thus bounded by $||X_i||\cdot||\widetilde{X}_i||\frac{1}{e^{2s}-1}\frac{2}{\pi}\sqrt{1+2\pi e^{-1}}$. Combining this bound with the bound computed above yields

$$|\eta_i| \le E\left[||T\widetilde{X}_i||\cdot||\widetilde{X}_i||\int_0^\infty e^{-s}\min\left\{\frac{4}{\sqrt{2\pi(e^{2s}-1)}}, \frac{||X_i||}{e^{2s}-1}\frac{2}{\pi}\sqrt{1+2\pi e^{-1}}\right\}\,ds\right] \qquad (9)$$

Because the expression for ζ_i will be similar, we derive a general bound for $\int_0^\infty e^{-s}\min\{\frac{1}{\sqrt{e^{2s}-1}}, \frac{\alpha}{e^{2s}-1}\}ds$. Note that the first term is less than the second term when $\sqrt{e^{2s}-1} < \alpha$, namely, when $s < \log\sqrt{\alpha^2+1}$. Further, it is straightforward to check that $\int\frac{e^{-s}}{\sqrt{e^{2s}-1}}ds = e^{-s}\sqrt{e^{2s}-1}$, and $\int\frac{e^{-s}}{e^{2s}-1}ds =$

$e^{-s} - \log \frac{e^s+1}{\sqrt{e^{2s}-1}}$. Thus we evaluate

$$\int_0^\infty e^{-s} \min\{\frac{1}{\sqrt{e^{2s}-1}}, \frac{\alpha}{e^{2s}-1}\} ds = \int_0^{\log \sqrt{\alpha^2+1}} \frac{e^{-s}}{\sqrt{e^{2s}-1}} ds + \alpha \int_{\log \sqrt{\alpha^2+1}}^\infty \frac{e^{-s}}{e^{2s}-1} ds$$
$$= \frac{\alpha}{\sqrt{\alpha^2+1}} + \alpha \left[\log \frac{\sqrt{\alpha^2+1}+1}{\alpha} - \frac{1}{\sqrt{\alpha^2+1}} \right]$$
$$= \alpha \log \frac{\sqrt{\alpha^2+1}+1}{\alpha} \leq \alpha \log \left(1 + \frac{2}{\alpha}\right) \tag{10}$$

We may thus bound $|\eta_i|$ from Equations 9 and 10 by setting $\alpha = \frac{1}{\sqrt{2\pi}} ||X_i|| \sqrt{1+2\pi e^{-1}}$. Since $\frac{2}{\pi}\sqrt{1+2\pi e^{-1}} < 1.16$ and $2 \cdot \frac{4}{\sqrt{2\pi}} / (\frac{2}{\pi}\sqrt{1+2\pi e^{-1}}) < 2.76$, we have that

$$|\eta_i| < 1.16 E \left[||T\widetilde{X}_i|| \cdot ||\widetilde{X}_i|| ||X_i|| \log \left(1 + \frac{2.76}{||X_i||}\right) \right]$$
$$= 1.16 E \left[||TX_i|| \cdot ||X_i|| \right] E \left[||X_i|| \log \left(1 + \frac{2.76}{||X_i||}\right) \right] \tag{11}$$

We now turn to bounding the last term of Equation 5, whose expectation we have denoted as ζ_i. Similarly to above, we have

$$\sum_{a,b,c=1}^k \int_0^1 (1-t) X_i(a) X_i(b) X_i(c) D_{abc} f_h(S_{-i} + t \cdot X_i)\, dt$$
$$= \int_0^1 (1-t) D^3_{X_i} f_h(S_{-i} + t \cdot X_i)\, dt$$
$$= \int_0^\infty \int_0^1 (1-t) D^3_{X_i} (g_s * \phi_{\sqrt{e^{2s}-1}})(S_{-i} + t \cdot X_i)\, dt\, ds$$
$$= \int_0^\infty \int_0^1 (1-t) (D_{X_i} g_s * D^2_{X_i} \phi_{\sqrt{e^{2s}-1}})(S_{-i} + t \cdot X_i)\, dt\, ds$$
$$= \int_0^\infty \int_{\mathbb{R}^k} D_{X_i} g_s(x) \int_0^1 (1-t) D^2_{X_i} \phi_{\sqrt{e^{2s}-1}}(S_{-i} + t \cdot X_i - x)\, dt\, dx\, ds$$
$$\leq ||TX_i|| e^{-s} \int_0^\infty \int_{\mathbb{R}^k} \left| \int_0^1 (1-t) D^2_{X_i} \phi_{\sqrt{e^{2s}-1}}(t \cdot X_i + x)\, dt \right| dx\, ds$$

As above, if we take an orthonormal basis that includes a vector in the direction of X_i then we can decompose $D^2_{X_i}\phi_{\sqrt{e^{2s}-1}}$ into the product of the corresponding expression for a univariate Gaussian in the direction of X_i, and univariate Gaussians along all the other basis directions. Thus, if we let $\bar{\phi}_r$ denote the univariate version of ϕ_r, namely, $\bar{\phi}_r(x) = \frac{1}{r \cdot \sqrt{2\pi}} e^{-\frac{x^2}{2r^2}}$, then the above integral over $\mathbb{R}^k$ equals exactly

$$||X_i||^2 \int_{-\infty}^\infty \left| \int_0^1 (1-t) \bar{\phi}''_{\sqrt{e^{2s}-1}}(x + ||X_i|| t)\, dt \right| dx \tag{12}$$

As above, we bound this expression in two ways. First, we bound it by moving the absolute values inside the integral, swapping the order of integration, and then making the substitution $y = x + ||X_i|| t$ to yield

$$||X_i||^2 \int_0^1 \int_{-\infty}^\infty \left| (1-t) \bar{\phi}''_{\sqrt{e^{2s}-1}}(y) \right| dy\, dt$$

The integral may thus be expressed as the product of separate integrals over t and y: since $\int_0^1 1-t\,dt = \frac{1}{2}$, and as we computed above, $\int_{-\infty}^{\infty} |\bar{\phi}''_{\sqrt{e^{2s}-1}}(y)|\,dy = \frac{4e^{-1/2}}{(e^{2s}-1)\sqrt{2\pi}}$, we have that Equation 12 is at most $||X_i||^2 \frac{2e^{-1/2}}{(e^{2s}-1)\sqrt{2\pi}}$.

For the second bound, we first note that we may simplify slightly by replacing $(1-t)$ by t in Equation 12 (this is the change of variables $t \to (1-t)$, $x \to -x - ||X_i||$, relying on the fact that $\bar{\phi}''$ is symmetric about 0). It will be convenient to consider the inner integral as being over $\mathbb{R}$ instead of just $[0,1]$, and we thus introduce the notation $(t)_{[0,1]}$ to represent t if $t \in [0,1]$ and 0 otherwise. Thus we bound Equation 12 as

$$\begin{aligned}
&||X_i||^2 \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} (t)_{[0,1]} \bar{\phi}''_{\sqrt{e^{2s}-1}}(x + ||X_i||t)\,dt \right| dx \\
&\quad = ||X_i||^2 \int_{-\infty}^{\infty} \left| \int_{-\infty}^{\infty} \left((t)_{[0,1]} - \left(-\frac{x}{||X_i||} \right)_{[0,1]} \right) \bar{\phi}''_{\sqrt{e^{2s}-1}}(x + ||X_i||t)\,dt \right| dx \\
&\quad \leq ||X_i||^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left| \left((t)_{[0,1]} - \left(-\frac{x}{||X_i||} \right)_{[0,1]} \right) \bar{\phi}''_{\sqrt{e^{2s}-1}}(x + ||X_i||t) \right| dx\,dt \\
&\quad = ||X_i||^2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left| \left((t)_{[0,1]} - \left(t - \frac{y}{||X_i||} \right)_{[0,1]} \right) \bar{\phi}''_{\sqrt{e^{2s}-1}}(y) \right| dy\,dt \\
&\quad = ||X_i||^2 \int_{-\infty}^{\infty} \left| \bar{\phi}''_{\sqrt{e^{2s}-1}}(y) \right| \int_{-\infty}^{\infty} \left| (t)_{[0,1]} - \left(t - \frac{y}{||X_i||} \right)_{[0,1]} \right| dt\,dy
\end{aligned}$$

where the first equality holds since ϕ'' has integral 0, and hence we can add any multiple of it (independent of t) to the inner integral; the second equality is just the substitution $x \to y - ||X_i||t$.

To bound this integral, we note the general fact that, if a function f has total variation a, then $\int_{-\infty}^{\infty} |f(x) - f(x-b)|\,dx \leq a|b|$. Thus since the function $(t)_{[0,1]}$ has total variation 2, the inner integral is bounded by $2\frac{y}{||X_i||}$. Since $\bar{\phi}''_r$ crosses 0 at $\pm r$, and integration by parts yields $\int y\bar{\phi}''_r(y)\,dy = y\bar{\phi}'_r(y) - \int \bar{\phi}'_r(y)\,dy = -\bar{\phi}_r(y)(1 + \frac{y^2}{r^2})$ and hence $\int_{-\infty}^{\infty} |y\bar{\phi}''_r(y)|\,dy = -2\int_0^r y\bar{\phi}''_r(y)\,dy + 2\int_r^{\infty} y\bar{\phi}''_r(y) = -2\bar{\phi}_r(0) + 8\bar{\phi}_r(r) = \frac{8e^{-1/2}-2}{r\cdot\sqrt{2\pi}}$ we may thus bound Equation 12 by $||X_i||\frac{16e^{-1/2}-4}{\sqrt{2\pi(e^{2s}-1)}}$.

Thus, similarly to above, we have

$$|\zeta_i| \leq ||TX_i|| \cdot ||X_i|| \int_0^{\infty} e^{-s} \min\left\{ \frac{16e^{-1/2}-4}{\sqrt{2\pi(e^{2s}-1)}}, \frac{||X_i|| \cdot 2e^{-1/2}}{(e^{2s}-1)\sqrt{2\pi}} \right\} ds.$$

Since $\frac{2e^{-1/2}}{\sqrt{2\pi}} < 0.49$ and $2 \cdot \frac{16e^{-1/2}-4}{\sqrt{2\pi}} / \frac{2e^{-1/2}}{\sqrt{2\pi}} < 9.41$, we have from Equation 10 that $|\zeta_i| < 0.49 \cdot E[||TX_i|| \cdot ||X_i||^2 \log(1 + \frac{9.41}{||X_i||})]$. Combining this and Equation 11 yields the theorem. □

B A Central Limit Theorem for Generalized Multinomial Distributions

In this section we leverage the central limit theorem of Theorem 2 to show our second central limit theorem that bounds the *statistical distance*, denoted by D_{tv} between generalized multinomial distributions and (discretized) Gaussian distributions. While Theorem 2 certainly applies to generalized multinomial distributions, the goal of this section is to derive a bound in terms of the rather more stringent statistical distance. The main hurdle is relating the "smooth" nature of the Gaussian distribution and earthmover distance metric to the "discrete" setting imposed by a statistical distance comparison with the discrete generalized multinomial distribution.

The analysis to compare a Gaussian to a generalized multinomial distribution proceeds in two steps. Given the earthmover distance bound provided by Theorem 2, we first smooth both sides via convolution with a suitably high-variance distribution to convert this bound into a statistical distance bound, albeit not between the original two distributions but between convolved versions of them. The second step is via a "deconvolution" lemma (Lemma 24) that relies on the unimodality in each coordinate of generalized multinomial distributions.

We begin by showing this unimodality via a result about homogeneous polynomials that generalizes the classic Newton inequalities.

Given a polynomial p in k variables, and a nonnegative integer vector $v \in \mathbb{Z}^k$, we denote by $p_{(v)}$ the coefficient of the term $x_1^{v(1)} x_2^{v(2)} \cdot \ldots \cdot x_k^{v(k)}$ in p.

Fact: Multivariate Newton Inequalities (Fact 1.10:2 of [21]). *Given a homogeneous polynomial p of degree n in k variables, with nonnegative coefficients, if it is the case that for any complex $x_1, \ldots, x_k$ with strictly positive real parts, $p(x_1, \ldots, x_k) \neq 0$, then for any nonnegative integer vector v and letting $\Delta = (1, -1, 0, \ldots, 0) \in \mathbb{Z}^k$, we have $p_{(v)}^2 \geq p_{(v+\Delta)} p_{(v-\Delta)}$.*

(We note that the actual result from [21], in analogy with Newton's inequalities, is tighter by a factor $\prod_i v(i)!^2 / \prod_i (v+\Delta)(i)!(v-\Delta)(i)! = \frac{v(1)v(2)}{(1+v(1))(1+v(2))}$, though for our purposes we need only the simpler bound.)

Definition 21. *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 22. *A function $f : \mathbb{Z} \to \mathbb{R}^+$ is log-concave if its support is an interval, and $\forall i \in \mathbb{Z}, f(i)^2 \geq f(i-1)f(i+1)$.*

The logarithm of a log-concave function is concave (interpreting $\log 0$ as $-\infty$); thus any log-concave function is unimodal (i.e., monotonically increasing to the left of some point, and monotonically decreasing to the right). We note that we consider "unimodal" in the non-strict sense, so that, for example, the constant function is unimodal.

Lemma 23. *Generalized multinomial distributions are log-concave—and in particular, unimodal—in any coordinate.*

Proof. Given a generalized multinomial distribution parameterized by ρ, where ρ has n rows and k columns, we define $\bar{\rho}$ to be the matrix whose columns are indexed 0 through k, and which consists of ρ extended so that for each $i \in \{1, \ldots n\}$, $\sum_{j=0}^k \bar{\rho}(i,j) = 1$.

Let p be the homogeneous polynomial of degree n in k variables defined as $p(x_1, \ldots, x_k) = \prod_{i=1}^n (\bar{\rho}(i,0)x_0 + \ldots + \bar{\rho}(i,k)x_k)$. We note that for any nonnegative integer vector v, the coefficient $p_{(v)}$ equals, by definition, the probability of drawing v from the multinomial distribution (ignoring the implicit "0th coordinate").

We invoke the multivariate Newton inequalities (with the coordinates renumbered as necessary) by noting that, first, p clearly has nonnegative coefficients, and second, if $x_0, \ldots, x_k$ are complex numbers with strictly positive real parts, then each term $(\bar{\rho}(i,0)x_0 + \ldots + \bar{\rho}(i,k)x_k)$ will have strictly positive real part, and hence be nonzero, which implies that $p(x_0, \ldots, x_k) \neq 0$. Thus the multivariate Newton inequalities imply that the multinomial distribution (with its "0th coordinate" ignored) is log-concave in its first coordinate; by symmetry, it is log-concave in every coordinate. □

Given this general structural result about the distributions at hand, we now construct the second ingredient of our proof, the "deconvolution" lemma. What this shows is that, given a convolution $f * g$ that closely approximates a third function h, we can leverage the unimodality of f under certain conditions to "deconvolve" by g and relate f and h directly. We will apply this univariate result in the proof of the central limit theorem by applying it inductively along lines in each of the k coordinate directions.

Lemma 24. *Given an integer $\ell > 0$, a unimodal function $f : \mathbb{Z} \to \mathbb{R}^+$, a function $g : \{-\ell, -\ell + 1 \ldots, \ell - 1, \ell\} \to \mathbb{R}^+$ with $\sum_i g(i) = 1$, and an arbitrary bounded function $h : \mathbb{Z} \to \mathbb{R}^+$ then, letting $f * g$ denote the convolution of f and g, we have*

$$\sum_{i=-\infty}^{\infty} |f(i) - h(i)| \leq 10\ell \left(\sup_i h(i)\right) + \sum_{i=-\infty}^{\infty} |(f * g)(i) - h(i)|.$$

Proof. Assume that we have scaled f and h so that $\sup_i h(i) = 1$. Let f^- denote the function that is the (pointwise) minimum of f and 1, and let f^+ denote $f - f^-$. We note that f^+ and f^- are unimodal. For the following inequality, we let $[[0, j]]$ denote the set of integers $\{0, \ldots, j-1\}$ when $j > 0$, the set $\{j, \ldots, -1\}$ when $j < 0$, and the empty set when $j = 0$: by the definition of convolution, two applications of the triangle inequality, and a rearrangement of terms we have

$$\begin{aligned}
\sum_{i=-\infty}^{\infty} |f^-(i) - (f^- * g)(i)| &= \sum_{i=-\infty}^{\infty} \left| \sum_{j=-\ell}^{\ell} g(j)(f^-(i) - f^-(i-j)) \right| \\
&\leq \sum_{i=-\infty}^{\infty} \sum_{j=-\ell}^{\ell} g(j) |f^-(i) - f^-(i-j)| \\
&\leq \sum_{i=-\infty}^{\infty} \sum_{j=-\ell}^{\ell} \sum_{k \in [[0,j]]} g(j) |f^-(i-k) - f^-(i-k+1)| \\
&= \left(\sum_{j=-\ell}^{\ell} g(j)|j| \right) \sum_{i=-\infty}^{\infty} |f^-(i) - f^-(i+1)| \\
&\leq \ell \sum_{i=-\infty}^{\infty} |f^-(i) - f^-(i+1)|.
\end{aligned}$$

Since f^- is unimodal and bounded between 0 and 1, $\sum_i |f^-(i) - f^-(i+1)| \leq 2$, and we thus bound the above inequality by 2ℓ.

We note that since f is unimodal, it exceeds 1 on a contiguous (possibly empty) interval, which we denote $[u, v]$. Since $f * g = f^- * g + f^+ * g$, we have the triangle inequality $|(f * g)(i) - h(i)| \leq |(f^+ * g)(i)| + |(f^- * g)(i) - h(i)|$. Since $f^- * g = 1$ on the interval $[u + \ell, v - \ell]$, and $f^+ * g$ is confined to the interval $[u - \ell, v + \ell]$, then we actually have equality everywhere *except* the intervals $[u - \ell, u + \ell - 1]$ and $[v - \ell + 1, v + \ell]$. On these intervals, we consider the reverse inequality (another triangle inequality) $|(f * g)(i) - h(i)| \geq |(f^+ * g)(i)| - |(f^- * g)(i) - h(i)|$ which, since $(f^- * g)(i) \in [0, 1]$, we bound as being at least $|(f^+ * g)(i)| + |(f^- * g)(i) - h(i)| - 2$ on these intervals. Thus

$$\begin{aligned}
\sum_{i=-\infty}^{\infty} |(f * g)(i) - h(i)| &\geq \sum_{i=-\infty}^{\infty} |(f^+ * g)(i)| + \sum_{i=-\infty}^{\infty} |(f^- * g)(i) - h(i)| + \sum_{i=u-\ell}^{u+\ell-1} (-2) + \sum_{i=v-\ell+1}^{v+\ell} (-2) \\
&= -8\ell + \sum_{i=-\infty}^{\infty} |f^+(i)| + \sum_{i=-\infty}^{\infty} |(f^- * g)(i) - h(i)| \\
&\geq -10\ell + \sum_{i=-\infty}^{\infty} |f^+(i)| + \sum_{i=-\infty}^{\infty} |f^-(i) - h(i)| \\
&= -10\ell + \sum_{i=-\infty}^{\infty} |f(i) - h(i)|,
\end{aligned}$$

where the last inequality is what we proved above, and the last equality is true term-by-term since the region where f^+ is nonzero is exactly the region where $f^-(i) = 1 \geq h(i)$, and thus we have the lemma. □

We are now equipped to assemble the components and prove the central limit theorem. Our central limit theorem related the generalized multinomial distribution to the "discretized" version of the Gaussian distribution of identical mean and covariance, as defined below.

Definition 25. *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.*

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.

Thus if $\sigma^2 = \omega(k^8 \log^4 n)$ then the multinomial distribution is well-approximated by the natural discrete Gaussian approximation.

Proof. Adopting the notation of Theorem 2, we let Z_i denote the distribution induced by the ith row of ρ, that is, a distribution over $(0, \ldots, 0), (1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots, (0, \ldots, 0, 1)$, where M^ρ is thus distributed as $\sum_{i=1}^n Z_i$. Since the range of Z_i has diameter $\sqrt{2}$, each sample from Z_i is within $\sqrt{2}$ of its mean. Theorem 2 implies that $d_W\left(M^\rho, \mathcal{N}(\mu, \Sigma)\right) < k\sqrt{2}(2.7 + 0.83 \log n)$.

For notational convenience, let $\phi = \mathcal{N}(\mu, \Sigma)$, and let $\phi^{disc} = \mathcal{N}^{disc}(\mu, \Sigma)$ denote the corresponding discretized Gaussian of Definition 25. We note that, since every point in $\mathbb{R}^k$ is within distance $\frac{\sqrt{k}}{2}$ from a lattice point, $d_W(\phi, \phi^{disc}) \leq \frac{\sqrt{k}}{2} \leq \frac{k}{2}$. Thus the triangle inequality yields $d_W(M^\rho, \phi^{disc}) < k\sqrt{2}(3.1 + 0.83 \log n)$.

Given positive integers d, ℓ, let $R_{d,\ell}$ denote the distribution over $\mathbb{Z}^k$ where the first d coordinates are each independent samples from the binomial distribution $B(2\ell, \frac{1}{2})$, shifted by $-\ell$ so as to lie in $\{-\ell, \ldots, \ell\}$ and the rest of the coordinates are 0.

The binomial distribution $B(2\ell, \frac{1}{2})$ is unimodal, with the probability of hitting its mode bounded by $\frac{1}{\sqrt{\pi\ell}}$, which implies that the statistical distance between $B(2\ell, \frac{1}{2})$ and a version shifted by an integer c is at most $\frac{c}{\sqrt{\pi\ell}}$; thus the same holds for shifting $R_{k,\ell}$ by c along any coordinate axis, since each coordinate is distributed as an independent (shifted) copy of $B(2\ell, \frac{1}{2})$. By the triangle inequality, if we shift by an integer vector x, then the statistical distance is at most $\frac{1}{\sqrt{\pi\ell}}\sum_{i=1}^{k}|x(i)|$. The Cauchy-Schwarz inequality yields $\sum_{i=1}^{k}|x(i)| \leq \sqrt{k}||x||$, yielding a bound on the statistical distance of $\frac{\sqrt{k}}{\sqrt{\pi\ell}}||x||$.

We are now prepared to make the key transformation from stating our central limit theorem in terms of earthmover distance, to stating a central limit theorem for statistical distance.

Consider a particular component of a "scheme to move earth" from M^ρ to ϕ^{disc}; for example, "move probability mass m from x to y". The bound of the previous paragraph implies that the statistical distance between copies of $R_{k,\ell}$ centered at x, and at y, respectively, is at most $\frac{\sqrt{k}}{\sqrt{\pi\ell}}||x-y||$. Thus, in this sense, convolution by $R_{k,\ell}$ converts earthmover bounds to statistical distance bounds, losing a factor of $\frac{\sqrt{k}}{\sqrt{\pi\ell}}$. We conclude that

$$d_{TV}(M^\rho * R_{k,\ell}, \phi^{disc} * R_{k,\ell}) \leq \frac{\sqrt{2k}\cdot k}{\sqrt{\pi\ell}}(3.1 + 0.83\log n). \qquad (13)$$

Were it not for the convolution by $R_{k,\ell}$ in the above expression, we could conclude here. We now consider how to "remove" these convolutions.

Consider ϕ (not ϕ^{disc}) shifted by a vector x. Since ϕ has variance at least σ^2 in every direction, then, when restricted to any line in the direction of x, ϕ will be a univariate normal distribution of variance at least σ^2. We may thus bound the statistical distance of ϕ and its shifted version by the corresponding univariate bound. Note that the univariate Gaussian is unimodal, and thus the statistical distance between itself and a version shifted $||x||$ is at most $||x||$ times the pdf at its mode, which is at most $\frac{1}{\sigma\sqrt{2\pi}}$. Applying this bound for each x drawn from $R_{k,\ell}$, where for each such x, $||x|| \leq \ell\sqrt{k}$ we have $d_{TV}(\phi, \phi * R_{k,\ell}) \leq \frac{\ell\sqrt{k}}{\sigma\sqrt{2\pi}}$. Since $R_{k,\ell}$ is a distribution on the lattice points, taking $\phi * R_{k,\ell}$ and rounding samples to the nearest integer is distributed identically to $\phi^{disc} * R_{k,\ell}$. Thus we have $d_{TV}(\phi^{disc}, \phi^{disc} * R_{k,\ell}) \leq \frac{\ell\sqrt{k}}{\sigma\sqrt{2\pi}}$, yielding, by the triangle inequality, $d_{TV}(M^\rho * R_{k,\ell}, \phi^{disc}) \leq \frac{\sqrt{2k}\cdot k}{\sqrt{\pi\ell}}(3.1 + 0.83\log n) + \frac{\ell\sqrt{k}}{\sigma\sqrt{2\pi}}$

Having "removed" the second convolution by $R_{k,\ell}$ in Equation 13, we now turn to the first. Recalling that $R_{i,\ell}$ is the distribution whose first i coordinates are distributed as (shifted) versions of the binomial distribution $B(2\ell, \frac{1}{2})$ where the remaining $k-i$ coordinates are 0, we aim to "deconvolve" by this binomial, coordinate-by-coordinate, so that when i reaches 0 we will have the desired central limit theorem. Our tool is Lemma 24, which we will use to show by induction that for every $i \in \{0, \ldots, k\}$ we have

$$d_{TV}(M^\rho * R_{i,\ell}, \phi^{disc}) \leq (k-i)\frac{5\ell}{\sigma\sqrt{2\pi}} + \frac{\ell\sqrt{k}}{\sigma\sqrt{2\pi}} + \frac{\sqrt{2k}\cdot k}{\sqrt{\pi\ell}}(3.1 + 0.83\log n) \qquad (14)$$

Letting $i = 0$ and $\ell = \frac{1}{6^{2/3}}\sigma^{2/3}k^{1/3}(3.1 + 0.83\log n)^{2/3}$ yields the theorem.

To prove Equation 14, we assume as our induction hypothesis that it holds for some $i > 0$ and will derive it for $i-1$. Consider $M^\rho * R_{i,\ell}$, $M^\rho * R_{i-1,\ell}$, and ϕ^{disc} restricted to a line L in the ith coordinate direction. We note that the pdf of ϕ restricted to this line will be a multiple of a univariate normal distribution of variance at least σ^2, and thus has the property that its maximum is at most $\frac{1}{\sigma\sqrt{2\pi}}$ times its integral; as this is true for every such line, it is also true in expectation for a distribution of lines, and is thus true for the distribution of lines that will be rounded to L. Thus ϕ^{disc} restricted to the line L has the property that its maximum is at most $\frac{1}{\sigma\sqrt{2\pi}}$ times its total. With a view towards applying Lemma 24, we note that $R_{i-1,\ell}$ is itself a generalized multinomial distribution, and hence so is $M^\rho * R_{i-1,\ell}$, from which we invoke Lemma 23 to see that $M^\rho * R_{i-1,\ell}$ is unimodal along L. We thus apply Lemma 24 with f equal to the restriction of $M^\rho * R_{i-1,\ell}$ to L, g equal to the binomial $B(2\ell, \frac{1}{2})$ shifted so as to have support on $\{-\ell, \ldots, \ell\}$, and h equal to the restriction of ϕ^{disc} to L. Since $f * g$ is the restriction of $M^\rho * R_{i,\ell}$ to L, we conclude that,

$$\begin{aligned}\sum_{x \in L} |(M^\rho * R_{i-1,\ell})(x) - \phi^{disc}(x)| &\leq 10\ell \left(\max_{x \in L} \phi^{disc}(x)\right) + \sum_{x \in L} |(M^\rho * R_{i,\ell})(x) - \phi^{disc}(x)| \\ &\leq \frac{10\ell}{\sigma\sqrt{2\pi}} \sum_{x \in L} \phi^{disc}(x) + \sum_{x \in L} |(M^\rho * R_{i,\ell})(x) - \phi^{disc}(x)|\end{aligned}$$

Summing over all such lines L yields the induction (since statistical distance has a normalizing factor of $\frac{1}{2}$). □

C Linear Combinations of Poisson Functions

In this section we show that one can closely approximate the function $poi(2x, i)$ as a sum $\sum_{j=0}^{\infty} \alpha_j \cdot poi(x, j)$, such that $\sum_j \alpha_j$ is not too large. We note that the Stone-Weierstrass theorem of Analysis trivially implies the convergence of this type of approximation; however, we require much stronger bounds on the relationship between the approximation factor and the coefficient sizes.

We prove these strong bounds via a Fourier analysis approach relying on properties of Hermite polynomials.

To see the intuition both behind the result, and our approach, consider the above problem but with Poisson functions replaced by Gaussians, and all errors evaluated in the L_2 sense: for each $\epsilon > 0$ there exists a function K_ϵ of L_2 norm $\frac{1}{\epsilon}$ that when convolved with $\mathcal{N}(0,1)$ approximates $\mathcal{N}(0, \frac{1}{2})$ to within ϵ, in the L_2 sense. Let $\hat{K}_\epsilon$ be the ratio of the Fourier transforms of the pdfs of $\mathcal{N}(0,1)$ and $\mathcal{N}(0, \frac{1}{2})$ respectively, restricted to be 0 outside the interval $[-2\sqrt{\log \frac{1}{\epsilon}}, 2\sqrt{\log \frac{1}{\epsilon}}]$ and let K_ϵ be the inverse Fourier transform of $\hat{K}_\epsilon$. By Parseval's theorem, we may bound the L_2 norm of K_ϵ and the L_2 norm of the error $||\mathcal{N}(0, \frac{1}{2}), K_\epsilon * \mathcal{N}(0,1)||_2$, as the L_2 norms of their corresponding Fourier transforms. As the Fourier transform of K_ϵ is $\hat{K}_\epsilon$, which grows as $e^{x^2/4}$ but is zero outside the interval $[-2\sqrt{\log \frac{1}{\epsilon}}, 2\sqrt{\log \frac{1}{\epsilon}}]$, its L_2 norm is roughly $\frac{1}{\epsilon}$. Further, the Fourier transform of $K_\epsilon * \mathcal{N}(0,1)$ equals $\hat{K}_\epsilon \cdot \mathcal{N}(0,1)$, which by construction is exactly the Fourier transform of $\mathcal{N}(0, \frac{1}{2})$ within the interval $[-2\sqrt{\log \frac{1}{\epsilon}}, 2\sqrt{\log \frac{1}{\epsilon}}]$, and zero outside this interval. Since the Fourier transform of $\mathcal{N}(0, \frac{1}{2})$ decays as $e^{-x^2/4}$, the L_2 norm of the portion outside this interval is thus roughly ϵ, the desired bound.

Our proof of the following lemma relies on the substitution $x \to x^2$ to make the Poisson functions "look like" Gaussians, where the relationship between the transformed Poisson functions and Gaussians is controlled by properties of Hermite polynomials. Additionally, since we require an L_1 bound on the coefficients, as opposed to the L_2 bound that comes more naturally (via Parseval's theorem), instead of a sharp cutoff outside a designated interval (as we had done in the previous paragraph in our construction of K_ϵ), we must use a smooth cutoff function T, constructed as the convolution of the indicator function of an interval with a Gaussian of carefully chosen width.

Lemma 18$\star$ *For any $\epsilon > 0$ and integer $i \geq 0$, one may approximate $poi(2x, i)$ as a linear combination $\sum_{j=0}^{\infty} \alpha(j) poi(x, j)$ such that*

1. *For all $x \geq 0$, $|poi(2x, i) - \sum_{j=0}^{\infty} \alpha(j) poi(x, j)| \leq \epsilon$; and*

2. *$\sum_{j=0}^{\infty} |\alpha(j)| \leq \frac{1}{\epsilon} \cdot 200 \max\{\sqrt[4]{i}, 24 \log^{3/2} \frac{1}{\epsilon}\}$.*

Proof. Let $g_k(x) := poi(x^2/2, k) = \frac{e^{-x^2/2}x^{2k}}{2^k k!}$. We consider the Fourier transform of $g_k(x)$, using the facts that the Fourier transform of $f(x) = e^{-x^2/2}$ is $\hat{f}(w) = e^{-w^2/2}$, and that if $f(x)$ is differentiable with Fourier transform $\hat{f}(w)$, then the Fourier transform of $\frac{d}{dx}f(x)$ is $-\mathbf{i}w\hat{f}(w)$:

$$\begin{aligned} \hat{g_k}(w) &= (-\mathbf{i})^{2k}\frac{d^{2k}}{dw^{2k}}\left(e^{-w^2/2}\right)\cdot\frac{1}{2^k k!} \\ &= \frac{(-1)^k e^{-w^2/2}H_{2k}(w)}{2^k k!}, \end{aligned}$$

where $H_j(x) := (-1)^j e^{x^2/2}\frac{d^j}{dx^j}e^{-x^2/2}$, is the jth Hermite polynomial. Since Hermite polynomials form an orthogonal basis with respect to the Gaussian measure $e^{-w^2/2}$, and the even numbered Hermite polynomials are even functions while the odd numbered Hermite polynomials are odd functions, we have that the even numbered Hermite polynomials form an orthogonal basis with respect to the Gaussian measure $e^{-w^2/2}$ for the set of even functions. Incorporating the (square root) of the normalizing function $e^{-w^2/2}$ into the basis yields that the set of functions $\hat{g}_k(w)e^{w^2/4}$ form an orthogonal basis for the set of even functions with respect to the *uniform* measure. In particular, since the set of functions $e^{-w^2/4}H_{2k}(w)/\sqrt{(2k)!\sqrt{2\pi}}$, sometimes known as the Hermite functions, are *orthonormal*, we define the orthonormal basis for even functions $\mathcal{G}_k(w) = \hat{g}_k(w)e^{w^2/4}\frac{2^k k!}{\sqrt{(2k)!\sqrt{2\pi}}}$.

Define $h_i(x) = g_i(x\sqrt{2})$. Recall our goal of approximating h_i as a linear combination of $\{g_j\}$. We work in Fourier space, and more specifically, to compute a linear combination of $\{\hat{g}_j\}$ which approximates $\hat{h}_i$, we multiply both sides by $e^{w^2/4}$ so that we may make use of the orthonormal basis $\{\mathcal{G}_j\}$. Explicitly, defining $T_{r,c}(w) = I_{[-r,r]}(w) * e^{-cw^2}\frac{\sqrt{c}}{\sqrt{\pi}}$, where $I_{[-r,r]}$ denotes the indicator function of the interval $[-r, r]$, for constants c and r to be specified later, and "$*$" denotes convolution, we use the basis $\{\mathcal{G}_j\}$ to express $T_{r,c}(w)\cdot e^{w^2/4}\cdot\hat{h}_i(w)$. Since $\{\mathcal{G}_j\}$ is orthonormal, the coefficient of $\mathcal{G}_j$ is exactly the inner product of $\mathcal{G}_j$ with this expression. That is, defining

$$\beta_{i,r,c}(j) \triangleq \int_{-\infty}^{\infty} T_{r,c}(w)\cdot e^{w^2/4}\cdot\hat{h}_i(w)\mathcal{G}_j(w)dw = \frac{2^j j!}{\sqrt{(2j)!\sqrt{2\pi}}}\int_{-\infty}^{\infty} T_{r,c}(w)\cdot e^{w^2/2}\cdot\hat{h}_i(w)\hat{g}_j(w)dw$$

we have expressed $T_{r,c}(w)\cdot e^{w^2/4}\cdot\hat{h}_i(w) = \sum_{j=0}^{\infty}\beta_{i,r,c}(j)\cdot\mathcal{G}_j(w)$. Invoking the definition of $\mathcal{G}_j$ and dividing both sides by $e^{w^2/4}$, we see that if we define

$$\alpha_{i,r,c}(j) \triangleq \frac{2^j j!}{\sqrt{(2j)!\sqrt{2\pi}}}\beta_{i,r,c}(j) = \frac{2^{2j}(j!)^2}{(2j)!\sqrt{2\pi}}\int_{-\infty}^{\infty} T_{r,c}(w)\cdot e^{w^2/2}\cdot\hat{h}_i(w)\hat{g}_j(w)dw, \tag{15}$$

then we have expressed

$$T_{r,c}(w)\cdot\hat{h}_i(w) = \sum_{j=0}^{\infty}\alpha_{i,r,c}(j)\cdot\hat{g}_j(w). \tag{16}$$

We bound $|\alpha_{i,r,c}(j)|$ in two ways from Equation 15.

We first note that since for a real number $a \neq 0$, the Fourier transform of a function $s(x) = f(a\cdot x)$ is $\hat{s}(w) = \frac{1}{a}\hat{f}(w/a)$, we have $\hat{h}_i(w) = \frac{1}{\sqrt{2}}\hat{g}_i(\frac{w}{\sqrt{2}})$. Further, we recall the basic fact that $|\mathcal{G}_j(w)|$ is maximized, over all j and w, when $j = w = 0$ (see [32] p. 190). Thus by definition of $\mathcal{G}_j(w)$, we bound $|e^{w^2/4}\hat{g}_j(w)| \leq \frac{\sqrt{(2j)!\sqrt{2\pi}}}{2^j j!}\mathcal{G}_0(0) = \frac{\sqrt{(2j)!}}{2^j j!}$, and thus since $\hat{h}_i(w) = \frac{1}{\sqrt{2}}\hat{g}_i(\frac{w}{\sqrt{2}})$, we have $|e^{w^2/8}\hat{h}_i(w)| \leq \frac{\sqrt{(2i)!}}{2^i i!\sqrt{2}}$. Thus we may bound

$$|\alpha_{i,r,c}(j)| \leq \frac{2^j j!}{\sqrt{(2j)!2\pi}}\frac{\sqrt{(2i)!}}{2^i i!\sqrt{2}}\int_{-\infty}^{\infty} T_{r,c}(w)\cdot e^{w^2/8}dw$$

To evaluate this integral, we use a trick which we will use twice more below: we "complete the square" and make the substitution (in this case) $s = t\frac{c}{c-1/8}$, yielding

$$\begin{aligned}
T_{r,c}(w) \cdot e^{w^2/8} &\triangleq \frac{\sqrt{c}}{\sqrt{\pi}} \int_{-r}^{r} e^{w^2/8} e^{-c(w-t)^2} dt \\
&= \frac{\sqrt{c}}{\sqrt{\pi}} \int_{-r}^{r} e^{-(w\sqrt{c-1/8} - tc/\sqrt{c-1/8})^2} e^{t^2 \frac{c}{8(c-1/8)}} dt \\
&= \frac{c-\frac{1}{8}}{\sqrt{c\pi}} \int_{-rc/(c-1/8)}^{rc/(c-1/8)} e^{-(w-s)^2 \cdot (c-1/8)} e^{s^2 \cdot \frac{c-1/8}{8c}} ds \\
&= \frac{c-\frac{1}{8}}{\sqrt{c\pi}} \left[I_{[-r\frac{c}{c-\frac{1}{8}}, r\frac{c}{c-\frac{1}{8}}]}(w) \cdot e^{\frac{c-1/8}{8c} \cdot w^2} \right] * e^{-(c-\frac{1}{8})w^2}.
\end{aligned}$$

We may thus integrate this over $\mathbb{R}$ as the product of the integrals of the terms on each side of the convolution, that is: $\frac{c-\frac{1}{8}}{\sqrt{c\pi}} \cdot \frac{\sqrt{8\pi c}}{\sqrt{c-1/8}} \text{erfi}\left(r\sqrt{\frac{c}{8(c-\frac{1}{8})}} \right) \cdot \frac{\sqrt{\pi}}{\sqrt{c-1/8}} = \sqrt{8\pi} \text{erfi}\left(r\sqrt{\frac{c}{8(c-\frac{1}{8})}} \right)$, where erfi is the imaginary error function, defined as $\text{erfi}(x) \triangleq \frac{2}{\sqrt{pi}} \int_0^x e^{y^2} dy$. Noting the bound that $\text{erfi}(x) \leq \frac{3}{4}\frac{1}{x}e^{x^2}$ (which can be derived by differentiating), we have

$$|\alpha_{i,r,c}(j)| \leq \frac{2^j j!}{\sqrt{(2j)!2\pi}} \frac{\sqrt{(2i)!}}{2^i i! \sqrt{2}} \sqrt{8\pi} \frac{3}{4} \frac{\sqrt{8}}{r} \sqrt{\frac{c-\frac{1}{8}}{c}} e^{\frac{r^2}{8} \frac{c}{c-1/8}} = \frac{2^j j!}{\sqrt{(2j)!}} \frac{\sqrt{(2i)!}}{2^i i!} \frac{3}{r} \sqrt{\frac{c-\frac{1}{8}}{c}} e^{\frac{r^2}{8} \frac{c}{c-1/8}} \tag{17}$$

To bound $\alpha_{i,r,c}(j)$ a second way, we first note that a second application of "completing the square" allows us to reexpress part of Equation 15 as

$$T_{r,c}(w) \cdot e^{w^2/2} = \frac{c-\frac{1}{2}}{\sqrt{c\pi}} \left[I_{[-r\frac{c}{c-\frac{1}{2}}, r\frac{c}{c-\frac{1}{2}}]}(w) \cdot e^{\frac{c-1/2}{2c} \cdot w^2} \right] * e^{-(c-\frac{1}{2})w^2}.$$

Let $f_{r,c}(x)$ be the inverse Fourier transform of $I_{[-r\frac{c}{c-\frac{1}{2}}, r\frac{c}{c-\frac{1}{2}}]}(w) \cdot e^{\frac{c-1/2}{2c} \cdot w^2}$. We thus evaluate Equation 15 by noting that inner products are preserved under Fourier transform, that the (inverse) Fourier transform of $e^{-(c-\frac{1}{2})w^2}$ equals $\frac{1}{\sqrt{2(c-\frac{1}{2})}} e^{-\frac{1}{4(c-1/2)}x^2}$, and that multiplication and convolution swap roles under the Fourier transform, we have that

$$\alpha_{i,r,c}(j) = \frac{2^{2j}(j!)^2}{(2j)!\sqrt{2\pi}} \frac{\sqrt{c-\frac{1}{2}}}{\sqrt{2c\pi}} \int_{-\infty}^{\infty} \left[\left(f_{r,c}(x) \cdot e^{-\frac{1}{4(c-1/2)}x^2} \right) * h_i(x) \right] \cdot g_j(x) dx \tag{18}$$

By definition of the Fourier transform, for any function f, we have $||f||_\infty \leq \frac{1}{\sqrt{2\pi}} ||\hat{f}||_1$. Thus we may bound the maximum value of $|f_{r,c}|$ by $\frac{1}{\sqrt{2\pi}}$ times the L_1 norm of its Fourier transform, that is,

$$|f_{r,c}(x)| \leq \frac{1}{\sqrt{2\pi}} \int_{-r\frac{c}{c-\frac{1}{2}}}^{r\frac{c}{c-\frac{1}{2}}} e^{\frac{c-1/2}{2c}w^2} dw = \sqrt{\frac{c}{c-\frac{1}{2}}} \text{erfi}\left(\frac{r}{\sqrt{2}} \sqrt{\frac{c}{c-\frac{1}{2}}} \right)$$

We now bound $g_j(x) = \frac{e^{-x^2/2} x^{2j}}{2^j j!}$, the final term of Equation 18, by noting that, since $x \leq e^{x-1}$ always, and replacing x by x/y yields $x \leq e^{x/y-1}y$, we set $y = \sqrt{2j}$ and raise both sides to the power $2j$ to yield that, for positive x,

$$g_j(x) = \frac{e^{-x^2/2} x^{2j}}{2^j j!} \leq \frac{e^{-x^2/2 + x\sqrt{2j} - 2j} j^j}{j!} = e^{-\frac{1}{2}(x-\sqrt{2j})^2} \frac{e^{-j} j^j}{j!}$$

Thus by definition of $h_i(x) = g_i(x\sqrt{2})$ we have $h_i(x) \leq e^{-(x-\sqrt{i})^2}\frac{e^{-i}i^i}{i!}$ for positive x. Generally, we may see that $h_i(x) \leq \left(e^{-(x-\sqrt{i})^2} + e^{-(x+\sqrt{i})^2}\right)\frac{e^{-i}i^i}{i!}$ for all x. We may thus bound Equation 18 as

$$|\alpha_{i,r,c}(j)| \leq \frac{2^{2j}(j!)^2}{(2j)!2\pi}\text{erfi}\left(\frac{r}{\sqrt{2}}\sqrt{\frac{c}{c-\frac{1}{2}}}\right)\frac{e^{-i}i^i}{i!}\frac{e^{-j}j^j}{j!}\sum_{\pm,\pm}\int_{-\infty}^{\infty}\left[e^{-\frac{1}{4(c-1/2)}x^2} * e^{-(x\pm\sqrt{i})^2}\right]\cdot e^{-\frac{1}{2}(x\pm\sqrt{2j})^2}dx,$$

where the summation is over the four possible combinations of the two choices of "±". We note that the integral is equal to the convolution of the three terms inside of it, evaluated at $x = 0$, namely $\sqrt{\frac{8(c-\frac{1}{2})}{4c+1}}e^{-\frac{1}{4c+1}(x\pm\sqrt{i}\pm\sqrt{2j})^2}\Big|_{x=0}$, since the denominators in the exponents of Gaussians add under convolution. Thus we bound

$$|\alpha_{i,r,c}(j)| \leq \frac{2^{2j}(j!)^2}{(2j)!2\pi}\text{erfi}\left(\frac{r}{\sqrt{2}}\sqrt{\frac{c}{c-\frac{1}{2}}}\right)\frac{e^{-i}i^i}{i!}\frac{e^{-j}j^j}{j!}\sqrt{\frac{8(c-\frac{1}{2})}{4c+1}}\cdot 4\cdot e^{-\frac{1}{4c+1}|\sqrt{i}-\sqrt{2j}|^2}$$

Since, as noted above, $\text{erfi}(x) \leq \frac{3}{4}\frac{1}{x}e^{x^2}$, we have

$$|\alpha_{i,r,c}(j)| \leq \frac{2^{2j}e^{-j}j^j j!}{(2j)!2\pi}\frac{e^{-i}i^i}{i!}\frac{4(c-\frac{1}{2})}{\sqrt{c(4c+1)}}\cdot\frac{3}{r}\cdot e^{\frac{r^2}{2}\frac{c}{c-1/2}-\frac{1}{4c+1}|\sqrt{i}-\sqrt{2j}|^2}$$

We bound $\frac{2^{2j}e^{-j}j^j j!}{(2j)!} \leq 1$ as a combination of Stirling's formula, $e^{-j}j^j \leq \frac{j!}{\sqrt{2\pi j}}$, and the bound on the middle binomial coefficient $\binom{2j}{j} \geq \frac{2^{2j}}{\sqrt{2\pi j}}$. A second application of Stirling's formula yields that $\frac{e^{-i}i^i}{i!} \leq \frac{1}{\sqrt{2\pi i}}$, and we trivially bound $\frac{4(c-\frac{1}{2})}{\sqrt{c(4c+1)}} \leq 2$ to yield

$$|\alpha_{i,r,c}(j)| \leq \frac{3}{\pi r\sqrt{2\pi i}}\cdot e^{\frac{r^2}{2}\frac{c}{c-1/2}-\frac{1}{4c+1}|\sqrt{i}-\sqrt{2j}|^2} \tag{19}$$

Having thus derived two bounds on $|\alpha_{i,r,c}(j)|$, that of Equation 17 and that of Equation 19, we now aim to bound $\sum_{j\geq 0}|\alpha_{i,r,c}(j)|$ via a combination of these bounds: using Equation 17 when $2j$ is near i, and using Equation 19 otherwise.

Let $c = r^2$, and consider two cases.

Case 1: $i \leq 2c^2$.

We first bound $\sum_{j\geq 4c^2}|\alpha_{i,r,c}(j)|$ from Equation 19. Specifically, consider $\sum_{j\geq 4c^2}e^{-\frac{1}{4c+1}|\sqrt{i}-\sqrt{2j}|^2}$. We note that the first term of the sum is at most $e^{-\frac{2c^2}{4c+1}} \leq e^{-\frac{c}{2}}e^{\frac{1}{8}}$. To bound the ratio between successive terms, we note that $\frac{d}{dj}(\sqrt{i}-\sqrt{2j})^2 = 2(1-\frac{\sqrt{i}}{\sqrt{2j}}) \geq 1$, which implies $\sum_{j\geq 4c^2}e^{-\frac{1}{4c+1}|\sqrt{i}-\sqrt{2j}|^2} \leq e^{-\frac{c}{2}}e^{\frac{1}{8}}\sum_{\ell\geq 0}e^{-\frac{1}{4c+1}\ell} = e^{-\frac{c}{2}}e^{\frac{1}{8}}\frac{1}{1-e^{-1/(4c+1)}}$. We note the general inequality $e^a \geq 1+a$, or equivalently, $e^{1/a} \geq 1+\frac{1}{a}$, which may be rearranged to $\frac{1}{1-e^{-1/a}} \leq a+1$, yielding a bound of $(4c+2)e^{-\frac{c}{2}}e^{\frac{1}{8}}$ on the sum. To bound the sum of Equation 19, we note that for $c \geq 1$, we have $\frac{r^2}{2}\frac{c}{c-1/2} \leq \frac{c}{2}+\frac{1}{2}$, leading to a bound of $\sum_{j\geq 4c^2}|\alpha_{i,r,c}(j)| \leq \frac{3}{\pi\sqrt{2\pi ic}}(4c+2)e^{5/8} < 5\sqrt{\frac{c}{i}}$

To bound $|\alpha_{i,r,c}(j)|$ for small j we instead use Equation 17. We note for $\ell \geq 1$ the bounds on the middle binomial coefficient of $\frac{1}{\sqrt{2\pi\ell}} \leq 2^{-2\ell}\binom{2\ell}{\ell} \leq 1$. Further, for $c \geq 1$ we have $\frac{r^2}{8}\frac{c}{c-1/8} \leq \frac{c}{8}+\frac{1}{56}$, yielding that $\sum_{j<4c^2}|\alpha_{i,r,c}(j)| \leq 4c^2\sqrt[4]{2\pi\cdot 4c^2}\frac{3}{r}e^{1/56}e^{c/8} < 28c^2e^{c/8}$. Combining this with the result of the previous paragraph yields $\sum_{j=0}^{\infty}|\alpha_{i,r,c}(j)| \leq 32c^2e^{c/8}$.

Case 2: $i > 2c^2$.

We use the bound of Equation 17 when $j \in (\frac{i}{2} - 2c\sqrt{i}, \frac{i}{2} + 3c\sqrt{i})$, and Equation 19 otherwise.

Consider $\sum_{j \geq \frac{i}{2}+3c\sqrt{i}} |\alpha_{i,r,c}(j)|$. Invoking Equation 19, we analyze $\sum_{j \geq \frac{i}{2}+3c\sqrt{i}} e^{-\frac{1}{4c+1}|\sqrt{i}-\sqrt{2j}|^2}$. We aim for $|\sqrt{i}-\sqrt{2j}| \geq \sqrt{2}c$, and show this by considering $(\sqrt{i}+\sqrt{2}c)^2 = i+2\sqrt{2}\sqrt{i}c+2c^2 < i+3\sqrt{2}\sqrt{i}c < 2j$, as desired. Thus the first term of this sum is at most $e^{-\frac{2c^2}{4c+1}} \leq e^{-\frac{c}{2}}e^{\frac{1}{8}}$. As above, we bound the ratio of successive terms by noting that $\frac{d}{dj}(\sqrt{i} - \sqrt{2j})^2 = 2(1 - \frac{\sqrt{i}}{\sqrt{2j}}) \geq \frac{c\sqrt{2}}{\sqrt{i}}$, which implies that $\sum_{j \geq \frac{i}{2}+3c\sqrt{i}} e^{-\frac{1}{4c+1}|\sqrt{i}-\sqrt{2j}|^2} \leq e^{-\frac{c}{2}}e^{\frac{1}{8}} \sum_{\ell \geq 0} e^{-\frac{c\sqrt{2}}{(4c+1)\sqrt{i}}} = e^{-\frac{c}{2}}e^{\frac{1}{8}} \frac{1}{1-e^{-c\sqrt{2}/((4c+1)\sqrt{i})}}$, which, as analyzed in the previous case, yields a bound of $e^{-\frac{c}{2}}e^{\frac{1}{8}}(\frac{(4c+1)\sqrt{i}}{c\sqrt{2}} + 1) \leq 4\sqrt{i}e^{-\frac{c}{2}}$ on $\sum_{j \geq \frac{i}{2}+3c\sqrt{i}} e^{-\frac{1}{4c+1}|\sqrt{i}-\sqrt{2j}|^2}$.

We now bound the small terms of the sum, $\sum_{j \leq \frac{i}{2}-2c\sqrt{i}} e^{-\frac{1}{4c+1}|\sqrt{i}-\sqrt{2j}|^2}$. As above, we show that $\sqrt{i} - \sqrt{2j} \geq \sqrt{2}c$ for such j by noting that $(\sqrt{i} - \sqrt{2}c)^2 = i - 2\sqrt{2}\sqrt{i} + 2c^2 > 2j$. Thus the last term in the sum is at most $e^{-\frac{2c^2}{4c+1}} \leq e^{-\frac{c}{2}}e^{\frac{1}{8}}$. As above, we bound the ratio of successive terms, this time as j decreases, by noting $\frac{d}{dj}(\sqrt{i} - \sqrt{2j})^2 = \frac{2}{\sqrt{2j}}(\sqrt{2j} - \sqrt{i})$, which since $2j < i$, has magnitude at least $\frac{2\sqrt{2}c}{\sqrt{i}}$. Thus the bound of the previous paragraph holds, yielding $\sum_{j \leq \frac{i}{2}-2c\sqrt{i}} e^{-\frac{1}{4c+1}|\sqrt{i}-\sqrt{2j}|^2} \leq 4\sqrt{i}e^{-\frac{c}{2}}$. As shown in Case 1, the remaining part of Equation 19 is bounded as $\frac{3}{\pi r\sqrt{2\pi i}} \cdot e^{\frac{r^2}{2}\frac{c}{c-1/2}} \leq \frac{3}{\pi r\sqrt{2\pi i}}e^{c/2}e^{1/2}$, yielding $\sum_{j \notin (\frac{i}{2}-2c\sqrt{i}, \frac{i}{2}+3c\sqrt{i})} |\alpha_{i,r,c}(j)| \leq 8\sqrt{i}\frac{3}{\pi r\sqrt{2\pi i}}e^{1/2} < 6$.

For intermediate $j \in (\frac{i}{2} - 2c\sqrt{i}, \frac{i}{2} + 3c\sqrt{i})$ we bound $|\alpha_{i,r,c}(j)|$ from Equation 17. From the fact that $i!$ lies between its Stirling estimate and 1.1 times its Stirling estimate, we have that $\frac{2^j j!}{\sqrt{(2j)!}} \in (\sqrt[4]{\pi j}, 1.1\sqrt[4]{\pi j})$. Thus, since $j < 6i$, we have $\frac{2^j j!}{\sqrt{(2j)!}}\frac{\sqrt{(2i)!}}{2^i i!} \leq 1.1\sqrt[4]{6} < 2$, and we thus bound Equation 17 as $|\alpha_{i,r,c}(j)| \leq 2\frac{3}{r}e^{1/56}e^{c/8}$, and the sum of the $5c\sqrt{i}$ of these terms as at most $31\sqrt{ci}e^{c/8}$. Combining this result with that of the previous paragraph yields $\sum_{j=0}^{\infty} |\alpha_{i,r,c}(j)| \leq 32\sqrt{ci}e^{c/8}$.■

Having bounded $\sum_{j=0}^{\infty} \alpha_{i,r,c}(j)|$, namely the second claim of the theorem, we now turn to bounding the first claim of the theorem—showing that the error of our approximation is small. As above, our expressions will involve the parameter c; as the final step of the proof, we choose c appropriately to obtain the claimed bounds.

Taking the inverse Fourier transform of both sides of Equation 16 yields that the difference between $h_i(w)$ and $\sum_{j=0}^{\infty} \alpha_{i,r,c}(j) \cdot g_j(w)$ equals the inverse Fourier transform of $(1 - T_{r,c}(w))\hat{h}_i(w)$; we thus aim to bound the absolute value of this, pointwise. We note that from the definition of the Fourier transform, for a function f, $||f||_\infty \leq \frac{1}{\sqrt{2\pi}}||\hat{f}||_1$, so thus the maximum error of our approximation is bounded by $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty}(1 - T_{r,c}(w))|\hat{h}_i(w)|dw \leq \frac{\sqrt{(2i)!}}{2^i i! 2\sqrt{\pi}} \int_{-\infty}^{\infty}(1 - T_{r,c}(w))e^{-w^2/8}dw$. Again using the "completing the square" trick yields that this integral equals $\sqrt{8\pi}\text{erfc}\left(r\sqrt{\frac{c}{8(c+\frac{1}{8})}}\right) \leq \sqrt{8\pi}\text{erfc}(\frac{\sqrt{c}}{\sqrt{8}})$, where erfc $= 1 -$ erf is the complementary error function. Noting the general bound that $\text{erfc}(x) \leq \frac{e^{-x^2}}{x\sqrt{\pi}}$, and from the above bound that $\frac{2^j j!}{\sqrt{(2j)!}} \geq \sqrt[4]{\pi j}$, the maximum error of our approximation is seen to be at most $\frac{8}{\sqrt[4]{\pi i}\sqrt{c}}e^{-c/8}$.

We have thus shown that $\sum_{j=0}^{\infty} \alpha_{i,r,c}(j) poi(x, j)$ approximates $poi(2x, i)$ to within $\frac{8}{\sqrt[4]{\pi i}\sqrt{c}}e^{-c/8}$, pointwise, while $\sum_{j=0}^{\infty} |\alpha_{i,r,c}(j)|$ is at most $32e^{c/8}\max\{c^2, \sqrt{ci}\}$, where c is arbitrary. Thus for desired error ϵ, we may choose $c \leq 8|\log \epsilon|$ so as to make $\frac{8}{\sqrt[4]{\pi i}\sqrt{c}}e^{-c/8} = \epsilon$, yielding that

$$\sum_{j=0}^{\infty} |\alpha_{i,r,c}(j)| \leq 32e^{c/8}\max\{c^2, \sqrt{ci}\} = \frac{1}{\epsilon} \cdot 200\max\{\sqrt[4]{i}, \frac{c\sqrt{c}}{\sqrt[4]{i}}\} \leq \frac{1}{\epsilon} \cdot 200\max\{\sqrt[4]{i}, 24\log^{3/2}\frac{1}{\epsilon}\},$$

as desired. $\square$

D Gaussian Facts

This section contains a straightforward analysis of the statistical distance between two multivariate Gaussians. The result in this section that is used in the main body of the paper is Proposition 30, which bounds this distance when the covariance matrices have no small eigenvalues, and are close element-by-element.

Fact 26. *Given independent real-valued random variables W, X, Y, Z the total variation distance satisfies $D_{tv}\left((W,X),(Y,Z)\right) \leq D_{tv}(W,Y) + D_{tv}(X,Z)$, where (W,X) and (Y,Z) denote joint distributions.*

Proof.

$$\begin{aligned} D_{tv}\left((W,X),(Y,Z)\right) &= \frac{1}{2}\int\int |P_W(a)P_X(b) - P_Y(a)P_Z(b)| da\ db \\ &= \frac{1}{4}\int\int |(P_W(a) - P_Y(a))(P_X(b) + P_Z(b)) + (P_W(a) + P_Y(a))(P_X(b) - P_Z(b))| da\ db \\ &\leq \frac{1}{4}\int\int |(P_W(a) - P_Y(a))(P_X(b) + P_Z(b))| da\ db \\ &\quad + \frac{1}{4}\int\int (P_W(a) + P_Y(a))(P_X(b) - P_Z(b))| da\ db \\ &= \frac{1}{2}\int |(P_W(a) - P_Y(a))| da + \frac{1}{2}\int (P_X(b) - P_Z(b))| db = D_{tv}(W,Y) + D_{tv}(X,Z). \end{aligned}$$

$\square$

Fact 27. *Letting $\mathcal{N}(\mu, \sigma^2)$ denote the univariate Gaussian distribution,*

$$D_{tv}(\mathcal{N}(\mu,1), \mathcal{N}(\mu+\alpha,1)) \leq |\alpha|/\sqrt{2\pi}.$$

Fact 28. *Letting $\mathcal{N}(\mu, \sigma^2)$ denote the univariate Gaussian distribution,*

$$D_{tv}(\mathcal{N}(\mu,1), \mathcal{N}(\mu,\sigma^2)) \leq \frac{\max(\sigma^2, 1/\sigma^2) - 1}{\sqrt{2\pi e}}.$$

Fact 29. *Given two Gaussian distributions in m dimensions $G_1 = \mathcal{N}(\mu_1, \Sigma_1)$, and $G_2 = \mathcal{N}(\mu_2, \Sigma_2)$, where $\Sigma_1 = TT'$, is the Cholesky decomposition of Σ_1, then*

$$D_{tv}(G_1, G_2) \leq \sum_{i=1}^{m} \frac{\max(\lambda_i, 1/\lambda_i) - 1}{\sqrt{2\pi e}} + \frac{||T^{-1}(\mu_1 - \mu_2)||}{\sqrt{2\pi}},$$

where λ_i is the ith eigenvalue of $T^{-1}\Sigma_2 T'^{-1}$.

Proof. Since variational distance is affine-invariant, applying the affine transformation T^{-1}, we have $D_{tv}(G_1, G_2) = D_{tv}\left(\mathcal{N}(0, T^{-1}\Sigma_1 T'^{-1}), \mathcal{N}(T^{-1}(\mu_1 - \mu_2), T^{-1}\Sigma_2 T'^{-1})\right)$, where we have $T^{-1}\Sigma_1 T'^{-1} = I$, the $m \times m$ identity. Thus, by the triangle inequality, this distance is at most

$$D_{tv}\left(\mathcal{N}(0,I), \mathcal{N}(T^{-1}(\mu_1 - \mu_2), I)\right) + D_{tv}\left(\mathcal{N}(0,I), \mathcal{N}(0, T^{-1}\Sigma_2 T'^{-1})\right).$$

Viewing $\mathcal{N}(T^{-1}(\mu_1 - \mu_2), I)$ as the joint distribution of m independent univariate Gaussians, where the first $m-1$ distributions are $\mathcal{N}(0,1)$, and the mth distribution is $\mathcal{N}(||T^{-1}(\mu_1-\mu_2)||, 1)$, by Facts 26 and 27 we get that

$$D_{tv}\left(\mathcal{N}(0,I), \mathcal{N}(T^{-1}(\mu_1 - \mu_2), I)\right) \leq \frac{||T^{-1}(\mu_1 - \mu_2)||}{\sqrt{2\pi}}.$$

To bound the other component, view $\mathcal{N}(0, T^{-1}\Sigma_2 T'^{-1})$ as the joint distribution of m independent univariate Gaussians, where the ith distribution is $\mathcal{N}(0, \lambda_i)$, with λ_i the ith eigenvalue of $T^{-1}\Sigma_2 T'^{-1}$, and use facts Facts 26 and 28, to yield the claimed result. □

Proposition 30. *Given two m-dimensional Gaussians $G_1 = \mathcal{N}(\mu_1, \Sigma_1), G_2 = \mathcal{N}(\mu_2, \Sigma_2)$ such that for all $i, j \in [m]$, $|\Sigma_1(i,j) - \Sigma_2(i,j)| \leq \alpha$, and min$(eig(\Sigma_1)) > \lambda$,*

$$D_{tv}(G_1, G_2) \leq \frac{||\mu_1 - \mu_2||}{\sqrt{2\pi\lambda}} + \frac{m\alpha}{\sqrt{2\pi e}(\lambda - \alpha)}.$$

Proof. Let $\Sigma_1 = PDDP'$, where D is a diagonal matrix, and P is a unitary matrix. Note that the minimum entry on the diagonal of D is $\sqrt{\lambda}$. We now write $\Sigma_2 = \Sigma_1 + A$, for some symmetric matrix A whose entries are bounded in magnitude by α. By Fact 29, the contribution to $D_{tv}(G_1, G_2)$ from the discrepancy in the means is at most

$$\frac{||D^{-1}P'(\mu_1 - \mu_2)||}{\sqrt{2\pi}} \leq \frac{||\mu_1 - \mu_2||}{\sqrt{2\pi\lambda}}.$$

We now consider the contribution to $D_{tv}(G_1, G_2)$ from the discrepancy in the covariance matrices. We consider the eigenvalues of $D^{-1}P'\Sigma_2 PD^{-1} = I + D^{-1}P'APD^{-1}$. We have $\max_v \frac{||D^{-1}P'APD^{-1}v||}{||v||} \leq \frac{\alpha}{\lambda}$, and thus the maximum eigenvalue of $I + D^{-1}P'APD^{-1}$ is at most $1 + \frac{\alpha}{\lambda}$, and the minimum eigenvalue is at least $1 - \frac{\alpha}{\lambda}$; thus from Fact 29 we have

$$\begin{aligned} D_{tv}(G_1, G_2) &\leq \frac{||\mu_1 - \mu_2||}{\sqrt{2\pi\lambda}} + \frac{m\left(\frac{1}{1-\alpha/\lambda}\right) - 1}{\sqrt{2\pi e}} \\ &= \frac{||\mu_1 - \mu_2||}{\sqrt{2\pi\lambda}} + \frac{m\alpha}{\sqrt{2\pi e}(\lambda - \alpha)}. \end{aligned}$$

□