Estimating the Unseen: Improved Estimators for Entropy and Other Properties

GREGORY VALIANT, Stanford University
PAUL VALIANT, Brown University

We show that a class of statistical properties of distributions, which includes such practically relevant properties as entropy, the number of distinct elements, and distance metrics between pairs of distributions, can be estimated given a *sublinear* sized sample. Specifically, given a sample consisting of independent draws from any distribution over at most k distinct elements, these properties can be estimated accurately using a sample of size $O(k/\log k)$. For these estimation tasks, this performance is *optimal*, to constant factors. Complementing these theoretical results, we also demonstrate that our estimators perform exceptionally well, in practice, for a variety of estimation tasks, on a variety of natural distributions, for a wide range of parameters. The key step in our approach is to first use the sample to characterize the "unseen" portion of the distribution—effectively reconstructing this portion of the distribution as accurately as if one had a logarithmic factor larger sample. This goes beyond such tools as the Good-Turing frequency estimation scheme, which estimates the total probability mass of the unobserved portion of the distribution: We seek to estimate the *shape* of the unobserved portion of the distribution. This work can be seen as introducing a robust, general, and theoretically principled framework that, for many practical applications, essentially amplifies the sample size by a logarithmic factor; we expect that it may be fruitfully used as a component within larger machine learning and statistical analysis systems.

CCS Concepts: • **Theory of computation** → **Sketching and sampling**; **Sample complexity and generalization bounds**; • **Mathematics of computing** → *Information theory*;

Additional Key Words and Pharses: Statistical property estimation, unseen species, entropy estimation, distinct elements

ACM Reference format:
Gregory Valiant and Paul Valiant. 2017. Estimating the Unseen: Improved Estimators for Entropy and Other Properties. *J. ACM* 64, 6, Article 37 (October 2017), 41 pages.
https://doi.org/10.1145/3125643

1 INTRODUCTION

What can one infer about an unknown distribution based on a random sample? If the distribution in question is relatively "simple" in comparison to the sample size—for example, if our sample

Preliminary versions of portions of this work appeared at the ACM Symposium on Theory of Computing (STOC), 2011 [40], and at Neural Information Processing Systems (NIPS), 2013 [42].
Gregory's contributions are supported by NSF CAREER Award CCF-1351108 and a Sloan Research Fellowship. Paul's contributions are supported by a Sloan Research Fellowship and NSF grant IIS-1562657.
Authors' addresses: G. Valiant, Computer Science Department, Stanford University, 353 Serra Mall, Stanford, CA 94305; email: gvaliant@cs.stanford.edu; P. Valiant, Department of Computer Science, 115 Waterman St, Providence, RI 02912; email: pvaliant@cs.brown.edu.

https://doi.org/10.1145/3125643

consists of 1,000 independent draws from a distribution supported on 100 domain elements—then the empirical distribution given by the sample will likely be an accurate representation of the true distribution. If, on the other hand, we are given a relatively small sample in relation to the size and complexity of the distribution—for example, a sample of size 100 drawn from a distribution supported on 1,000 domain elements—then the empirical distribution may be a poor approximation of the true distribution. In this case, can one still extract accurate estimates of various properties of the true distribution?

Many real-world machine-learning and data analysis tasks face this challenge; indeed there are many large datasets where the data only represent a tiny fraction of an underlying distribution we hope to understand. This challenge of inferring properties of a distribution given a "too small" sample is encountered in a variety of settings, including text data (typically, no matter how large the corpus, around 30% of the observed vocabulary only occurs once), customer data (many customers or website users are only seen a small number of times), the analysis of neural spike trains [35], and the study of genetic mutations across a population.[1] Additionally, many database management tasks employ sampling techniques to optimize query execution; improved estimators would allow for either smaller sample sizes or increased accuracy, leading to improved efficiency of the database system (see, e.g., References [21, 30]).

We introduce a general and robust approach for using a sample to characterize the "unseen" portion of the distribution. Without any *a priori* assumptions about the distribution, one cannot know what the unseen domain elements are. Nevertheless, one can still hope to estimate the "shape" or *histogram* of the unseen portion of the distribution—essentially, we estimate how many unseen domain elements occur in various probability ranges. Given such a reconstruction, one can then use it to estimate any property/functional of the distribution that only depends on the shape/histogram; such properties are termed *symmetric* and include entropy and support size. In light of the long history of work on estimating entropy by the neuroscience, statistics, computer science, and information theory communities, it is compelling that our approach (which is agnostic to the property in question) outperforms these entropy-specific estimators (see Section 3).

Additionally, we extend this intuition to develop estimators for properties of pairs of distributions, the most important of which are the *distance metrics*. We demonstrate that our approach can accurately estimate the total variational distance (also known as *statistical distance* or ℓ_1 distance) between distributions using small samples. To illustrate the challenge of estimating variational distance (between distributions over discrete domains) given small samples, consider drawing two samples, each consisting of 1,000 draws from a uniform distribution over 10,000 distinct elements. Each sample can contain at most 10% of the domain elements, and their intersection will likely contain only 1% of the domain elements; yet from this, one would like to conclude that these two samples must have been drawn from nearly identical distributions.

For clarity, we summarize the performance guarantees of our approach in terms of the following three concrete and practically relevant questions, each defined with respect to an arbitrarily small constant error parameter $\epsilon > 0$:

- **Distinct Elements:** Given n buckets, each of which contains one object that is not necessarily distinct from those in the other buckets, how many buckets must one inspect to estimate the total number of distinct objects to within $\pm\epsilon n$, with high probability?
- **Entropy Estimation:** Given a sample obtained by taking independent draws from a distribution, p, of support size at most k, how large does the sample need to be to estimate

[1]For example, three different 2012 studies found that rare genetic mutations are especially abundant in humans and observed that better statistical tools are needed to characterize this "rare events" regime. A better understanding of these distributions of rare mutations would shed light on our evolutionary process and selective pressures [25, 29, 38].

the entropy of the distribution, $H(p) := -\sum_{x:p(x)>0} p(x) \log p(x)$, to within $\pm\epsilon$, with high probability?

- **Distance:** Given two samples obtained by taking independent draws from two distributions, p_1, p_2 of support size at most k, how large do the samples need to be to estimate the total variation distance between the distributions (also referred to as ℓ_1 distance or "statistical distance"), $D_{tv}(p_1, p_2) = \frac{1}{2} \sum_{x:p_1(x)+p_2(x)>0} |p_1(x) - p_2(x)|$, to within $\pm\epsilon$, with high probability?

We show that our approach performs the above three estimation tasks when given a sample (or two samples in the case of distance estimation) of size $n = O(\frac{k}{\log k})$, where the constant is dependent on the error parameter ϵ. This performance is information theoretically optimal to constant factors, as shown in Reference [39]. Prior to this work, no explicit estimators were known to solve any of these problems using samples of size $o(k)$, even for $\epsilon = 0.49$. See Section 1.4 for formal statements of our more general result on recovering a representation of the distribution, from which the estimation results follow immediately.

1.1 Previous Work: Estimating Distributions and Estimating Properties

There is a long line of work on inferring information about the unseen portion of a distribution, beginning with independent contributions from both R. A. Fisher and Alan Turing during the 1940s. Fisher was presented with data on butterflies collected over a 2-year expedition in Malaysia and sought to estimate the number of *new* species that would be discovered if a second 2-year expedition were conducted [17]. (His answer was "≈75.") This question was later revisited by I. J. Good and Toulmin [18], who offered a nonparametric alternative to Fisher's parametric model. At nearly the same time, as part of the British WWII effort to understand the statistics of the German enigma ciphers, Turing and Good were working on the related problem of estimating the total probability mass accounted for by the unseen portion of a distribution [19, 37]. This resulted in the Good-Turing frequency estimation scheme, which continues to be employed, analyzed, and extended (see, e.g., References [26, 32, 33, 46, 47]).

More recently, in similar spirit to this work, Orlitsky et al. [2, 31] posed the following natural question: Given a sample, what distribution maximizes the likelihood of seeing the observed species frequencies, that is, the number of species observed once, twice, and so on? (What Orlitsky et al. term the *pattern* of a sample, we call the *fingerprint*, as in Definition 1.1.) Orlitsky et al. show that such likelihood maximizing distributions can be found in some specific settings, though the problem of finding or approximating such distributions for typical patterns/fingerprints may be difficult. Recently, Acharya *et al.* showed that this maximum likelihood approach can be used to yield a near-optimal algorithm for deciding whether two samples originated from *identical* distributions versus distributions that have large distance [1].

In contrast to this approach of trying to estimate the "shape/histogram" of a distribution, there has been nearly a century of work proposing and analyzing estimators for particular properties (functionals) of distributions. A large portion of this literature focuses on analyzing the asymptotic consistency and distribution of natural estimators, such as the "plug-in" estimator or variants thereof (e.g., References [3, 5]. In Section 3, we describe several standard, and some recent, estimators for entropy, though we refer the reader to Reference [35] for a thorough treatment. There is also a large literature on the "unseen species" problem and the closely related "distinct elements" problem, including the efforts of Efron and Thisted to estimate the total number of words that Shakespeare knew (though might not have used in his extant works) [16]. Much of this work is based heavily on heuristic arguments or strong assumptions on the true distribution from which the sample is drawn and thus lies beyond the scope of our work; we refer the reader

to Reference [12] and to Reference [11] for several hundred references. We end Section 3 by demonstrating that our approach can accurately estimate the total number of distinct words that appear in *Hamlet* based on a short contiguous passage from the text.

Since the early 2000s, the theoretical computer science community has spent significant effort developing estimators and establishing worst-case information-theoretic lower bounds on the sample size required for various distribution estimation tasks, including entropy and support size (e.g., [4, 6–10, 14, 20, 44]). In contrast to the more traditional analysis of asymptotic rates of convergence for various estimators, this body of work aims to provide tight bounds on the sample size required to ensure that, with high probability over the randomness of the sampling, a desired error is achieved.

1.2 Subsequent Work

Subsequent to the initial dissemination of the preliminary versions of this work, there have been several relevant followup works. The approach of this work—namely to use the sample to recover a representation of the true distribution and then return the desired property value of the recovered distribution—is quite different from the more typical approach towards property estimation. The vast majority of estimators for entropy, for example, are *linear* functions of the summary statistics of the sample, $\mathcal{F}_1, \mathcal{F}_2, \ldots$, where $\mathcal{F}_i$ denotes the number of domain elements that occur exactly i times in the sample. For example, the plug-in estimator is the linear estimator $\sum_i c_i \mathcal{F}_i$ for $c_i = -\frac{i}{n} \log \frac{i}{n}$ and the "best-upper bound" estimator of Paninski can be viewed as an effort to heuristically find the "best" coefficients c_i [35]. The work of this current article prompted the question of whether there exist near optimal linear estimators or whether the more powerful computation involved in the estimators of this work are necessary to achieving constant-factor optimal entropy estimation. In Reference [41], we showed that, for a broad class of properties (functionals) of distributions, there are constant factor optimal linear estimators. Similar results were also independently obtained more recently [23, 48].

Despite the comparable theoretical performance for entropy estimation of the approach of this work, and the subsequent linear estimators of References [23, 41, 48], the approach of this work seems to yield superior performance in practice, particularly in the "hard" regime in which the sample size is smaller than the true support size of the distribution. The proof approach of Reference [41] provides some explanation for this disparity in performance: The linear estimators of Reference [41] are (roughly) defined as duals (via linear programming duality) to the worst-case instances for which entropy estimation is hardest. In this sense, the linear estimators are catering to worst-case instances. In contrast, the approaches of this current work are not based on worst-case instances (though also achieve constant factor optimal minimax error rates) and hence might perform better on more typical "easy" instances. Additionally, it should be stressed that this approach also yields a histogram representation (i.e., an unlabeled representation) of the distribution from which the sample is drawn. Such a representation can be used to reveal many further aspects of the distribution, beyond estimating the value of a specific property.

The approach of this article—to recover the unlabeled histogram representation of the distribution—is also useful for tasks that depend on the labels of the domain elements. In Reference [43], the authors consider the problem of learning an arbitrary distribution over a discrete support and develop an "instance optimal" algorithm that de-noises the empirical distribution of a set of samples, whenever such a denoising is possible. Specifically, given n i.i.d. samples from any discrete distribution, the learning algorithm outputs a labeled vector whose expected ℓ_1 distance to the true distribution is at most a factor of $1 + o_n(1)$ worse than the minimal expected error that is achievable by *any* algorithm that takes as input the set of samples and is agnostic among relabelings of the domain. One of the components of that algorithm is a slight strengthening of

the histogram recovery algorithm of this work that removes the requirement that the samples be drawn from a distribution of bounded support size. Among other consequences, that result implies the following clean result: given n i.i.d. samples from any discrete distribution, one can accurately estimate the number of new domain elements that will be observed in a second set of samples of size up to $\Theta(n \log n)$. This result was independently and simultaneously obtained by Reference [34] via a linear estimator.

On the practical side, an adaptation of this approach of recovering the histogram was applied to a large dataset of 60,000 human genomes to understand the relative frequencies of *unobserved* genetic mutations of various types (synonymous, missense, and medically relevant mutations such as loss-of-function mutations) [50]. Additionally, this recovered histogram allowed us to quantify the value of sequencing additional genomes by making accurate predictions regarding the number of new mutations of these various types that would likely be observed in larger sequenced cohorts.

1.3 Definitions and Examples

We begin by defining the *fingerprint* of a sample, which essentially removes all the label-information from the sample. For the remainder of this article, we will work with the fingerprint of a sample rather than the with the sample itself.

Definition 1.1. Given a sample $X = (x_1, \ldots, x_n)$, the associated *fingerprint*, $\mathcal{F} = (\mathcal{F}_1, \mathcal{F}_2, \ldots)$, is the "histogram of the histogram" of the sample. Formally, $\mathcal{F}$ is the vector whose ith component, $\mathcal{F}_i$, is the number of elements in the domain that occur exactly i times in sample X.

For estimating entropy, or any other property whose value is invariant to relabeling the distribution support, the fingerprint of a sample contains all the relevant information (see Reference [9], for a formal proof of this fact). We note that in some of the literature, the fingerprint is alternately termed the *pattern*, *histogram*, *histogram of the histogram*, or *collision statistics* of the sample.

In analogy with the fingerprint of a sample, we define the *histogram* of a distribution, a representation in which the labels of the (finite or countably infinite) domain have been removed.

Definition 1.2. The *histogram* of a distribution p, with a finite or countably infinite support, is a mapping $h_p : (0, 1] \to \mathbb{N} \cup \{0\}$, where $h_p(x)$ is equal to the number of domain elements that each occur in distribution p with probability x. Formally, $h_p(x) = |\{\alpha : p(\alpha) = x\}|$, where $p(\alpha)$ is the probability mass that distribution p assigns to domain element α. We will also allow for "generalized histograms" in which h_p does not necessarily take integral values.

Since $h(x)$ denotes the number of elements that have probability x, we have $\sum_{x:h(x)\neq 0} x \cdot h(x) = 1$, as the total probability mass of a distribution is 1.

Definition 1.3. Let $\mathcal{D}$ denote the set of all distributions and $\mathcal{D}^k$ denote the set of distributions over the domain $[k] = \{1, \ldots, k\}$.

Definition 1.4. A *symmetric* distribution property $\pi : \mathcal{D} \to \mathbb{R}$ is a function that depends only on the histogram of the distribution and hence is invariant to permuting the labels of the domain.

Both entropy and support size are symmetric distribution properties as follows:

- The *Shannon entropy* $H(p)$ of a distribution p is defined to be
$$H(p) := -\sum_{\alpha \in sup(p)} p(\alpha) \log_2 p(\alpha) = -\sum_{x:h_p(x)\neq 0} h_p(x) x \log_2 x.$$
- The *support size* is the number of domain elements that occur with positive probability:
$$|sup(p)| := |\{\alpha : p(\alpha) > 0\}| = \sum_{x:h_p(x)\neq 0} h_p(x).$$

We provide an example to illustrate the above definitions:

Example 1.5. Consider a sequence of animals, obtained as a sample from the distribution of animals on a certain island, $X = (mouse, mouse, bird, cat, mouse, bird, bird, mouse, dog, mouse)$. We have $\mathcal{F} = (2, 0, 1, 0, 1)$, indicating that two species occurred exactly once (cat and dog), one species occurred exactly 3 times (bird), and one species occurred exactly 5 times (mouse).

Consider the following distribution of animals:

$$Pr(mouse) = 1/2, \quad Pr(bird) = 1/4, \quad Pr(cat) = Pr(dog) = Pr(bear) = Pr(wolf) = 1/16.$$

The associated *histogram* of this distribution is $h : (0, 1] \to \mathbb{Z}$ defined by $h(1/16) = 4$, $h(1/4) = 1$, $h(1/2) = 1$, and for all $x \notin \{1/16, 1/4, 1/2\}$, $h(x) = 0$.

Our main theorem will apply to any symmetric distribution property that is sufficiently continuous with respect to changes in the distribution. To formalize this notion, we now define what it means for two distributions to be "close." In particular, distributions that are close under the following metric will have similar histograms and hence similar values of entropy, support size, and other symmetric properties.

Definition 1.6. For two distributions p_1, p_2 with respective histograms h_1, h_2, we define the *relative earthmover distance* between them, $R(p_1, p_2) := R(h_1, h_2)$, as the minimum over all schemes of moving the probability mass of the first histogram to yield the second histogram of the cost of moving that mass, where the per-unit mass cost of moving mass from probability x to y is $|\log(x/y)|$. Formally, for $x, y \in (0, 1]$, the cost of moving $x \cdot h(x)$ units of mass from probability x to y is $x \cdot h(x)|\log \frac{x}{y}|$.

One can also define the relative earthmover distance via the following dual formulation (given by the Kantorovich-Rubinstein theorem [24], though it can be intuitively seen as exactly what one would expect from linear programming duality):

$$R(h_1, h_2) = \sup_{f \in \mathcal{R}} \sum_{x : h_1(x) + h_2(x) \neq 0} f(x) \cdot x \left(h_1(x) - h_2(x)\right),$$

where $\mathcal{R}$ is the set of differentiable functions $f : (0, 1] \to \mathbb{R}$, s.t. $|\frac{d}{dx} f(x)| \leq \frac{1}{x}$.

We provide a clarifying example of the above definition:

Example 1.7. Let $p_1 = Unif[m], p_2 = Unif[\ell]$ be the uniform distributions over m and ℓ distinct elements, respectively. $R(p_1, p_2) = |\log m - \log \ell|$, since we must take all the probability mass at probability $x = 1/m$ in the histogram corresponding to p_1, and move it to probability $y = 1/\ell$, at a per-unit mass cost of $|\log \frac{m}{\ell}| = |\log m - \log \ell|$.

As mentioned above, relative earthmover distance is the metric through which we state our main results describing how, given a sample from a distribution p, we can reconstruct a distribution $\hat{p}$ that approximates p (namely, $R(p, \hat{p})$ is small). To motivate our choice of relative earthmover distance, we note that relative earthmover distance is independent of element relabeling and essentially declares two distributions close if the multiset of probabilities with which the elements of one distribution occur can be approximately matched to the multiset of the second distribution—with a logarithmically increasing cost of matching up more distant probabilities. Example 1.7 above shows that the logarithmic weighting means that the relative earthmover distance between two *uniform* distributions is exactly their difference in entropy. Further, relative earthmover distance is similar to a label-invariant modification of the ℓ_1 distance between two distributions, where the labels of the supports of the distributions are permuted to minimize the resulting ℓ_1 distance. The following fact, whose elementary proof is given in Reference [43] characterizes this relationship:

Fact 1 (Fact 1 in Reference [43]). *Given two distributions p_1, p_2, supported on the integers, there exists a relabeling π of the support of p_2 such that*

$$\frac{1}{2}\sum_i |p_1(i) - p_2(\pi(i))| \le R(p_1, p_2).$$

Minor variants of the $|\log \frac{x}{y}|$ weighting in Definition 1.6 would not significantly affect the analysis, though the logarithm is a simple and natural choice.

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

Definition 1.8. A symmetric distribution property π is (ϵ, δ)-*continuous* if for all distributions p_1, p_2 with respective histograms h_1, h_2 satisfying $R(h_1, h_2) \le \delta$ it follows that $|\pi(p_1) - \pi(p_2)| \le \epsilon$.

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

Fact 2. *For a distribution $p \in \mathcal{D}^k$, and $\delta > 0$*

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

As we will see in Example 1.10 below, the fingerprint of a sample is intimately related to the Binomial distribution; the theoretical analysis will be greatly simplified by reasoning about the related Poisson distribution, which we now define:

Definition 1.9. We denote the Poisson distribution of expectation λ as $Poi(\lambda)$ and write $poi(\lambda, j) := \frac{e^{-\lambda}\lambda^j}{j!}$ to denote the probability that a random variable with distribution $Poi(\lambda)$ takes value j.

Example 1.10. Let D be the uniform distribution with support size 1,000. Then $h_D(1/1{,}000) = 1{,}000$, and for all $x \neq 1/1{,}000$, $h_D(x) = 0$. Let X be a sample consisting of 500 independent draws from D. Each element of the domain, in expectation, will occur 1/2 times in X, and thus the number of occurrences of each domain element in the sample X will be roughly distributed as $Poi(1/2)$—of course, the exact distribution will be $Binomial(500, 1/1{,}000)$. By linearity of expectation, the expected fingerprint satisfies $E[\mathcal{F}_i] \approx 1000 \cdot poi(1/2, i)$. Thus we expect to see roughly 303 elements once, 76 elements twice, 13 elements 3 times, and so on, and, in expectation, 607 domain elements will not be seen at all.

1.4 Statement of Main Theorems

Our main theorem guarantees the performance of a novel algorithm for approximating an arbitrary unknown discrete distribution given a sample whose size is *sublinear* in the support size of the distribution. The performance is described in terms of the *relative earthmover* distance metric R (Definition 1.6), which is a distance metric between distributions that captures the similarity of distribution *up to relabeling the supports* and has the property that two distributions that are close in relative earthmover distance have similar values of entropy, support size, and other well-behaved symmetric properties.

Theorem 1.11. *There exist absolute positive constants α, β such that for any $c > 0$ and any $k > k_c$ (where k_c is a constant dependent on c), given a sample of size $n = c\frac{k}{\log k}$ consisting of independent*

draws from a distribution $p \in \mathcal{D}^k$, with probability at least $1 - e^{-k^\alpha}$ over the randomness in the selection of the sample, our algorithm returns a distribution $\hat{p}$ such that

$$R(p, \hat{p}) \leq \frac{\beta}{\sqrt{c}}.$$

In other words, for any desired accuracy, $\epsilon > 0$, up to constant factors, a sample of size $\frac{k}{\epsilon^2 \log k}$ is sufficient to estimate the histogram of any distribution supported on at most k elements. While our results are stated in terms of the error ϵ and an upper bound on the support size, k, the algorithm does not depend on either of these parameters and is given only the sample as input; hence both Theorem 1.11 and its corollaries below can naturally be interpreted as bounds on convergence rates.

In subsequent work [43], essentially the same algorithm was shown to obtain an analogous result in the more general setting where there is no bound (k) on the support size of the distribution in question. Instead, the theorem (Theorem 2 in References [43]) shows that given n independent draws, the algorithm will accurately recover the portion of the histogram corresponding to the probability values larger than $\Omega(n/\log n)$.

For estimating entropy and the support size, Theorem 1.11 together with Fact 2 yields the following:

Corollary 1.12. *There exist absolute positive constants α, γ such that for any positive $\epsilon < 1$, there exists k_ϵ such that for any $k > k_\epsilon$, given a sample of size at least $\frac{\gamma}{\epsilon^2}\frac{k}{\log k}$ drawn from any $p \in \mathcal{D}^k$, our estimator will output a pair of real numbers $(\hat{H}, \hat{S})$ such that with probability at least $1 - e^{-k^\alpha}$*

- *$\hat{H}$ is within ϵ of the entropy of p, and*
- *$\hat{S}$ is within $k\epsilon$ of the support size of p, provided none of the probabilities in p lie in $(0, \frac{1}{k})$.*

For the distinct elements problem, the above corollary implies that by randomly selecting (with replacement) $\frac{\gamma}{\epsilon^2}\frac{k}{\log k}$ buckets to inspect, our algorithm will return an estimate of the number of distinct elements accurate to within $\pm\epsilon k$, with probability of failure at most e^{-k^α}.

These estimators have the optimal dependence on k, up to constant factors. We show the following information theoretic lower bounds in Reference [39]:

Theorem. *There exists a constant c and integer k_0 such that for any $k \geq k_0$, no estimator has the property that, when given a sample of size $c\frac{k}{\log k}$ drawn from any $p \in \mathcal{D}^k$, it can estimate the entropy of p to within accuracy $\pm\frac{\log 2}{2}$ with probability of success at least 0.51. The analogous statement holds for estimating the support size to $\pm\frac{k}{4}$, for distributions $p \in \mathcal{D}^k$ such that for all i, $p(i) \notin (0, \frac{1}{k})$.*

Phrased differently, let S denote a sample of size n, with $S \underset{n}{\leftarrow} p$ denoting the process of assigning a sample of size n via independent draws from $p \in \mathcal{D}^k$, and let $\hat{H} : [k]^n \to \mathbb{R}$ denote an arbitrary estimator that maps a sample S to an estimate of the entropy of the distribution from which the sample was drawn. The above theorem states that there exists a constant c such that for $n = c\frac{k}{\log k}$,

$$\inf_{\hat{H}} \sup_{p \in \mathcal{D}^k} \Pr_{S \underset{n}{\leftarrow} p}\left[|\hat{H}(S) - H(p)| > \frac{\log 2}{2}\right] > 0.49,$$

where the infimum is taken over all possible estimators.

Our entire estimation framework generalizes to estimating properties of pairs of distributions. As in the setting described above for properties of a single distribution, given a pair of samples drawn independently from two (possibly different) distributions, we can characterize the performance of our estimators in terms of returning a representation of the pair of distributions. For

clarity, we state our performance guarantees for estimating total variation distance (ℓ_1 distance); see Theorem 5.6 in Section 5 for the more general formulation.

THEOREM 1.13. *There exist absolute positive constants α, γ such that for any positive $\epsilon < 1$, there exists k_ϵ such that for any $k > k_\epsilon$, given a pair of samples of size $n = \frac{\gamma}{\epsilon^2}\frac{k}{\log k}$ drawn independently, respectively, from $p, q \in \mathcal{D}^k$, our estimator will output a number $\hat{d}$ such that with probability at least $1 - e^{-k^\alpha}$*

$$|\hat{d} - D_{tv}(p, q)| \leq \epsilon,$$

where $D_{tv}(p, q) = \sum_i \frac{1}{2}|p(i) - q(i)|$ is half the ℓ_1 distance between distributions p and q.

In Reference [39], we show that the above performance is optimal in its dependence on k, up to constant factors:

THEOREM. *There exist a constant c and integer k_0 such that for any $k > k_0$, no estimator, when given a pair of samples of size $c\frac{k}{\log k}$ drawn from any $p, q \in \mathcal{D}^k$ can estimate $D_{tv}(p, q)$ to within accuracy ± 0.49 with probability of success at least 0.51.*

1.5 Outline

In Section 2, we motivate and describe our approach of posing the inverse problem "given a sample, what is the histogram of the distribution from which it was drawn" as an explicit optimization problem. We show, perhaps surprisingly, that we can capture the essential features of this problem via a *linear program*—rendering it both computationally tractable, as well as amenable to a rich set of analysis tools. Furthermore, our general linear program formulation allows for considerable flexibility in tailoring both the objective function and constraints for specific estimation tasks.

In Section 3, we illustrate the performance and robustness of our approach for several estimation tasks on both synthetic, and real data. Section 4 summarizes the structure and main components of the proof of Theorem 1.11. Section 5 describes how to extend our approach to the two distribution setting, which yields our results for estimating the total variation distance between pairs of distributions, Theorem 1.13. Section 6 gives a self-contained proof of Theorem 1.11. The proof of our two-distribution analog of Theorem 1.13 closely parallels the proof in the one distribution setting, and we defer this proof to Appendix A. Appendix B contains some additional empirical results demonstrating that the performance of our approach is robust to different implementation decisions and choices of parameters. Appendix C provides a Matlab implementation of our approach, which was used to produce our empirical results.

2 ESTIMATING THE UNSEEN

Given the fingerprint $\mathcal{F}$ of a sample of size n, drawn from a distribution with histogram h, our high-level approach is to find a histogram h' that has the property that if one were to take n independent draws from a distribution with histogram h', the fingerprint of the resulting sample would be similar to the observed fingerprint $\mathcal{F}$. The hope is then that h and h' will be similar and, in particular, have similar entropies, support sizes, and so on.

As an illustration of this approach, suppose we are given a sample of size $n = 500$, with fingerprint $\mathcal{F} = (301, 78, 13, 1, 0, 0, \ldots)$; recalling Example 1.10, we recognize that $\mathcal{F}$ is very similar to the expected fingerprint that we would obtain if the sample had been drawn from the uniform distribution over support $1,000$. Although the sample only contains 391 unique domain elements, one might be inclined to conclude that the true distribution is close to the uniform distribution over 1,000 elements, and the entropy is roughly $H(Unif(1,000)) = \log_2(1,000)$, for example. Our results show that this intuition is justified, and rigorously quantify the extent to which such reasoning may be applied.

In general, how does one obtain a "plausible" histogram from a fingerprint in a principled fashion? We must start by understanding how to obtain a plausible fingerprint from a histogram.

Given a distribution D, and some domain element α occurring with probability $x = D(\alpha)$, the probability that it will be drawn exactly i times in n independent draws from D is $Pr[Binomial(n, x) = i] \approx poi(nx, i)$. By linearity of expectation, the expected ith fingerprint entry will roughly satisfy

$$E[\mathcal{F}_i] \approx \sum_{x:h_D(x)\neq 0} h(x)poi(nx, i). \tag{1}$$

This mapping between histograms and expected fingerprints is linear in the histogram, with coefficients given by the Poisson probabilities. Additionally, it is not hard to show that $Var[\mathcal{F}_i] \leq E[\mathcal{F}_i]$, and thus the fingerprint is tightly concentrated about its expected value. This motivates a "first moment" approach. We will, roughly, invert the linear map from histograms to expected fingerprint entries, to yield a map from observed fingerprints, to plausible histograms h'.

There is one additional component of our approach. For many fingerprints, there will be a large space of equally plausible histograms. To illustrate, suppose we obtain fingerprint $\mathcal{F} = (10, 0, 0, 0, \ldots)$ and consider the two histograms given by the uniform distributions with respective support sizes 10,000, and 100,000. Given either distribution, the probability of obtaining the observed fingerprint from a set of 10 samples is $> .99$, yet these distributions are quite different and have very different entropy values and support sizes. They are both very plausible—which distribution should we return?

To resolve this issue in a principled fashion, we strengthen our initial goal of "returning a histogram that could have plausibly generated the observed fingerprint": We instead return the *simplest* histogram that could have plausibly generated the observed fingerprint. Recall the example above, where we observed only 10 distinct elements, but to explain the data, we could either infer an additional 9,990 unseen elements or an additional 99,990. In this sense, inferring "only" 9,990 additional unseen elements is the simplest explanation that fits the data, in the spirit of Occam's razor.[2]

2.1 The Algorithm

We pose this problem of finding the simplest plausible histogram as a pair of linear programs. The first linear program will return a histogram h' that minimizes the distance between its expected fingerprint and the observed fingerprint, where we penalize the discrepancy between $\mathcal{F}_i$ and $E[\mathcal{F}_i^{h'}]$ in proportion to the inverse of the standard deviation of $\mathcal{F}_i$, which we estimate as $1/\sqrt{1 + \mathcal{F}_i}$, since Poisson distributions have variance equal to their expectation. The constraint that h' corresponds to a histogram simply means that the total probability mass is 1, and all probability values are nonnegative. The second linear program will then find the histogram h'' of minimal support size, subject to the constraint that the distance between its expected fingerprint, and the observed fingerprint, is not much worse than that of the histogram found by the first linear program.

To make the linear programs finite, we consider a fine mesh of values $x_1, \ldots, x_\ell \in (0, 1]$ that between them discretely approximate the potential support of the histogram. The variables of the linear program, $h'_1, \ldots, h'_\ell$ will correspond to the histogram values at these mesh points, with variable h'_i representing the number of domain elements that occur with probability x_i, namely $h'(x_i)$.

[2]The practical performance seems virtually unchanged if one returns the "plausible" histogram of minimal entropy instead of minimal support size (see Appendix B).

A minor complicating issue is that this approach is designed for the challenging "rare events" regime, where there are many domain elements each seen only a handful of times. By contrast if there is a domain element that occurs very frequently, say, with probability $1/2$, then the number of times it occurs will be concentrated about its expectation of $n/2$ (and the trivial empirical estimate will be accurate), though fingerprint $\mathcal{F}_{n/2}$ will not be concentrated about its expectation, as it will take an integer value of 0, 1, or 2. Hence we will split the fingerprint into the "easy" and "hard" portions and use the empirical estimator for the easy portion, and our linear programming approach for the hard portion. The full algorithm is below (see our websites or Appendix C for Matlab code).[3]

Algorithm 1. *Estimate Unseen*

Input: Fingerprint $\mathcal{F} = \mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_m$, derived from a sample of size n,
vector $x = x_1, \ldots, x_\ell$ with $0 < x_i \leq 1$, and error parameter $\delta > 0$.
Output: List of pairs $(y_1, h'_{y_1}), (y_2, h'_{y_2}), \ldots,$ with $y_i \in (0, 1]$, and $h'_{y_i} \geq 0$.

- Initialize the output list of pairs to be empty, and initialize a vector $\mathcal{F}'$ to be equal to $\mathcal{F}$.
- For $i = 1$ to n,
  - If $\sum_{j \in \{i - \lceil \sqrt{i} \rceil, \ldots, i + \lceil \sqrt{i} \rceil\}} \mathcal{F}_j \leq 2\sqrt{i}$ (i.e., if the fingerprint is "sparse" at index i)
    Set $\mathcal{F}'_i = 0$, and append the pair $(i/n, \mathcal{F}_i)$ to the output list.[4]
- Let v_{opt} be the objective function value returned by running Linear Program 1 on input $\mathcal{F}', x$.
- Let h be the histogram returned by running Linear Program 2 on input $\mathcal{F}', x, v_{opt}, \delta$.
- For all i s.t. $h_i > 0$, append the pair (x_i, h_i) to the output list.

Linear Program 1. Find Plausible Histogram

Input: Fingerprint $\mathcal{F} = \mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_m$, derived from a sample of size n,
vector $x = x_1, \ldots, x_\ell$ consisting of a fine mesh of points in the interval $(0, 1]$.
Output: vector $h' = h'_1, \ldots, h'_\ell$, and objective value $v_{opt} \in \mathbb{R}$.

Let $h'_1, \ldots, h'_\ell$ and v_{opt} be, respectively, the solution assignment, and corresponding objective function value of the solution of the following linear program, with variables $h'_1, \ldots, h'_\ell$:

$$\text{Minimize: } \sum_{i=1}^{m} \frac{1}{\sqrt{1+\mathcal{F}_i}} \left| \mathcal{F}_i - \sum_{j=1}^{\ell} h'_j \cdot poi(nx_j, i) \right|$$

$$\text{Subject to: } \sum_{j=1}^{\ell} x_j h'_j = \sum_i \mathcal{F}_i / n, \text{ and } \forall j, \quad h'_j \geq 0.$$

Linear Program 2. Find Simplest Plausible Histogram

Input: Fingerprint $\mathcal{F} = \mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_m$, derived from a sample of size n,
vector $x = x_1, \ldots, x_\ell$ consisting of a fine mesh of points in the interval $(0, 1]$,
optimal objective function value v_{opt} from Linear Program 1, and error parameter $\delta > 0$.
Output: vector $h' = h'_1, \ldots, h'_\ell$.
Let $h'_1, \ldots, h'_\ell$ be the solution assignment of the following linear program, with variables $h'_1, \ldots, h'_\ell$:

$$\text{Minimize: } \sum_{j=1}^{\ell} h'_j \quad \text{Subject to: } \sum_{i=1}^{m} \frac{1}{\sqrt{1+\mathcal{F}_i}} \left| \mathcal{F}_i - \sum_{j=1}^{\ell} h'_j \cdot poi(nx_j, i) \right| \leq v_{opt} + \delta,$$
$$\sum_{j=1}^{\ell} x_j h'_j = \sum_i \mathcal{F}_i / n, \text{ and } \forall j, \quad h'_j \geq 0.$$

[3] A unified approach is possible, using an earthmover distance metric as part of the linear programs to cleanly circumvent these issues. Such an approach yields comparable theoretical performance guarantees, though the experimental results this approach yielded were indistinguishable from those presented here and thus do not seem to justify the additional computational expense.

[4] This scheme for partitioning the fingerprint into the "easy" regime (on which we use the empirical distribution) and "hard" regime (for which we employ the linear programs) is what we recommend in practice and produced the experimental results of Section 3. For simplicity of exposition, we prove Theorem 1.11 for the slight variant with a fixed transition point s—that is, the linear programs are run on $\{\mathcal{F}_i : i \leq s\}$.

The following restatement of our main theorem characterizes the worst-case performance guarantees of the above algorithm, establishing the constant-factor optimal guarantees for entropy estimation and the distinct elements problem and implying bounds on the error of estimating any symmetric distribution property that is Lipschitz continuous with respect to the relative earth-mover distance metric. While the theorem characterizes the performance of Algorithm 1 in terms of the support size, k, we stress that the algorithm does not depend on k and hence can be applied in settings where k is unknown.

THEOREM 1.11. *There exist absolute positive constants α, β and an assignment of the parameters $\delta = \delta(n)$ and $x = x_1, \ldots, x_\ell$ in Algorithm 1 such that for any $c > 0$ and k sufficiently large, given a sample of size $n = c\frac{k}{\log k}$ consisting of independent draws from a distribution $p \in \mathcal{D}^k$, with probability at least $1 - e^{-k^\alpha}$ over the randomness in the selection of the sample, Algorithm 1 returns a distribution $\hat{p}$ such that*

$$R(p, \hat{p}) \leq \frac{\beta}{\sqrt{c}}.$$

The proof of Theorem 1.11 is rather technical, with the cornerstone being the construction of an explicit earthmoving scheme via a Chebyshev polynomial construction. We give a detailed overview of the proof in Section 4 and give the complete proof in Section 6.

3 EMPIRICAL RESULTS

In this section, we demonstrate that Algorithm 1 performs well, in practice. We begin by briefly discussing the five entropy estimators to which we compare our estimator in Figure 1. The first three are standard and are, perhaps, the most commonly used estimators [35]. We then describe two more recently proposed estimators that have been shown to perform well in some practical settings [45].

The "naive" estimator: the entropy of the empirical distribution, namely, given a fingerprint $\mathcal{F}$ derived from a sample of size n, $H^{naive}(\mathcal{F}) := -\sum_i \mathcal{F}_i \frac{i}{n} |\log_2 \frac{i}{n}|$.

The Miller-Madow corrected estimator [27]: the naive estimator H^{naive} corrected to try to account for the second derivative of the logarithm function, namely $H^{MM}(\mathcal{F}) := H^{naive}(\mathcal{F}) + \frac{(\sum_i \mathcal{F}_i)-1}{2n}$, though we note that the numerator of the correction term is sometimes replaced by various related quantities, see Reference [36].

The jackknifed naive estimator [15, 49]: $H^{JK}(\mathcal{F}) := k \cdot H^{naive}(\mathcal{F}) - \frac{n-1}{n} \sum_{j=1}^{n} H^{naive}(\mathcal{F}^{-j})$, where $\mathcal{F}^{-j}$ is the fingerprint given by removing the contribution of the jth sample.

The coverage adjusted estimator (CAE) [13]: Chao and Shen proposed the CAE, which is specifically designed to apply to settings in which there is a significant component of the distribution that is unseen and was shown to perform well in practice in Reference [45].[5] Given a fingerprint $\mathcal{F}$ derived from a set of n samples, let $P_s := 1 - \mathcal{F}_1/n$ be the Good-Turing estimate of the probability mass of the "seen" portion of the distribution [19]. The CAE adjusts the empirical probabilities according to P_s and then applies the Horvitz-Thompson estimator for population totals [22] to take into account the probability that the elements were seen. This yields:

$$H^{CAE}(\mathcal{F}) := -\sum_i \mathcal{F}_i \frac{(i/n)P_s \log_2((i/n)P_s)}{1 - (1 - (i/n)P_s)^n}.$$

[5] One curious weakness of the CAE, is that its performance is exceptionally poor on some simple large instances. Given a sample of size n from a uniform distribution over n elements, it is not hard to show that the bias of the CAE is unbounded, growing proportionally to $\log n$. For comparison, even the naive estimator has error bounded by a constant in the limit as $n \to \infty$ in this setting. This bias of the CAE is easily observed in our experiments as the "hump" in the top row of Figure 1.

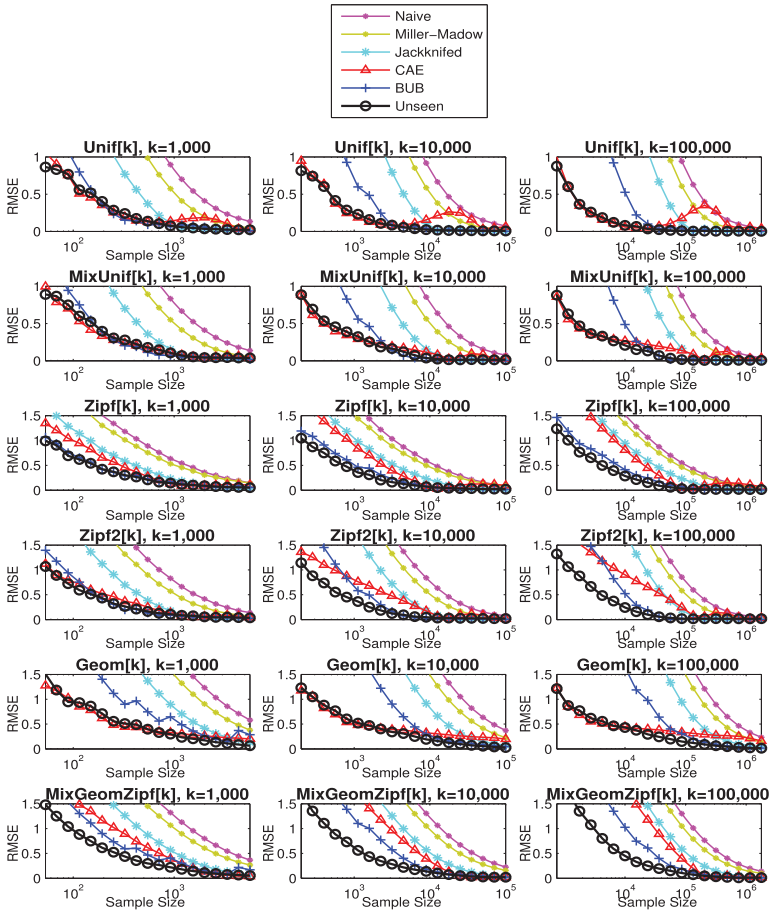

Fig. 1. Plots depicting the RMSE of each entropy estimator over 500 trials, plotted as a function of the sample size; note the logarithmic scaling of the x-axis. The samples are drawn from six classes of distributions: the uniform distribution, $Unif[k]$ that assigns probability $p_i = 1/k$ for $i = 1, 2, \ldots, k$; an even mixture of $Unif[\frac{k}{5}]$ and $Unif[\frac{4k}{5}]$, which assigns probability $p_i = \frac{5}{2k}$ for $i = 1, \ldots, \frac{k}{5}$ and probability $p_i = \frac{5}{8k}$ for $i = \frac{k}{5} + 1, \ldots, k$; the Zipf distribution $Zipf[k]$ that assigns probability $p_i = \frac{1/i}{\sum_{j=1}^{k} 1/j}$ for $i = 1, 2, \ldots, k$ and is commonly used to model naturally occurring "power law" distributions, particularly in natural language processing; a modified Zipf distribution with power–law exponent 0.6, $Zipf2[k]$, that assigns probability $p_i = \frac{1/i^{0.6}}{\sum_{j=1}^{k} 1/j^{0.6}}$ for $i = 1, 2, \ldots, k$; the geometric distribution $Geom[k]$, which has infinite support and assigns probability $p_i = (1/k)(1 - 1/k)^i$, for $i = 1, 2 \ldots$; and, last, an even mixture of $Geom[k/2]$ and $Zipf[k/2]$. For each distribution, we considered three settings of the parameter k: $k = 1,000$ (left column), $k = 10,000$ (center column), and $k = 100,000$ (right column). In each plot, the sample size, n, ranges over the interval $[k^{0.6}, k^{1.25}]$. Appendix B contains additional empirical results showing that the performance of our estimator is extremely robust to varying the parameters of the algorithm and changing the specifics of the implementation of our high-level approach.

The *Best Upper Bound*estimator [35]: The final estimator to which we compare ours is the *Best Upper Bound* (BUB) estimator of Paninski. This estimator is obtained by searching for a minimax linear estimator, with respect to a certain error metric. The linear estimators of Reference [41] can be viewed as a variant of this estimator with provable performance bounds.[6] The BUB estimator requires, as input, an upper bound on the support size of the distribution from which the samples are drawn; if the bound provided is inaccurate, the performance degrades considerably, as was also remarked in Reference [45]. In our experiments, we used Paninski's implementation of the BUB estimator (publicly available on his website), with default parameters. For the distributions with finite support, we gave the true support size as input, and thus we are arguably comparing our estimator to the best-case performance of the BUB estimator.

Figure 1 compares the root-mean-squared error (RMSE) of these estimators with the estimator obtained by returning the entropy of the histogram returned by Algorithm 1, which we refer to as the *unseen estimator*. All experiments were run in Matlab, with the RMSE errors calculated based on 500 independent trials. The error parameter α in Algorithm 1 was set to be 0.5 for all trials, and the vector $x = x_1, x_2, \ldots$ used as the support of the returned histogram was chosen to be a coarse geometric mesh, with $x_1 = 1/n^2$, and $x_i = 1.1x_{i-1}$. The experimental results are essentially unchanged if the parameter α varied within the range $[0.25, 1]$, if x_1 is decreased, or if the mesh is made more fine (see Appendix B). Appendix C contains our Matlab implementation of Algorithm 1 (also available from our websites).

The *unseen* estimator performs far better than the three standard estimators, dominates the CAE estimator for larger sample sizes and on samples from the Zipf distributions and also dominates the BUB estimator, even for the uniform and Zipf distributions for which the BUB estimator received the true support sizes as input. The consistently good performance of the *unseen* estimator over all the classes of distributions is especially startling given that Algorithm 1 is designed to compute a representation of the distribution rather than specifically tailored to estimate entropy.

3.1 Estimating ℓ_1 Distance and Number of Words in Hamlet

The other two properties that we consider do not have such widely accepted estimators as entropy, and thus our evaluation of the unseen estimator will be more qualitative. We include these two examples here, because they are of a substantially different flavor from entropy estimation, and highlight the flexibility of our approach.

Figure 2 shows the results of estimating the total variation distance (ℓ_1 distance). Because total variation distance is a property of two distributions instead of one, fingerprints and histograms are two-dimensional objects in this setting (see Definitions 5.1 and 5.2 in Section 5), and Algorithm 1 and the linear programs are extended accordingly, replacing single indices by pairs of indices, and Poisson coefficients by corresponding products of Poisson coefficients.

Finally, in contrast to the synthetic tests above, we also evaluated our estimator on a real-data problem which may be seen as emblematic of the challenges in a wide gamut of natural language processing problems: *Given a (contiguous) fragment of Shakespeare's* Hamlet*, estimate the number of distinct words in the whole play*. We use this example to showcase the flexibility of our linear programming approach—our estimator can be customized to particular domains in powerful and principled ways by adding or modifying the constraints of the linear program. To estimate the histogram of word frequencies in *Hamlet*, we note that the play is of length $\approx$25,000, and thus the minimum probability with which any word can occur is $\frac{1}{25,000}$. Thus in contrast to our previous approach of using Linear Program 2 to bound the support of the returned histogram, we instead simply modify the input vector x of Linear Program 1 to contain only probability values $\geq \frac{1}{25,000}$,

[6] We also implemented the linear estimators of Reference [41], though found that the BUB estimator performed better.

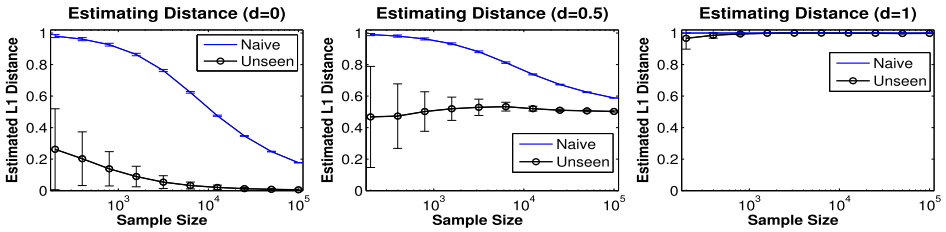

Fig. 2. Plots depicting the estimated total variation distance (ℓ_1 distance) between two uniform distributions on $k = 10,000$ points, in three cases: the two distributions are identical (left plot, $d = 0$), the supports overlap on *half* their domain elements (center plot, $d = 0.5$), and the distributions have disjoint supports (right plot, $d = 1$). The estimate of the distance is plotted along with error bars at plus and minus one standard deviation; our results are compared with those for the naive estimator (the distance between the empirical distributions). The *unseen* estimator can be seen to reliably distinguish between the $d = 0$, $d = \frac{1}{2}$, and $d = 1$ cases even for samples as small as several hundred.

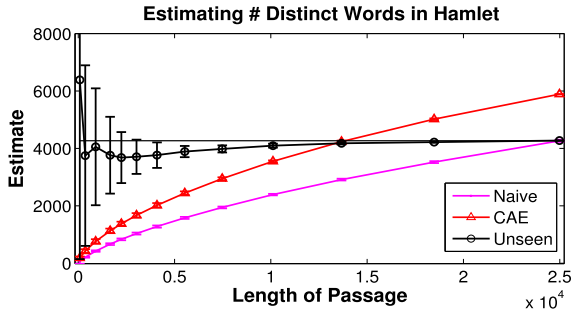

Fig. 3. Estimates of the total number of distinct word forms in Shakespeare's *Hamlet* (excluding stage directions and proper nouns) as a functions of the length of the passage from which the estimate is inferred. The true value, 4,268, is shown as the horizontal line.

and forgo running Linear Program 2. The results are plotted in Figure 3. The estimates converge towards the true value of 4,268 distinct words extremely rapidly and are slightly negatively biased, perhaps reflecting the fact that words appearing close together are correlated.

In contrast to Hamlet's charge that "there are more things in heaven and earth...than are dreamt of in your philosophy," we can say that there are almost exactly as many things in *Hamlet* as can be dreamt of from 10% of *Hamlet*.

4 OVERVIEW OF PROOF OF THEOREM 1.11

In this section, we give a detailed high-level overview of the proof of Theorem 1.11. The complete proof is given in Section 6 and fleshes out the scaffold described here. The proof of Theorem 1.11 decomposes into three main parts, described in the following three sections.

4.1 Compartmentalizing the Probabilistic Portion of the Proof

The first part of the proof argues that with high probability (over the randomness in the independent draws of the sample) the sample will be a "faithful" sample from the distribution—no domain element occurs too much more frequently than one would expect, and the fingerprint entries are reasonably close to their expected values. This part of the proof is intuitively obvious and will follow trivially from a union bound over tail bounds on Poisson random variables and Chernoff

tail bounds. Having thus compartmentalized the probabilistic component of our theorem, we will then argue that the algorithm will *always* be successful whenever it receives a "faithful" sample as input.

The following condition defines what it means for a sample from a distribution to be "faithful" with respect to positive constants $\mathcal{B}, \mathcal{D} \in (0,1)$:

Definition 4.1. A sample of size n with fingerprint $\mathcal{F}$, drawn from a distribution p with histogram h, is said to be *faithful* with respect to positive constants $\mathcal{B}, \mathcal{D} \in (0,1)$ if the following conditions hold:

- For all i,
$$\left| \mathcal{F}_i - \sum_{x:h(x)\neq 0} h(x) \cdot poi(nx, i) \right| \leq \max\left(\mathcal{F}_i^{\frac{1}{2}+\mathcal{D}}, n^{\mathcal{B}(\frac{1}{2}+\mathcal{D})} \right).$$
- For all domain elements i, letting $p(i)$ denote the true probability of i, the number of times i occurs in the sample from p differs from $n \cdot p(i)$ by at most
$$\max\left((n \cdot p(i))^{\frac{1}{2}+\mathcal{D}}, n^{\mathcal{B}(\frac{1}{2}+\mathcal{D})} \right).$$

The following lemma follows easily from basic tail bounds on Poisson random variables and Chernoff bounds.

Lemma 4.2. *For any constants $\mathcal{B}, \mathcal{D} \in (0,1)$, there is a constant $\alpha > 0$ and integer n_0 such that for any $n \geq n_0$, a sample of size n consisting of independent draws from a distribution is "faithful" with respect to $\mathcal{B}, \mathcal{D}$ with probability at least $1 - e^{-n^\alpha}$.*

4.2 The Existence of a "Good" Feasible Point of the Linear Program

The second component of the proof argues that (provided the sample in question is "faithful"), the histogram of the true distribution, rounded to be supported at values in the set X of probabilities corresponding to the linear program variables, is a feasible point, v, of the linear program Find Plausible Histogram with reasonably small objective function value. Recall that the linear program aims to find distributions that "could reasonably have generated" the observed fingerprint $\mathcal{F}$; this portion of the proof guarantees that, provided the sample is faithful, the true distribution, h, minimally modified, will in fact be such a feasible point, v. This portion of the proof is also intuitively clear—the objective function measures the deviation between the expected fingerprint entries (given by the process of drawing the sample from the returned histogram) and the observed fingerprint of the sample; because we are considering the objective function value corresponding to the true histogram (rounded slightly to be supported at probability values in set X), we expect that the observed fingerprint entries will be closely concentrated about these expectations.

Lemma 4.3. *Given constants $\mathcal{B}, \mathcal{D}$, there is an integer n_0 such that for any $n \geq n_0$ and $k < n^{1+\mathcal{B}/2}$ the following holds: Given a distribution of support size at most k with histogram h, and a "faithful" sample of size n with respect to the constants $\mathcal{B}, \mathcal{D}$ with fingerprint $\mathcal{F}$, linear program Find Plausible Histogram has a feasible point $v = v_1, \ldots, v_\ell$ with objective value*
$$\sum \frac{1}{\sqrt{1+\mathcal{F}_i}} \left| \mathcal{F}_i - \sum_{j=1}^{\ell} v_j \cdot poi(nx_j, i) \right| \leq n^{2\mathcal{B}},$$
such that $\sum_i v_i \leq k$ and v is close in relative earthmover distance to the true histogram of the distribution, h, namely if h_v is the histogram obtained by appending the "large probability" portion of the

empirical fingerprint to v, then

$$R(h, v) \leq \frac{1}{n^{c_{\mathcal{B},\mathcal{D}}}} = o(1),$$

where $c_{\mathcal{B},\mathcal{D}} > 0$ is a constant that is dependent on $\mathcal{B}, \mathcal{D}$.

4.3 The Chebyshev Earthmoving Scheme

The final component of the proof, which is the technical heart of the proof, will then argue that given *any* two feasible points of linear program Find Plausible Histogram that both have reasonably small objective function values and both have similar support sizes, they must be close in relative earthmover distance. Since we have already established that the histogram of the true distribution (appropriately rounded) will be a feasible point with small objective function value, it will follow that the solution output by the algorithm must also have small objective function value, and correspond to a distribution of comparable (or smaller) support size and hence must be close in relative earthmover distance to the true distribution from which the sample was drawn. This component of the proof gives rise to the logarithmic term in the $n = O(\frac{k}{\log k})$ bounds on the sample size necessary for accurate estimation of distributions supported on a most k elements.

To establish this component of the proof, we define a class of earthmoving schemes, which will allow us to directly relate the relative earthmover distance between two distributions to the discrepancy in their respective fingerprint expectations. The main technical tool is a Chebyshev polynomial construction, though for clarity, we first describe a simpler scheme that provides some intuition for the Chebyshev construction. We begin by describing the form of our earthmoving schemes; since we hope to relate the cost of such schemes to the discrepancy in expected fingerprints, we will require that the schemes be formulated in terms of the Poisson functions $poi(nx, i)$.

Definition 4.4. For a given n, a *β-bump earthmoving scheme* is defined by a sequence of positive real numbers $\{c_i\}$, the *bump centers*, and a sequence of functions $\{f_i\} : (0, 1] \to \mathbb{R}$ such that $\sum_{i=0}^{\infty} f_i(x) = 1$ for each x, and each function f_i may be expressed as a linear combination of Poisson functions, $f_i(x) = \sum_{j=0}^{\infty} a_{ij} poi(nx, j)$, such that $\sum_{j=0}^{\infty} |a_{ij}| \leq \beta$.

Given a histogram h, the scheme works as follows: For each x such that $h(x) \neq 0$, and each integer $i \geq 0$, move $xh(x) \cdot f_i(x)$ units of probability mass from x to c_i. We denote the histogram resulting from this scheme by $(c, f)(h)$.

Definition 4.5. A bump earthmoving scheme (c, f) is $[\epsilon, k]$-*good* if, for any generalized histogram h of support size $\sum_x h(x) \leq k$, the relative earthmover distance between h and $(c, f)(h)$ is at most ϵ.

The crux of the proof of correctness of our estimator is the explicit construction of a surprisingly good earthmoving scheme. We will show that for any sufficiently large n and $k = \delta n \log n$ for a $\delta \in [1/\log n, 1]$, there exists an $[O(\sqrt{\delta}), k]$-good $O(n^{0.3})$-bump earthmoving scheme. In fact, we will construct a single scheme for all such δ. We begin by defining a simple scheme that illustrates the key properties of a bump earthmoving scheme, and its analysis.

Perhaps the most natural bump earthmoving scheme is where the bump functions $f_i(x) = poi(nx, i) = \frac{e^{-nx}(nx)^i}{i!}$ and the bump centers $c_i = \frac{i}{n}$. For $i = 0$, we may, for example, set $c_0 = \frac{1}{2n}$ to avoid a logarithm of 0 when evaluating relative earthmover distance. This is a valid earthmoving scheme, since $\sum_{i=0}^{\infty} f_i(x) = 1$ for any x.

The motivation for this construction is the fact that, for any i, the amount of probability mass that ends up at c_i in $(c, f)(h)$ is exactly $\frac{i+1}{n}$ times the expectation of the $i+1$st fingerprint in a $Poi(n)$-sample from h:

$$\begin{aligned}((c, f)(h))(c_i) = \sum_{x:h(x)\neq 0} h(x)x \cdot f_i(x) &= \sum_{x:h(x)\neq 0} h(x)x \cdot poi(nx, i) \\ &= \sum_{x:h(x)\neq 0} h(x) \cdot poi(nx, i+1)\frac{i+1}{n} \\ &= \frac{i+1}{n}\,\mathrm{E}[\mathcal{F}_{i+1}].\end{aligned}$$

Consider applying this earthmoving scheme to two histograms h, g with nearly identical fingerprint expectations. Letting $h' = (c, f)(h)$ and $g' = (c, f)(g)$, by definition both h' and g' are supported at the bump centers c_i, and by the above equation, for each i, $|h'(c_i) - g'(c_i)| = \frac{i+1}{n}|\sum_x (h(x) - g(x))poi(nx, i+1)|$, where this expression is exactly $\frac{i+1}{n}$ times the difference between the $i+1$st fingerprint expectations of h and g. In particular, if h and g have nearly identical fingerprint expectations, then h' and g' will be very similar. Analogs of this relation between $R((c, f)(g), (c, f)(h))$ and the discrepancy between the expected fingerprint entries corresponding to g and h will hold for any bump earthmoving scheme, (c, f). Sufficiently "good" earthmoving schemes (guaranteeing that $R(h, h')$ and $R(g, g')$ are small) thus provides a powerful way of bounding the relative earthmover distance between two distributions in terms of the discrepancy in their fingerprint expectations.

The problem with the "Poisson bump" earthmoving scheme described in the previous paragraph is that it not very "good": It incurs a very large relative earthmover cost, particularly for small probabilities. This is due to the fact that most of the mass that starts at a probability below $\frac{1}{n}$ will end up in the zeroth bump, no matter if it has probability nearly $\frac{1}{n}$ or the rather lower $\frac{1}{k}$. Phrased differently, the problem with this scheme is that the first few "bumps" are extremely fat. The situation gets significantly better for higher Poisson functions: Most of the mass of $Poi(i)$ lies within relative distance $O(\frac{1}{\sqrt{i}})$ of i, and hence the scheme is relatively cheap for larger probabilities $x \gg \frac{1}{n}$. We will therefore construct a scheme that uses regular Poisson functions $poi(nx, i)$ for $i \geq O(\log n)$ but takes great care to construct "skinnier" bumps below this region.

The main tool of this construction of skinnier bumps is the Chebyshev polynomials. For each integer $i \geq 0$, the ith Chebyshev polynomial, denoted $T_i(x)$, is the polynomial of degree i such that $T_i(\cos(y)) = \cos(i \cdot y)$. Thus, up to a change of variables, any linear combination of cosine functions up to frequency s may be re-expressed as the same linear combination of the Chebyshev polynomials of orders 0 through s. Given this, constructing a "good" earth-moving scheme is an exercise in trigonometric constructions.

Before formally defining our bump earthmoving scheme, we give a rough sketch of the key features. We define the scheme with respect to a parameter $s = O(\log n)$. For $i > s$, we use the fat Poisson bumps: That is, we define the bump centers $c_i = \frac{i}{n}$ and functions $f_i = poi(nx, i)$. For $i \leq s$, we will use skinnier "Chebyshev bumps"; these bumps will have roughly quadratically spaced bump centers $c_i \approx \frac{i^2}{n \log n}$, with the width of the ith bump roughly $\frac{i}{n \log n}$ (as compared to the larger width of $\frac{\sqrt{i}}{n}$ of the ith Poisson bump). At a high level, the logarithmic factor improvement in our $O(\frac{k}{\log k})$ bound on the sample size necessary to achieve accurate estimation arises, because the first few Chebyshev bumps have width $O(\frac{1}{n \log n})$, in contrast to the first Poisson bump, $poi(nx, 1)$, which has width $O(\frac{1}{n})$.

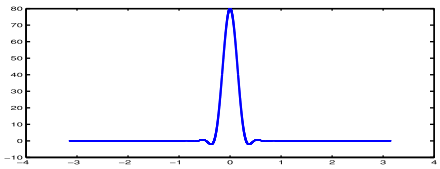

Fig. 4. A plot of the "skinny" function $g_2(y)$ (without the scaling factor) for the value $s = 12$. This is the main ingredient in the Chebyshev bumps construction of Definition 4.6.

Definition 4.6. The *Chebyshev bumps* are defined in terms of n as follows. Let $s = 0.2 \log n$. Define $g_1(y) = \sum_{j=-s}^{s-1} \cos(jy)$. Define

$$g_2(y) = \frac{1}{16s}\left(g_1\left(y - \frac{3\pi}{2s}\right) + 3g_1\left(y - \frac{\pi}{2s}\right) + 3g_1\left(y + \frac{\pi}{2s}\right) + g_1\left(y + \frac{3\pi}{2s}\right)\right),$$

and, for $i \in \{1, \ldots, s-1\}$, define $g_3^i(y) := g_2(y - \frac{i\pi}{s}) + g_2(y + \frac{i\pi}{s})$, $g_3^0 = g_2(y)$, and $g_3^s = g_2(y + \pi)$. Let $t_i(x)$ be the linear combination of Chebyshev polynomials so $t_i(\cos(y)) = g_3^i(y)$. We thus define $s + 1$ functions, the "skinny bumps," to be $B_i(x) = t_i(1 - \frac{xn}{2s}) \sum_{j=0}^{s-1} poi(xn, j)$, for $i \in \{0, \ldots, s\}$. That is, $B_i(x)$ is related to $g_3^i(y)$ by the coordinate transformation $x = \frac{2s}{n}(1 - \cos(y))$, and scaling by $\sum_{j=0}^{s-1} poi(xn, j)$.

See Figure 4 for a plot of $g_2(y)$, illustrating a "skinny Chebyshev bump." The Chebyshev bumps of Definition 4.6 are "third order"; If, instead, we had considered the analogous less-skinny "second order" bumps by defining $g_2(y) := \frac{1}{8s}(g_1(y - \frac{\pi}{s}) + 2g_1(y) + g_1(y + \frac{\pi}{s}))$, then the results would still hold, though the proofs are slightly more cumbersome.

Definition 4.7. The *Chebyshev earthmoving scheme* is defined in terms of n as follows: As in Definition 4.6, let $s = 0.2 \log n$. For $i \geq s + 1$, define the ith bump function $f_i(x) = poi(nx, i - 1)$ and associated bump center $c_i = \frac{i-1}{n}$. For $i \in \{0, \ldots, s\}$ let $f_i(x) = B_i(x)$, and for $i \in \{1, \ldots, s\}$, define their associated bump centers $c_i = \frac{2s}{n}(1 - \cos(\frac{i\pi}{s}))$, and let $c_0 := c_1$.

The following lemma characterizes the key properties of the Chebyshev earthmoving scheme. Namely (1) that the scheme is, in fact, an earthmoving scheme, (2) that each bump can be expressed as a low-weight linear combination of Poisson functions, and (3) that the scheme incurs a small relative-earthmover cost.

Lemma 4.8. *The Chebyshev earthmoving scheme, of Definition 4.7 has the following properties:*

(1) For any $x \geq 0$,

$$\sum_{i \geq 0} f_i(x) = 1,$$

hence the Chebyshev earthmoving scheme is a valid earthmoving scheme.

(2) Each $B_i(x)$ may be expressed as $\sum_{j=0}^{\infty} a_{ij} poi(nx, j)$ for a_{ij} satisfying

$$\sum_{j=0}^{\infty} |a_{ij}| \leq 2n^{0.3}.$$

(3) There is an absolute constant C such that the Chebyshev earthmoving scheme is $[C\sqrt{\delta}, k]$-good, for $k = \delta n \log n$, and $\delta \geq \frac{1}{\log n}$.

4.4 Putting the Pieces Together

Given the lemmas described in the above sections, we now sketch how to assemble these pieces into a proof of Theorem 1.11. Let p denote the true histogram of the distribution of support size at most k from which the sample of size $n = c\frac{k}{\log k}$ was drawn. Lemma 4.2 guarantees that with probability at least $1 - e^{-n^{\Theta(1)}}$, the sample will be "faithful" (see Definition 4.1), in which case Lemma 4.3 guarantees that there exists a feasible point of the linear program FIND PLAUSIBLE HISTOGRAM with objective function value at most $n^{2\beta}$, such that the corresponding histogram (after the "large probability" portion of the empirical fingerprint is appended), h_v, satisfies $R(p, h_v) \leq 1/n^{\Theta(1)} = o(1)$, and the effective support size of histogram h_v satisfies $\sum h_v \leq 2k$. Hence if we set the error parameter, δ of algorithm FIND SIMPLEST PLAUSIBLE HISTOGRAM, to equal $n^{2\beta}$, then we are guaranteed that this linear program will output a point $\hat{h}_{LP}$ that has effective support size at most $2k$, and would yield an objective value of at most $v_{opt} + \delta \leq 2n^{2\beta}$ for the linear program FIND PLAUSIBLE HISTOGRAM.

Let $\hat{h}$ denote the histogram returned by the whole algorithm—consisting of the solution to linear program FIND SIMPLEST PLAUSIBLE HISTOGRAM, $\hat{h}_{LP}$, with the large probability portion of the empirical fingerprint appended. Note that we aim to show that $R(h_v, \hat{h}) = O(1/\sqrt{c})$, from which, by the triangle inequality, it will follow that $R(p, \hat{h}) = O(1/\sqrt{c})$, as desired.

To show this, we leverage the Chebyshev earthmoving scheme (Definition 4.7). First, note that the "large probability" regions of $\hat{h}$ and h_v are identical, thus it remains to bound the relative earthmover distance between their small-probability regions. To this order, let g_v and g denote the results of applying the Chebyshev earthmoving schemes to h_v and $\hat{h}$, respectively. The third condition of Lemma 4.8 guarantees that $R(g_v, h_v) = O(1/\sqrt{c})$, and $R(g, \hat{h}) = O(1/\sqrt{c})$. Hence, all that remains, is to bound $R(g_v, g)$.

The high-level idea is that we know that $\hat{h}$ and h_v have similar fingerprint expectations, because they both have small values for the objective function of linear program FIND PLAUSIBLE HISTOGRAM. The second condition of Lemma 4.8 shows that, essentially, one can translate this discrepancy in fingerprint expectations to a bound on the relative earthmover distance at a cost of a factor of $O(n^{0.3})$, and a normalizing factor of $O(\frac{\log n}{n})$. Formally, letting c_i denote one of the first $O(\log n)$ bump centers, with $f_j(x) = \sum_{\ell \geq 0} a_{\ell,j} \cdot poi(xn, \ell)$ denoting the jth bump function of the earthmoving scheme, we have the following where $\sum_x$ is shorthand for $\sum_{x:\hat{h}(x)+h_v(x)\neq 0}$:

$$
\begin{aligned}
|g(c_i) - g_v(c_i)| &= \left| \sum_x (\hat{h}(x) - h_v(x)) x f_i(x) \right| \\
&= \left| \sum_x (\hat{h}(x) - h_v(x)) x \sum_{j \geq 0} a_{ij} poi(xn, j) \right| \\
&= \left| \sum_{j \geq 0} a_{ij} \sum_x (\hat{h}(x) - h_v(x)) x poi(xn, j) \right| \\
&= \left| \sum_{j \geq 1} a_{i,j-1} \frac{j}{n} \sum_x (\hat{h}(x) - h_v(x)) poi(xn, j) \right| \\
&\leq \left| \sum_{i,j} a_{i,j} \right| \left| \sum_{j \geq 1} \frac{j}{n} \sum_x (\hat{h}(x) - h_v(x)) poi(xn, j) \right|.
\end{aligned}
$$

Because the large probability portions of $\hat{h}$ and h_v are identical, the bulk of the above discrepancy is accounted for by the first $O(\log n)$ fingerprint expectations, hence the above sum is effectively over $j \in [1, O(\log n)]$, in which case the above quantity is bounded by the discrepancy in fingerprint expectations, multiplied by a factor of at most $|\sum_{i,j} a_{i,j}\frac{j}{n} = O(n^{0.3}\frac{\log n}{n}) = O(n^{-0.7}\log n)$. The proof concludes by noting that the bounds of $O(n^{2\beta})$ on the objective function values of $\hat{h}$ and h_v, which are the discrepancies in fingerprint expectations normalized by a factor of $\frac{1}{\sqrt{1+\mathcal{F}_i}} \geq 1/\sqrt{n+1}$, immediately implies that the discrepancies in fingerprint expectations (unnormalized) are bounded by $O(n^{1/2+2\mathcal{B}})$. Hence, choosing $2\mathcal{B}$ to be a sufficiently small constant yields that $O(n^{1/2+2\mathcal{B}}n^{-0.7}\log n) = o(1)$.

Hence we have the following:

$$R(p,\hat{h}) \leq R(p,h_v) + R(h_v,g_v) + R(g_v,g) + R(g,\hat{h}) = o(1) + O(1/\sqrt{c}) + o(1) + O(1/\sqrt{c}) = O(1/\sqrt{c}),$$

where the "o" and "O" notation is with respect to n. The details of this high-level proof overview are given in a self-contained fashion in Section 6.

5 PROPERTIES OF PAIRS OF DISTRIBUTIONS

Our general approach for constructing constant-factor optimal estimators for symmetric properties of distributions can be extended to yield constant-factor optimal estimators for many natural symmetric properties of *pairs* of distributions, including total variation distance (ℓ_1 distance). In analogy with the single-distribution setting, given a pair of distributions over a common domain, a property of the pair of distributions is symmetric if its value is invariant to permutations of the domain.

For properties of pairs of distributions, an estimator receives two samples as input, one drawn from the first distribution and one drawn independently from the second distribution. As with the analysis of estimators for properties of a single distribution, we begin by extending our definitions of *fingerprints* and *histograms* to this two-distribution setting.

Definition 5.1. The *fingerprint* $\mathcal{F}$ of a sample of size n_1 from distribution p_1 and a sample of size n_2 from distribution p_2 is a $n_1 \times n_2$ matrix, whose entry $\mathcal{F}(i,j)$ is given by the number of domain elements that are seen exactly i times in the sample from p_1 and exactly j times in the sample from p_2.

Definition 5.2. The *histogram* $h_{p_1,p_2} : [0,1]^2 \setminus \{(0,0)\} \to \mathbb{N} \cup 0$ of a pair of distributions p_1, p_2 is defined by letting $h_{p_1,p_2}(x,y)$ be the number of domain elements that occur with probability x in distribution p_1 and probability y in distribution p_2.

Thus for any two-dimensional histogram h corresponding to a pair of distributions, we have

$$\sum_{x,y:h(x,y)\neq 0} x \cdot h(x,y) = \sum_{x,y:h(x,y)\neq 0} y \cdot h(x,y) = 1.$$

As in the case with symmetric properties of single distributions, symmetric properties of pairs of distributions are functions of only the histogram of the pair of distributions, and, given any estimator that takes as input the actual pair of samples, there is an estimator of equivalent performance that takes as input the fingerprint $\mathcal{F}$ derived from such a pair of samples.

Both total variation distance (ℓ_1 distance), and Kullback-Leibler divergence are symmetric properties:

Example 5.3. Consider a pair of distributions p_1, p_2 with histogram h:

- The total variation distance (ℓ_1 distance) is given by
$$D_{tv}(p_1, p_2) = \frac{1}{2} \sum_{(x,y):h(x,y)\neq 0} h(x,y) \cdot |x - y|.$$
- The Kullback-Leibler divergence is given by
$$D_{KL}(p_1||p_2) = \sum_{(x,y):h(x,y)\neq 0} h(x,y) \cdot x \log \frac{x}{y}.$$

We will use the following two-dimensional earthmover metric on the set of two-dimensional generalized histograms. Note that it does not make sense to define a strict analog of the relative earthmover distance of Definition 1.6, since a given histogram entry $h(x, y)$ does not correspond to a single quantity of probability mass—it corresponds to $xh(x, y)$ mass in one distribution and $yh(x, y)$ mass in the other distribution. Thus the following metric is in terms of moving *histogram entries* rather than probability mass.

Definition 5.4. Given two two-dimensional generalized histograms h_1, h_2, their *histogram distance*, denoted $W(h_1, h_2)$, is defined to be the minimum over all schemes of moving the histogram values in h_1 to yield h_2, where the cost of moving histogram value c at location x, y to location x', y' is $c\,(|x - x'| + |y - y'|)$. To ensure that such a scheme always exists, in the case that $\sum_{x,y:x+y>0} h_1(x, y) < \sum_{x,y:x+y>0} h_2(x, y)$, one proceeds as if
$$h_1(0, 0) = \sum_{x,y:x+y>0} h_2(x, y) - \sum_{x,y:x+y>0} h_1(x, y)$$
and analogously for the case in which h_2 contains fewer histogram entries.

We provide an example of the above definitions:

Example 5.5. Define distributions $p_1 = Unif[k]$, and $p_2 = Unif[k/2]$, where the $k/2$ support elements of distribution p_2 are contained in the support of p_1. The corresponding histogram h_{p_1,p_2}, is defined as $h_{p_1,p_2}(\frac{1}{k}, \frac{2}{k}) = \frac{k}{2}$, $h_{p_1,p_2}(\frac{1}{k}, 0) = \frac{k}{2}$, and $h_{p_1,p_2}(x, y) = 0$ for all other values of x, y.

Considering a second pair of distributions, $q_1 = q_2 = Unif[k/4]$, with histogram $h_{q_1,q_2}(\frac{4}{k}, \frac{4}{k}) = \frac{k}{4}$, we have
$$W(h_{p_1,p_2}, h_{q_1,q_2}) = \frac{k}{4}\left(\left|\frac{1}{k} - \frac{4}{k}\right| + \left|\frac{2}{k} - \frac{4}{k}\right|\right) + \frac{k}{4}\left(\left|\frac{1}{k} - 0\right| + \left|\frac{2}{k} - 0\right|\right) + \frac{k}{2}\left(\left|\frac{1}{k} - 0\right| + \left|0 - 0\right|\right) = \frac{5}{2},$$
since the optimal scheme is to move $k/4$ histogram entries in h_{p_1,p_2} from $(1/k, 2/k)$ to location $(4/k, 4/k)$, and all the remaining histogram entries must be moved to $(0, 0)$ to yield histogram h_{q_1,q_2}.

We note that ℓ_1 distance is 1-Lipschitz with respect to the above distance metric:

Fact 3. *For any pair of two-dimensional generalized histograms, h, h'*
$$W(h, h') \geq \left| \sum_{x,y:h(x,y)\neq 0} h(x, y)|x - y| - \sum_{x,y:h'(x,y)\neq 0} h'(x, y)|x - y| \right|.$$
Hence if $h = h_{p_1,p_2}$ and $h' = h_{q_1,q_2}$ are histograms corresponding to pairs of distributions, then $W(h_{p_1,p_2}, h_{q_1,q_2}) \geq |D_{tv}(p_1, p_2) - D_{tv}(q_1, q_2)|$.

Both our algorithm for estimating properties of pairs of distributions and its analysis parallel their analogs in the one-distribution setting. For simplicity, we restrict our attention to the setting in which one obtains samples of size n from both distributions—though our approach extends naturally to the setting in which one obtains samples of different sizes from the two distributions.

THEOREM 5.6. *There exist absolute constants $\alpha, \gamma > 0$ such that for any $c > 0$, for sufficiently large k, given two samples of size $n = c\frac{k}{\log k}$ consisting of independent draws from each of two distributions, $p, q \in \mathcal{D}^k$ with a two-dimensional histogram $h_{p,q}$, with probability at least $1 - e^{-n^\alpha}$ over the randomness in the selection of the sample, our algorithm returns a two-dimensional generalized histogram g_{LP} such that*

$$W(g_{LP}, h_{p,q}) \leq \frac{\gamma}{\sqrt{c}}.$$

Together with Fact 3, this immediately implies our $O(k/logk)$ sample estimator for total variation distance, Theorem 1.13. The proof of Theorem 5.6 closely parallels that of its one distribution analog, Theorem 1.11, and the complete proof is provided in Appendix A.

6 PROOF OF THEOREM 1.11

We begin by restating Algorithm 1 in a form that our proofs can more easily reference. The one difference between this algorithm, and Algorithm 1 (beyond relabeling variables) is the manner in which the fingerprint is partitioned into the "easy" regime for which the empirical estimate is applied, and the "hard" regime for which the linear programming approach is applied. Here, for simplicity, we analyze the partitioning scheme that simply chooses a fixed cutoff and applies the naive empirical estimator to any fingerprint entry $\mathcal{F}_i$ for i above the cutoff and applies the linear programming approach to the smaller fingerprint indices.

For clarity of exposition, we state the algorithm in terms of three positive constants, $\mathcal{B}, C$, and $\mathcal{D}$, which can be defined arbitrarily provided the following inequalities hold:

$$0.1 > \mathcal{B} > C > \mathcal{B}\left(\frac{1}{2} + \mathcal{D}\right) > \frac{\mathcal{B}}{2} > \mathcal{D} > 0.$$

LINEAR PROGRAM 3.

Given a n-sample fingerprint $\mathcal{F}$:

- Define the set $X := \{\frac{1}{n^2}, \frac{2}{n^2}, \frac{3}{n^2}, \ldots, \frac{n^{\mathcal{B}}+n^C}{n}\}$.
- For each $x \in X$, define the associated LP variable v_x.

The linear program is defined as follows:

$$\text{Minimize} \sum_{i=1}^{n^{\mathcal{B}}} \frac{1}{\sqrt{1+\mathcal{F}_i}} \left| \mathcal{F}_i - \sum_{x \in X} poi(nx, i) \cdot v_x \right|$$

Subject to:

- $\sum_{x \in X} x \cdot v_x + \sum_{i=n^{\mathcal{B}}+2n^C}^{n} \frac{i}{n} \mathcal{F}_i = 1$ (total prob. mass = 1)
- $\forall x \in X, v_x \geq 0$ (histogram entries are non-negative)

Linear Program 4.

Given a n-sample fingerprint $\mathcal{F}$ and value val:

- Define the set $X := \{\frac{1}{n^2}, \frac{2}{n^2}, \frac{3}{n^2}, \ldots, \frac{n^{\mathcal{B}}+n^C}{n}\}$.
- For each $x \in X$, define the associated LP variable v_x.

The linear program is defined as follows:

$$\text{Minimize} \sum_{x \in X} v_x, \text{ (minimize support size of histogram corresponding to } v_x)$$

Subject to:

- $\sum_{i=1}^{n^{\mathcal{B}}} \frac{1}{\sqrt{\mathcal{F}_i+1}} |\mathcal{F}_i - \sum_{x \in X} poi(nx, i)v_x| \leq val + n^{2\mathcal{B}}$ (expected fingerprints are close to $\mathcal{F}$)
- $\sum_{x \in X} x \cdot v_x + \sum_{i=n^{\mathcal{B}}+2n^C}^{n} \frac{i}{n}\mathcal{F}_i = 1$ (total prob. mass = 1)
- $\forall x \in X, v_x \geq 0$ (histogram entries are non-negative)

Algorithm 2. *Estimate Unseen*

Input: n-sample fingerprint $\mathcal{F}$.
Output: Histogram g_{LP}.

- Let val be the objective function value of the solution to Linear Program 3, on input $\mathcal{F}$.
- Let $v = (v_{x_1}, v_{x_2}, \ldots)$ be the solution to Linear Program 4, on input $\mathcal{F}$ and val.
- Let g_{LP} be the histogram formed by setting $g_{LP}(x_i) = v_{x_i}$ for all i, and then for each integer $j \geq n^{\mathcal{B}} + 2n^C$, incrementing $g_{LP}(\frac{j}{n})$ by $\mathcal{F}_j$.

For convenience, we restate Theorem 1.11 in terms of the above algorithm.

Theorem 1.11. *For any choice of constants $\mathcal{B}, C$, and $\mathcal{D}$ that satisfy $0.1 > \mathcal{B} > C > \mathcal{B}(\frac{1}{2} + \mathcal{D}) > \frac{\mathcal{B}}{2} > \mathcal{D} > 0$, there exist absolute constants $a, b > 0$ such that for any $c > 0$, there is a constant k_c such that given a sample of size $n = c\frac{k}{\log k}$ consisting of independent draws from a distribution $p \in \mathcal{D}^k$ with $k > k_c$, with probability at least $1 - e^{-k^a}$ over the randomness in the selection of the sample, Algorithm 2 returns a histogram g_{LP} such that*

$$R(p, g_{LP}) \leq \frac{b}{\sqrt{c}}.$$

The proof of Theorem 1.11 decomposes into three main parts, addressed in the following three sections.

6.1 Compartmentalizing the Probabilistic Portion of the Proof

We first argue that with high probability (over the randomness in the independent draws of the sample) the sample will be a "faithful" sample from the distribution—no domain element occurs too much more frequently than one would expect, and the fingerprint entries are reasonably close to their expected values. This part will follow from a union bound over tail bounds on Poisson random variables and Chernoff tail bounds. The remainder of the proof will then argue that the algorithm will *always* be successful whenever it receives a "faithful" sample as input.

The following condition defines what it means for a sample from a distribution to be "faithful" with respect to positive constants $\mathcal{B}, \mathcal{D} \in (0, 1)$:

Definition 6.1. A sample of size n with fingerprint $\mathcal{F}$, drawn from a distribution p with histogram h, is said to be *faithful* with respect to positive constants $\mathcal{B}, \mathcal{D} \in (0,1)$ if the following conditions hold:

- For all i,
$$\left| \mathcal{F}_i - \sum_{x:h(x)\neq 0} h(x) \cdot poi(nx, i) \right| \leq \max\left(\mathcal{F}_i^{\frac{1}{2}+\mathcal{D}}, n^{\mathcal{B}(\frac{1}{2}+\mathcal{D})} \right).$$
- For all domain elements i, letting $p(i)$ denote the true probability of i, the number of times i occurs in the sample from p differs from $n \cdot p(i)$ by at most
$$\max\left((n \cdot p(i))^{\frac{1}{2}+\mathcal{D}}, n^{\mathcal{B}(\frac{1}{2}+\mathcal{D})} \right).$$

The following lemma is proven via the standard "Poissonization" technique (see, e.g., Reference [28]).

Lemma 6.2. *For any constants $\mathcal{B}, \mathcal{D} \in (0,1)$, there is a constant $\alpha > 0$ and integer n_0 such that for any $n \geq n_0$, a sample of size n consisting of independent draws from a distribution is "faithful" with respect to $\mathcal{B}, \mathcal{D}$ with probability at least $1 - e^{-n^\alpha}$.*

Proof. We first analyze the case of a $Poi(n)$-sized sample drawn from a distribution with histogram h. Thus
$$\mathrm{E}[\mathcal{F}_i] = \sum_{x:h(x)\neq 0} h(x) poi(nx, i).$$
Additionally, the number of times each domain element occurs is independent of the number of times the other domain elements occur, and thus each fingerprint entry $\mathcal{F}_i$ is the sum of independent random 0/1 variables, representing whether each domain element occurred exactly i times in the sample (i.e., contributing 1 towards $\mathcal{F}_i$). By independence, Chernoff bounds apply.

We split the analysis into two cases, according to whether $\mathrm{E}[\mathcal{F}_i] \geq n^{\mathcal{B}}$. In the case that $\mathrm{E}[\mathcal{F}_i] < n^{\mathcal{B}}$, we leverage the basic Chernoff bound that if X is the sum of independent 0/1 random variables with $\mathrm{E}[X] \leq S$, then for any $\delta \in (0,1)$,
$$\Pr[|X - \mathrm{E}[X]| \geq \delta S] \leq 2e^{-\delta^2 S/3}.$$
Applied to our present setting where $\mathcal{F}_i$ is a sum of independent 0/1 random variables, provided $\mathrm{E}[\mathcal{F}_i] < n^{\mathcal{B}}$, we have
$$\Pr\left[|\mathcal{F}_i - \mathrm{E}[\mathcal{F}_i]| \geq (n^{\mathcal{B}})^{\frac{1}{2}+\mathcal{D}} \right] \leq 2e^{-(\frac{1}{(n^{\mathcal{B}})^{1/2-\mathcal{D}}})^2 \frac{n^{\mathcal{B}}}{3}} = 2e^{-n^{2\mathcal{B}\mathcal{D}}/3}.$$
In the case that $\mathrm{E}[\mathcal{F}_i] \geq n^{\mathcal{B}}$, the same Chernoff bound yields
$$\Pr\left[|\mathcal{F}_i - \mathrm{E}[\mathcal{F}_i]| \geq \mathrm{E}[\mathcal{F}_i]^{\frac{1}{2}+\mathcal{D}} \right] \leq 2e^{-(\frac{1}{\mathrm{E}[\mathcal{F}_i]^{1/2-\mathcal{D}}})^2 \frac{\mathrm{E}[\mathcal{F}_i]}{3}} = 2e^{-(\mathrm{E}[\mathcal{F}_i]^{2\mathcal{D}})/3} \leq 2e^{-n^{2\mathcal{B}\mathcal{D}}/3}.$$
A union bound over the first n fingerprints shows that the probability that given a sample (consisting of $Poi(n)$ draws), the probability that any of the fingerprint entries violate the first condition of *faithful* is at most $n \cdot 2e^{-\frac{n^{2\mathcal{B}\mathcal{D}}}{3}} \leq e^{-n^{\Omega(1)}}$ as desired.

For the second condition of "faithful," in analogy with the above argument, for any $\lambda \leq S$, and $\delta \in (0,1)$,
$$\Pr[|Poi(\lambda) - \lambda| > \delta S] \leq 2e^{-\delta^2 S/3}.$$

Hence for $x = n \cdot p(i) \geq n^{\mathcal{B}}$, the probability that the number of occurrences of domain element i differs from its expectation of $n \cdot p(i)$ by at least $(n \cdot p(i))^{\frac{1}{2}+\mathcal{D}}$ is bounded by $2e^{-(n\cdot p(i))^{2\mathcal{D}}/3} \leq e^{-n^{\Omega(1)}}$. Similarly, in the case that $x = n \cdot p(i) < n^{\mathcal{B}}$,

$$\Pr\left[|Poi(x) - x| > n^{\mathcal{B}(\frac{1}{2}+\mathcal{D})}\right] \leq e^{-n^{\Omega(1)}}.$$

Thus we have shown that provided we are considering a sample of size $Poi(n)$, the probability that the conditions hold is at least $1 - e^{-n^{\Omega(1)}}$. To conclude, note that $\Pr[Poi(n) = n] > \frac{1}{3\sqrt{n}}$, and hence the probability that the conditions do not hold for a sample of size exactly n (namely the probability that they do not hold for a sample of size $Poi(n)$, conditioned on the sample size being exactly n) is at most a factor of $3\sqrt{n}$ larger, and hence this probability of failure is still $e^{-n^{\Omega(1)}}$, as desired.

6.2 The Existence of a "Good" Feasible Point of the Linear Program

The second component of the proof argues that (provided the sample in question is "faithful"), the histogram of the true distribution, rounded to be supported at values in the set X of probabilities corresponding to the linear program variables, is a feasible point, v of Linear Program 3 with objective function value at most $n^{\mathcal{B}}$. This portion of the proof is also intuitively clear—the objective function measures the deviation between the expected fingerprint entries (given by the process of drawing the sample from the returned histogram) and the observed fingerprint of the sample; because we are considering the objective function value corresponding to the true histogram (rounded slightly to be supported at probability values in set X), we expect that the observed fingerprint entries will be closely concentrated about these expectations.

Lemma 6.3. *Given constants $\mathcal{B}, \mathcal{D}$, there is an integer n_0 such that for any $n \geq n_0$ and $k < n^{1+\mathcal{B}/2}$ the following holds: Given a distribution of support size at most k with histogram h, and a "faithful" sample of size n with respect to the constants $\mathcal{B}, \mathcal{D}$ with fingerprint $\mathcal{F}$, linear program* Find Plausible Histogram *has a feasible point $v = v_1, \ldots, v_\ell$ with objective value*

$$\sum \frac{1}{\sqrt{1+\mathcal{F}_i}} \left| \mathcal{F}_i - \sum_{j=1}^{\ell} v_j \cdot poi(nx_j, i) \right| \leq n^{2\mathcal{B}},$$

such that $\sum_i v_i \leq k$, and v is close in relative earthmover distance to the true histogram of the distribution, h, namely if h_v is the histogram obtained by appending the "large probability" portion of the empirical fingerprint to v, then

$$R(h, v) \leq \frac{1}{n^{c_{\mathcal{B},\mathcal{D}}}} = o(1),$$

where $c_{\mathcal{B},\mathcal{D}} > 0$ is a constant that is dependent on $\mathcal{B}, \mathcal{D}$.

Before giving a formal proof, we describe the high-level intuition of the proof. Roughly, we construct the desired v by taking the portion of h with probabilities at most $\frac{n^{\mathcal{B}}+n^{C}}{n}$ and rounding the support of h to the closest multiple of $1/n^2$, to be supported at points in the set $X = \{1/n^2, 2/n^2, \ldots\}$. We will then need to adjust the total probability mass accounted for in v to ensure that the first constraint of the linear program is satisfied, namely the total (implicit) probability mass is 1; this adjusting of mass must be accomplished while ensuring that the fingerprint expectations do not change significantly, to ensure that objective function value remains small.

The "support size" of v, $\sum_x v_x$, will easily be bounded by $2k$, since we are assuming that the support size of the distribution corresponding to the true histogram, h, is bounded by k, and the rounding will at most double this value. To argue that v is a feasible point of the linear program, we note that the mesh X is sufficiently fine to guarantee that the rounding of the support of a

histogram to probabilities that are integer multiples of $1/n^2$ does not greatly change the expected fingerprints, and hence the expected fingerprint entries associated with v will be close to those of h. Our definition of "faithful" guarantees that all fingerprint entries are close to their expectations, and hence the objective function will be small. (Intuitively, the reader should be convinced that there is *some* suitably fine mesh for which rounding issues are benign; there is nothing special about $1/n^2$ except that it simplifies some of the proof.)

To bound the relative earthmover distance between the true histogram h and the histogram h_v associated to v, we first note that the portion of h_v corresponding to probabilities below $\frac{n^{\mathcal{B}}+n^C}{n}$ will be extremely similar to h, because it was created from h. For probabilities above $\frac{n^{\mathcal{B}}+2n^C}{n}$, h_v and h will be similar, because these "frequently occurring" elements will appear close to their expected number of times, by the second condition of "faithful" and hence the relative earthmover distance between the empirical histogram and the true histogram in this frequently occurring region will also be small. Finally, the only remaining region is the relatively narrow intermediate region of probabilities, which is narrow enough so probability mass can be moved arbitrarily within this intermediate region while incurring minimal relative earthmover cost. The formal proof of Lemma 6.3 containing the details of this argument is given below.

Proof of Lemma 6.3. We explicitly define v as a function of the true histogram h and fingerprint of the sample, $\mathcal{F}$, as follows:

(1) Define h' such that $h'(x) = h(x)$ for all $x \le \frac{n^{\mathcal{B}}+n^C}{n}$, and $h'(x) = 0$ for all $x > \frac{n^{\mathcal{B}}+n^C}{n}$.
(2) Initialize v to be 0, and for each $x \ge 1/n^2$ s.t. $h'(x) \neq 0$ increment $v_{\bar{x}}$ by $h'(x)$, where $\bar{x} = \max(z \in X : z \le x)$ is x rounded down to the closest point in the set $X = \{1/n^2, 2/n^2, \ldots\}$.
(3) Let $m := \sum_{x\in X} x v_x + m_{\mathcal{F}}$, where $m_{\mathcal{F}} := \sum_{i \ge n^{\mathcal{B}}+2n^C} \frac{i}{n}\mathcal{F}_i$. If $m < 1$, then increment v_y by $(1-m)/y$, where $y = \frac{n^{\mathcal{B}}+n^C}{n}$. Otherwise, if $m \ge 1$, for all $x \in X$ scale v_x by a factor of $s = \frac{1-m_{\mathcal{F}}}{m-m_{\mathcal{F}}}$, after which the total probability mass $m_{\mathcal{F}} + \sum_{x\in X} x v_x$ will be 1.

We first note that the above procedure is well defined, since $m_{\mathcal{F}} \le 1$, and, hence, when $m > 1$ and the scaling factor s is applied, s will be positive.

Note that by construction, the first and second conditions of the linear program are trivially satisfied. We now consider the objective function value. Note that since $C > \frac{1}{2}\mathcal{B}$, we have $\sum_{i\le n^{\mathcal{B}}} poi(n^{\mathcal{B}} + n^C, i) = o(1/n)$, so the fact that we are truncating h at probability $\frac{n^{\mathcal{B}}+n^C}{n}$ in the first step in our construction of v, has little effect on the first $n^{\mathcal{B}}$ "expected fingerprints": specifically, for $i \le n^{\mathcal{B}}$,

$$\sum_{x:h(x)\neq 0} (h'(x) - h(x))\, poi(nx, i) = o(1).$$

Together with the first condition of the definition of faithful, by the triangle inequality, for each i,

$$\frac{1}{\sqrt{\mathcal{F}_i + 1}} \left| \mathcal{F}_i - \sum_{x:h'(x)\neq 0} h'(x) poi(nx, i) \right| \le \max\left(\mathcal{F}_i^{\mathcal{D}}, n^{\mathcal{B}(\frac{1}{2}+\mathcal{D})}\right) + o(1).$$

We now analyze how the discretization contributes to the expected fingerprints. To this end, note that $|\frac{d}{dx} poi(nx, i)| \le n$, and since we are discretizing to multiples of $1/n^2$, the discretization alters the contribution of each domain element to each "expected fingerprint" by at most $n/n^2 = 1/n$ (including those domain elements with probability $< 1/n^2$ which are effectively rounded to 0). Thus, since the support size is bounded by k, the discretization alters each "expected fingerprint" by at most k/n, and thus contributes at most $n^{\mathcal{B}}\frac{k}{n}$ to the quantity $\sum_{i=1}^{n^{\mathcal{B}}} \frac{1}{\sqrt{\mathcal{F}_i+1}} \left|\mathcal{F}_i - \sum_{x\in X} poi(nx, i) v_x\right|$.

To conclude our analysis of the objective function of the linear program for the point v, we consider the effect of the final adjustment of probability mass in the construction of v. In the case that $m \le 1$, where m is the amount of mass in v before the final adjustment (as defined in the final step in the construction of v), mass is added to v_y, where $y = \frac{n^{\mathcal{B}}+n^{C}}{n}$, and thus since $\sum_{i \le n^{\mathcal{B}}} poi(ky, i) = o(1/n)$, this added mass—no matter how much—alters each $\sum_{x \in X} v_x poi(kx, i)$ by at most $o(1)$.

In the case where $m > 1$ and we must scale down the low-frequency portion of the distribution by the quantity $s < 1$, we must do a more delicate analysis. We first bound s in such a way that we can leverage the definition of "faithful." Recall that by definition at the start of the third step of the construction of v, we have $s = \frac{1-m_{\mathcal{F}}}{m-m_{\mathcal{F}}} = \frac{\sum_{i<n^{\mathcal{B}+2n^{C}}} \frac{i}{n} \mathcal{F}_i}{\sum_{x \in X} x v_x}$. We lowerbound this expression via an upperbound on the denominator, noting that $\sum_{x \in X} x v_x$ is at most the total probability mass below frequency $\frac{n^{\mathcal{B}}+n^{C}}{n}$ in the true histogram h, which by Poisson tail bounds is at most $o(1/n)$ less than the total mass implied by expected fingerprints up to $n^{\mathcal{B}} + 2n^{C}$. Namely, letting $\mathrm{E}[\mathcal{F}_i] = \sum_{x:h(x)\neq 0} h(x) \cdot poi(nx, i)$ be the expected fingerprints of sampling from the true distribution, we have $s \ge \frac{\sum_{i<n^{\mathcal{B}+2n^{C}}} \frac{i}{n} \mathcal{F}_i}{\sum_{i<n^{\mathcal{B}+2n^{C}}} \frac{i}{k} \mathrm{E}[\mathcal{F}_i]} - o(1/n)$.

We bound this expression using the definition of "faithful": For each i, we have $\mathrm{E}[\mathcal{F}_i] \le \mathcal{F}_i + \max(\mathcal{F}_i^{\frac{1}{2}+\mathcal{D}}, n^{\mathcal{B}(\frac{1}{2}+\mathcal{D})}) \le \mathcal{F}_i + \mathcal{F}_i^{\frac{1}{2}+\mathcal{D}} + n^{\mathcal{B}(\frac{1}{2}+\mathcal{D})}$. To bound s, we must bound the sum of these terms, each scaled by $\frac{i}{n}$. Because $x^{\frac{1}{2}+\mathcal{D}}$ is a concave function, and letting $z := \sum_{i<n^{\mathcal{B}+2n^{C}}} \frac{i}{n} = O(\frac{n^{2\mathcal{B}}}{n})$, Jensen's inequality gives that $\sum_{i<n^{\mathcal{B}+2n^{C}}} \frac{i}{n} \mathcal{F}_i^{\frac{1}{2}+\mathcal{D}} \le z(\frac{1}{z} \sum_{i<n^{\mathcal{B}+2n^{C}}} \frac{i}{n} \mathcal{F}_i)^{\frac{1}{2}+\mathcal{D}}$. Thus, defining the mass implied by the low-frequency fingerprints to be $m_S := \sum_{i<n^{\mathcal{B}+2n^{C}}} \frac{i}{n} \mathcal{F}_i$, we bound one over the expression in our bound for s as $\frac{\sum_{i<n^{\mathcal{B}+2n^{C}}} \frac{i}{n} \mathrm{E}[\mathcal{F}_i]}{\sum_{i<n^{\mathcal{B}+2n^{C}}} \frac{i}{n} \mathcal{F}_i} \le 1 + (\frac{z}{m_S})^{\frac{1}{2}-\mathcal{D}} + n^{\mathcal{B}(\frac{1}{2}+\mathcal{D})} \frac{z}{m_S}$. Thus s is at least 1 over this last expression, minus $o(1/n)$, which we bound via the inequality $\frac{1}{1+x} \ge 1 - x$ (for positive x) as: $s \ge 1 - O(n^{(2\mathcal{B}-1)(\frac{1}{2}-\mathcal{D})}) m_S^{-(\frac{1}{2}-\mathcal{D})} - O(n^{2\mathcal{B}+\mathcal{B}(\frac{1}{2}+\mathcal{D})-1})/m_S$.

Recall that v is scaled by s at the end of the third step of its construction, and thus to analyze the contribution of this scaling to the objective function value, we bound the total quantity which will be scaled, $\sum_{i=1}^{n^{\mathcal{B}}} \frac{1}{\sqrt{\mathcal{F}_i+1}} \sum_{x \in X} poi(nx, i) v_x$ at the beginning of step 3. We make use of the bounds on the first constraint derived above, for each i:

$$\frac{1}{\sqrt{\mathcal{F}_i + 1}} \left| \mathcal{F}_i - \sum_{x:h'(x)\neq 0} poi(nx, i) v_x \right| \le \max\left(\mathcal{F}_i^{\mathcal{D}}, n^{\mathcal{B}(\frac{1}{2}+\mathcal{D})}\right) + \frac{k}{n} + o(1),$$

which can be rearranged to

$$\begin{aligned} \frac{1}{\sqrt{\mathcal{F}_i + 1}} \sum_{x:h'(x)\neq 0} poi(nx, i) v_x &\le \frac{\mathcal{F}_i}{\sqrt{\mathcal{F}_i + 1}} + \max\left(\mathcal{F}_i^{\mathcal{D}}, n^{\mathcal{B}(\frac{1}{2}+\mathcal{D})}\right) + \frac{k}{n} + o(1) \\ &\le \sqrt{\mathcal{F}_i} + O(n^{\mathcal{B}(\frac{1}{2}+\mathcal{D})}). \end{aligned}$$

The Cauchy-Schwarz inequality yields that $\sum_{i \le n^{\mathcal{B}}} \sqrt{\mathcal{F}_i} \le \sqrt{\sum_{i \le n^{\mathcal{B}}} \frac{i}{n} \mathcal{F}_i} \sqrt{\sum_{i \le n^{\mathcal{B}}} \frac{n}{i}}$, which is bounded by $\sqrt{m_S} O(\sqrt{n \log n})$.

Thus scaling by s in step 3 modifies the first constraint of the linear program by at most the product of $s - 1$ and $\frac{1}{\sqrt{\mathcal{F}_i+1}} \sum_{x:h'(x)\neq 0} poi(nx, i) v_x$, which we have thus bounded as

$$\min\left(1, O(n^{(2\mathcal{B}-1)(\frac{1}{2}-\mathcal{D})}) m_S^{-(\frac{1}{2}-\mathcal{D})} + O(n^{2\mathcal{B}+\mathcal{B}(\frac{1}{2}+\mathcal{D})-1})/m_S\right) \left(\sqrt{m_S} O(\sqrt{n \log n}) + O(n^{\mathcal{B}(\frac{3}{2}+\mathcal{D})})\right).$$

When $m_S < n^{3\mathcal{B}-1}$, we bound the left parenthetical expression by 1 and the right expression is bounded by $O(\sqrt{n^{3\mathcal{B}}\log k} + n^{\mathcal{B}(\frac{3}{2}+\mathcal{D})}) = O(n^{\mathcal{B}(\frac{3}{2}+\mathcal{D})})$.

Otherwise, when $m_S \in [n^{3\mathcal{B}-1}, 1]$, we bound the product of the first parenthetical with the rightmost term $O(n^{\mathcal{B}(\frac{3}{2}+\mathcal{D})})$ by simply $O(n^{\mathcal{B}(\frac{3}{2}+\mathcal{D})})$. We bound the remaining two cross-terms as

$$O\Big(n^{(2\mathcal{B}-1)(\frac{1}{2}-\mathcal{D})}\Big)m_S^{-(\frac{1}{2}-\mathcal{D})}\sqrt{m_S}O(\sqrt{n\log n}) \le O(n^{\mathcal{B}+\mathcal{D}})$$

and

$$O\Big(n^{2\mathcal{B}+\mathcal{B}(\frac{1}{2}+\mathcal{D})-1}\Big)/m_S\sqrt{m_S}O(\sqrt{n\log n}) \le O\Big(n^{\mathcal{B}(1+\mathcal{D})}\Big).$$

Thus the total contribution of the scaling by s to the objective function is $O(n^{\mathcal{B}(\frac{3}{2}+\mathcal{D})})$.

Thus for sufficiently large n, the objective function value of the constructed point will be bounded by $n^{2\mathcal{B}}$.

We now turn to analyzing the relative earthmover distance $R(h, h_v)$. Consider applying the following earthmoving scheme to h_v to yield a new histogram g. The following scheme applies in the case that no probability mass was scaled down from v in the final step of its construction; in the case that v was scaled down, we consider applying the same earthmoving scheme, with the modification that one never moves more than $xh_v(x)$ mass from location x.

- For each $x \le \frac{n^{\mathcal{B}}+n^C}{n}$, move $\bar{x}h(x)$ units of probability from location $\bar{x}$ to x, where as above, $\bar{x} = \max(z \in X : z \le x)$ is x rounded down to the closest point in set $X = \{1/n^2, 2/n^2, \ldots\}$.
- For each domain element i that occurs $j \ge n^{\mathcal{B}} + 2k^C$ times, move $\frac{j}{n}$ units of probability mass from location $\frac{j}{n}$ to location $p(i)$, where $p(i)$ is the true probability of domain element i.

By our construction of h_v, it follows that the above earthmoving scheme is a valid scheme to apply to h_v, in the sense that it never tries to move more mass from a point than was at that point. And g is the histogram resulting from applying this scheme to h_v. We first show that $R(h_v, g)$ is small, since probability mass is only moved relatively small distances. We will then argue that $R(g, h)$ is small: Roughly, this follows from first noting that g and h will be very similar below probability value $\frac{k^{\mathcal{B}}+n^C}{n}$, and from the second condition of "faithful" g and h will also be quite similar above probability $\frac{n^{\mathcal{B}}+4n^{\mathcal{B}}}{n}$. Thus the bulk of the disparity between g and h is in the very narrow intermediate region, within which mass may be moved at the small per-unit-mass cost of $\log\frac{n^{\mathcal{B}}+O(n^C)}{n^{\mathcal{B}}} \le O(n^{C-\mathcal{B}})$.

We first seek to bound $R(h_v, g)$. To bound the cost of the first component of the scheme, consider some $x \ge \frac{n^{1/2}}{n^2}$. The per-unit-mass cost of applying the scheme at location x is bounded by $\log\frac{x}{x-1/n^2} < 2n^{-1/2}$. From the bound on the support size of h and the construction of h_v, the total probability mass in h_v at probabilities $x \le \frac{n^{1/2}}{n^2}$ is at most $\frac{n}{n^{3/2}} < n^{\mathcal{B}/2-1/2}$, and hence this mass can be moved anywhere at cost $n^{\mathcal{B}/2-1/2}\log(n^2)$. To bound the second component of the scheme, by the second condition of "faithful" for each of these frequently occurring domain elements that occur $j \ge n^{\mathcal{B}} + 2n^C$ times with true probability $p(i)$, we have that $|n \cdot p(i) - j| \le (n \cdot p(i))^{\frac{1}{2}+\mathcal{D}}$, and hence the per-unit-mass cost of this portion of the scheme is bounded by $\log\frac{n^{\mathcal{B}}-n^{\mathcal{B}(\frac{1}{2}+\mathcal{D})}}{n^{\mathcal{B}}} \le O(n^{\mathcal{B}(-\frac{1}{2}+\mathcal{D})})$, which dominates the cost of the first portion of the scheme. Hence

$$R(h_v, g) \le O\Big(n^{\mathcal{B}(-\frac{1}{2}+\mathcal{D})}\Big).$$

We now consider $R(h, g)$. To this end, we will show that

$$\sum_{x \notin [n^{\mathcal{B}-1}, \frac{n^{\mathcal{B}}+4n^C}{k}]} x|h(x) - g(x)| \leq O\Big(n^{\mathcal{B}(-1/2+\mathcal{D})}\Big).$$

First, consider the case that there was no scaling down of v in the final step of the construction. For $x \leq n^{\mathcal{B}-1}$, we have $g(x) = \frac{\bar{x}}{x}h(x)$, and hence for $x > \frac{n^{1/2}}{n^2}$, $|h(x) - g(x)| \leq h(x)n^{-1/2}$. On the other hand, $\sum_{x \leq \frac{n^{1/2}}{n^2}} xh(x) \leq n^{-1/2+\mathcal{B}/2}$, since the support size of h is at most $n \leq n^{1+\mathcal{B}/2}$. Including the possible removal of at most $n^{-1/2+\mathcal{D}}$ units of mass during the scaling in the final step of constructing v, we have that

$$\sum_{x \leq n^{\mathcal{B}-1}} x|h(x) - g(x)| \leq O\Big(n^{-1/2+\mathcal{B}/2}\Big).$$

We now consider the "high probability" regime. From the second condition of "faithful," for each domain element i whose true probability is $p(i) \geq \frac{n^{\mathcal{B}}+4n^C}{n}$, the number of times i occurs in the faithful sample will differ from its expectation $n \cdot p(i)$ by at most $(n \cdot p(i))^{\frac{1}{2}+\mathcal{D}}$. Hence from our condition that $C > \mathcal{B}(\frac{1}{2} + \mathcal{D})$ this element will occur at least $n^{\mathcal{B}} + 2n^C$ times, in which case it will contribute to the portion of h_v corresponding to the empirical distribution. Thus for each such domain element, the contribution to the discrepancy $|h(x) - g(x)|$ is at most $(n \cdot p(i))^{-1/2+\mathcal{D}}$. Hence $\sum_{x \geq n^{\mathcal{B}-1}+4n^{C-1}} x|h(x) - g(x)| \leq n^{\mathcal{B}(-1/2+\mathcal{D})}$, yielding the claim that

$$\sum_{x \notin [n^{\mathcal{B}-1}, \frac{n^{\mathcal{B}}+4n^C}{n}]} x|h(x) - g(x)| \leq O\Big(n^{\mathcal{B}(-1/2+\mathcal{D})}\Big).$$

To conclude, note that all the probability mass in g and h at probabilities below $1/n^2$ can be moved to location $1/n^2$ incurring a relative earthmover cost bounded by $\max_{x \leq 1/n^2} kx|\log xn^2| \leq \frac{k}{n^2} \leq \frac{n^{\mathcal{B}/2}}{n}$. After such a move, the remaining discrepancy between $g(x)$ and $h(x)$ for $x \notin [\frac{n^{\mathcal{B}}}{n}, \frac{n^{\mathcal{B}}+4n^C}{n}]$ can be moved to probability $n^{\mathcal{B}}/n$ at a per-unit-mass cost of at most $\log n^2$, and hence a total cost of at most $O(n^{\mathcal{B}(-1/2+\mathcal{D})} \log n^2)$. After this move, the only region for which $g(x)$ and $h(x)$ differ is the narrow region with $x \in [\frac{n^{\mathcal{B}}}{n}, \frac{n^{\mathcal{B}}+4n^C}{n}]$, within which mass may be moved arbitrarily at a total cost of $\log(1 + 4n^{C-\mathcal{B}}) \leq O(n^{C-\mathcal{B}})$. Hence we have

$$R(h, h_v) \leq R(h, g) + R(g, h_v) \leq O\Big(n^{C-\mathcal{B}} + n^{\mathcal{B}(-1/2+\mathcal{D})} \log n\Big). \qquad \square$$

6.3 Similar Expected Fingerprints Imply Similar Histograms

In this section, we argue that if two histograms h_1, h_2 corresponding to distributions with support size at most $2k$ have the property that their expected fingerprints derived from $Poi(n)$-sized samples are very similar, then $R(h_1, h_2)$ must be small. This will guarantee that any two feasible points of Linear Program 4 that both have small objective function values correspond to histograms that are close in relative earthmover distance. The previous section established the existence of a feasible point with small objective function value that is close to the true histogram, hence by the triangle inequality, all such feasible points must be close to the true histogram; in particular, the optimal point—the solution to the linear program—will correspond to a histogram that is close to the true histogram of the distribution from which the sample was drawn, completing our proof of Theorem 1.11.

We define a class of earthmoving schemes, which will allow us to directly relate the relative earthmover cost of two distributions to the discrepancy in their respective fingerprint expectations. The main technical tool is a Chebyshev polynomial construction, though, for clarity, we

first describe a simpler scheme that provides some intuition for the Chebyshev construction. We begin by describing the form of our earthmoving schemes; since we hope to relate the cost of such schemes to the discrepancy in expected fingerprints of $Poi(n)$-sized samples, we will require that the schemes be formulated in terms of the Poisson functions $poi(nx, i)$.

Definition 6.4. For a given n, a *β-bump earthmoving scheme* is defined by a sequence of positive real numbers $\{c_i\}$, the *bump centers*, and a sequence of functions $\{f_i\} : (0, 1] \to \mathbb{R}$ such that $\sum_{i=0}^{\infty} f_i(x) = 1$ for each x, and each function f_i may be expressed as a linear combination of Poisson functions, $f_i(x) = \sum_{j=0}^{\infty} a_{ij} poi(nx, j)$, such that $\sum_{j=0}^{\infty} |a_{ij}| \leq \beta$.

Given a generalized histogram h, the scheme works as follows: For each x such that $h(x) \neq 0$, and each integer $i \geq 0$, move $xh(x) \cdot f_i(x)$ units of probability mass from x to c_i. We denote the histogram resulting from this scheme by $(c, f)(h)$.

Definition 6.5. A bump earthmoving scheme (c, f) is $[\epsilon, k]$-*good* if for any generalized histogram h of support size $\sum_x h(x) \leq k$, the relative earthmover distance between h and $(c, f)(h)$ is at most ϵ.

The crux of the proof of correctness of our estimator is the explicit construction of a surprisingly good earthmoving scheme. We will show that for any n and $k = \delta n \log n$ for some $\delta \in [1/\log n, 1]$, there exists an $[O(\sqrt{\delta}), k]$-good $O(n^{0.3})$-bump earthmoving scheme. In fact, we will construct a single scheme for all δ. We begin by defining a simple scheme that illustrates the key properties of a bump earthmoving scheme, and its analysis.

Perhaps the most natural bump earthmoving scheme is where the bump functions $f_i(x) = poi(nx, i)$ and the bump centers $c_i = \frac{i}{n}$. For $i = 0$, we may, for example, set $c_0 = \frac{1}{2n}$ to avoid a logarithm of 0 when evaluating relative earthmover distance. This is a valid earthmoving scheme since $\sum_{i=0}^{\infty} f_i(x) = 1$ for any x.

The motivation for this construction is the fact that, for any i, the amount of probability mass that ends up at c_i in $(c, f)(h)$ is exactly $\frac{i+1}{n}$ times the expectation of the $i + 1$st fingerprint in a $Poi(n)$-sample from h:

$$\begin{aligned}((c, f)(h))(c_i) = \sum_{x:h(x)\neq 0} h(x)x \cdot f_i(x) &= \sum_{x:h(x)\neq 0} h(x)x \cdot poi(nx, i) \\ &= \sum_{x:h(x)\neq 0} h(x) \cdot poi(nx, i+1)\frac{i+1}{n} \\ &= \frac{i+1}{n} \sum_{x:h(x)\neq 0} h(x) \cdot poi(nx, i+1).\end{aligned}$$

Consider applying this earthmoving scheme to two histograms h, g with nearly identical fingerprint expectations. Letting $h' = (c, f)(h)$ and $g' = (c, f)(g)$, by definition both h' and g' are supported at the bump centers c_i, and by the above equation, for each i, $|h'(c_i) - g'(c_i)| = \frac{i+1}{n} |\sum_x (h(x) - g(x)) poi(nx, i + 1)|$, where this expression is exactly $\frac{i+1}{n}$ times the difference between the $i + 1$st fingerprint expectations of h and g. In particular, if h and g have nearly identical fingerprint expectations, then h' and g' will be very similar. Analogs of this relation between $R((c, f)(g), (c, f)(h))$ and the discrepancy between the expected fingerprint entries corresponding to g and h will hold for any bump earthmoving scheme, (c, f). Sufficiently "good" earthmoving schemes (guaranteeing that $R(h, h')$ and $R(g, g')$ are small) thus provides a powerful way of bounding the relative earthmover distance between two distributions in terms of the discrepancy in their fingerprint expectations.

The problem with the "Poisson bump" earthmoving scheme described in the previous paragraph is that it not very "good": It incurs a very large relative earthmover cost, particularly for small

probabilities. This is due to the fact that most of the mass that starts at a probability below $\frac{1}{n}$ will end up in the zeroth bump, no matter if it has probability nearly $\frac{1}{n}$, or the rather lower $\frac{1}{k}$. Phrased differently, the problem with this scheme is that the first few "bumps" are extremely fat. The situation gets significantly better for higher Poisson functions: Most of the mass of $Poi(i)$ lies within relative distance $O(\frac{1}{\sqrt{i}})$ of i, and hence the scheme, is relatively cheap for larger probabilities $x \gg \frac{1}{n}$. We will therefore construct a scheme that uses regular Poisson functions $poi(nx, i)$ for $i \geq O(\log n)$ but takes great care to construct "skinnier" bumps below this region.

The main tool of this construction of skinnier bumps is the Chebyshev polynomials. For each integer $i \geq 0$, the ith Chebyshev polynomial, denoted $T_i(x)$, is the polynomial of degree i such that $T_i(\cos(y)) = \cos(i \cdot y)$. Thus, up to a change of variables, any linear combination of cosine functions up to frequency s may be re-expressed as the same linear combination of the Chebyshev polynomials of orders 0 through s. Given this, constructing a "good" earthmoving scheme is an exercise in trigonometric constructions.

Before formally defining our bump earthmoving scheme, we give a rough sketch of the key features. We define the scheme with respect to a parameter $s = O(\log n)$. For $i > s$, we use the fat Poisson bumps: that is, we define the bump centers $c_i = \frac{i}{n}$ and functions $f_i = poi(nx, i)$. For $i \leq s$, we will use skinnier "Chebyshev bumps"; these bumps will have roughly quadratically spaced bump centers $c_i \approx \frac{i^2}{n \log n}$, with the width of the ith bump roughly $\frac{i}{n \log n}$ (as compared to the larger width of $\frac{\sqrt{i}}{n}$ of the ith Poisson bump). At a high level, the logarithmic factor improvement in our $O(\frac{k}{\log k})$ bound on the sample size necessary to achieve accurate estimation arises, because the first few Chebyshev bumps have width $O(\frac{1}{n \log n})$, in contrast to the first Poisson bump, $poi(nx, 1)$, which has width $O(\frac{1}{n})$.

Definition 6.6. The *Chebyshev bumps* are defined in terms of n as follows. Let $s = 0.2 \log n$. Define $g_1(y) = \sum_{j=-s}^{s-1} \cos(jy)$. Define

$$g_2(y) = \frac{1}{16s}\left(g_1\left(y - \frac{3\pi}{2s}\right) + 3g_1\left(y - \frac{\pi}{2s}\right) + 3g_1\left(y + \frac{\pi}{2s}\right) + g_1\left(y + \frac{3\pi}{2s}\right)\right),$$

and, for $i \in \{1, \ldots, s-1\}$ define $g_3^i(y) := g_2(y - \frac{i\pi}{s}) + g_2(y + \frac{i\pi}{s})$, and $g_3^0 = g_2(y)$, and $g_3^s = g_2(y + \pi)$. Let $t_i(x)$ be the linear combination of Chebyshev polynomials so $t_i(\cos(y)) = g_3^i(y)$. We thus define $s+1$ functions, the "skinny bumps," to be $B_i(x) = t_i(1 - \frac{xk}{2s}) \sum_{j=0}^{s-1} poi(xk, j)$, for $i \in \{0, \ldots, s\}$. That is, $B_i(x)$ is related to $g_3^i(y)$ by the coordinate transformation $x = \frac{2s}{n}(1 - \cos(y))$, and scaling by $\sum_{j=0}^{s-1} poi(xn, j)$.

The Chebyshev bumps of Definition 6.6 are "third order"; if, instead, we had considered the analogous less skinny "second-order" bumps by defining $g_2(y) := \frac{1}{8s}(g_1(y - \frac{\pi}{s}) + 2g_1(y) + g_1(y + \frac{\pi}{s}))$, then the results would still hold, though the proofs are slightly more cumbersome.

Definition 6.7. The *Chebyshev earthmoving scheme* is defined in terms of n as follows: As in Definition 6.6, let $s = 0.2 \log n$. For $i \geq s+1$, define the ith bump function $f_i(x) = poi(nx, i-1)$ and associated bump center $c_i = \frac{i-1}{n}$. For $i \in \{0, \ldots, s\}$ let $f_i(x) = B_i(x)$, and for $i \in \{1, \ldots, s\}$, define their associated bump centers $c_i = \frac{2s}{n}(1 - \cos(\frac{i\pi}{s}))$, and let $c_0 := c_1$.

The following lemma characterizes the key properties of the Chebyshev earthmoving scheme. Namely, that the scheme is, in fact, an earthmoving scheme, that each bump can be expressed as a low-weight linear combination of Poisson functions, and that the scheme incurs a small relative-earthmover cost.

Lemma 6.8. *The Chebyshev earthmoving scheme, of Definition 6.7 has the following properties:*

- *For any $x \geq 0$,*
$$\sum_{i \geq 0} f_i(x) = 1,$$
hence the Chebyshev earthmoving scheme is a valid earthmoving scheme.
- *Each $B_i(x)$ may be expressed as $\sum_{j=0}^{\infty} a_{ij} poi(nx, j)$ for a_{ij} satisfying*
$$\sum_{j=0}^{\infty} |a_{ij}| \leq 2n^{0.3}.$$
- *The Chebyshev earthmoving scheme is $[O(\sqrt{\delta}), n]$-good, for $n = \delta n \log n$, and $\delta \geq \frac{1}{\log n}$, where the O notation hides an absolute constant factor.*

The proof of the above lemma is quite involved, and we split its proof into a series of lemmas. The first lemma below shows that the Chebyshev scheme is a valid earthmoving scheme (the first bullet in the above lemma):

Lemma 6.9. *For any x*
$$\sum_{i=-s+1}^{s} g_2\left(x + \frac{\pi i}{s}\right) = 1, \text{ and } \sum_{i=0}^{\infty} f_i(x) = 1.$$

Proof. $g_2(y)$ is a linear combination of cosines at integer frequencies j, for $j = 0, \ldots, s$, shifted by $\pm\pi/2s$ and $\pm 3\pi/s2$. Since $\sum_{i=-s+1}^{s} g_2(x + \frac{\pi i}{s})$ sums these cosines over all possible multiples of π/s, we note that all but the frequency 0 terms will cancel. The $\cos(0y) = 1$ term will show up once in each g_1 term, and thus $1 + 3 + 3 + 1 = 8$ times in each g_2 term, and thus $8 \cdot 2s$ times in the sum in question. Together with the normalizing factor of $16s$, the total sum is thus 1, as claimed.

For the second part of the claim,
$$\begin{aligned}\sum_{i=0}^{\infty} f_i(x) &= \left(\sum_{j=-s+1}^{s} g_2\left(\cos^{-1}\left(\frac{xn}{2s} - 1\right) + \frac{\pi j}{s}\right)\right)\sum_{j=0}^{s-1} poi(xn, j) + \sum_{j \geq s} poi(xn, j) \\ &= 1 \cdot \sum_{j=0}^{s-1} poi(xn, j) + \sum_{j \geq s} poi(xn, j) = 1.\end{aligned}$$

□

We now show that each Chebyshev bump may be expressed as a low-weight linear combination of Poisson functions.

Lemma 6.10. *Each $B_i(x)$ may be expressed as $\sum_{j=0}^{\infty} a_{ij} poi(nx, j)$ for a_{ij} satisfying*
$$\sum_{j=0}^{\infty} |a_{ij}| \leq 2n^{0.3}.$$

Proof. Consider decomposing $g_3^i(y)$ into a linear combination of $\cos(\ell y)$, for $\ell \in \{0, \ldots, s\}$. Since $\cos(-\ell y) = \cos(\ell y)$, $g_1(y)$ consists of one copy of $\cos(sy)$, two copies of $\cos(\ell y)$ for each ℓ between 0 and s, and one copy of $\cos(0y)$; $g_2(y)$ consists of ($\frac{1}{16s}$ times) eight copies of different $g_1(y)$'s, with some shifted to introduce sine components, but these sine components are canceled out in the formation of $g_3^i(y)$, which is a symmetric function for each i. Thus since each g_3 contains at most two g_2's, each $g_3^i(y)$ may be regarded as a linear combination $\sum_{\ell=0}^{s} \cos(\ell y) b_{i\ell}$ with the coefficients bounded as $|b_{i\ell}| \leq \frac{2}{s}$.

Since t_i was defined so $t_i(\cos(y)) = g_3^i(y) = \sum_{\ell=0}^{s} \cos(\ell y) b_{i\ell}$, by the definition of Chebyshev polynomials we have $t_i(z) = \sum_{\ell=0}^{s} T_\ell(z) b_{i\ell}$. Thus the bumps are expressed as

$$B_i(x) = \left(\sum_{\ell=0}^{s} T_\ell\left(1 - \frac{xn}{2s}\right) b_{i\ell} \right) \left(\sum_{j=0}^{s-1} poi(xn, j) \right).$$

We further express each Chebyshev polynomial via its coefficients as $T_\ell(1 - \frac{xn}{2s}) = \sum_{m=0}^{\ell} \beta_{\ell m} (1 - \frac{xn}{2s})^m$ and then expand each term via binomial expansion as $(1 - \frac{xn}{2s})^m = \sum_{q=0}^{m} (-\frac{xn}{2s})^q \binom{m}{q}$ to yield

$$B_i(x) = \sum_{\ell=0}^{s} \sum_{m=0}^{\ell} \sum_{q=0}^{m} \sum_{j=0}^{s-1} \beta_{\ell m} \left(-\frac{xn}{2s}\right)^q \binom{m}{q} b_{i\ell}\, poi(xn, j).$$

We note that in general we can reexpress $x^q\, poi(xn, j) = x^q \frac{x^j n^j e^{-xn}}{j!} = poi(xn, j+q) \frac{(j+q)!}{j! n^q}$, which finally lets us express B_i as a linear combination of Poisson functions, for all $i \in \{0, \ldots, s\}$:

$$B_i(x) = \sum_{\ell=0}^{s} \sum_{m=0}^{\ell} \sum_{q=0}^{m} \sum_{j=0}^{s-1} \beta_{\ell m} \left(-\frac{1}{2s}\right)^q \binom{m}{q} \frac{(j+q)!}{j!} b_{i\ell}\, poi(xn, j+q).$$

It remains to bound the sum of the absolute values of the coefficients of the Poisson functions. That is, by the triangle inequality, it is sufficient to show that

$$\sum_{\ell=0}^{s} \sum_{m=0}^{\ell} \sum_{q=0}^{m} \sum_{j=0}^{s-1} \left| \beta_{\ell m} \left(-\frac{1}{2s}\right)^q \binom{m}{q} \frac{(j+q)!}{j!} b_{i\ell} \right| \le 2n^{0.3}.$$

We take the sum over j first: The general fact that $\sum_{m=0}^{\ell} \binom{m+i}{i} = \binom{i+\ell+1}{()i+1}$ implies that $\sum_{j=0}^{s-1} \frac{(j+q)!}{j!} = \sum_{j=0}^{s-1} \binom{j+q}{q} q! = q! \binom{s+q}{q+1} = \frac{1}{q+1} \frac{(s+q)!}{(s-1)!}$, and, further, since $q \le m \le \ell \le s$ we have $s + q \le 2s$, which implies that this final expression is bounded as $\frac{1}{q+1} \frac{(s+q)!}{(s-1)!} = s \frac{1}{q+1} \frac{(s+q)!}{s!} \le s \cdot (2s)^q$. Thus we have

$$\begin{aligned} \sum_{\ell=0}^{s} \sum_{m=0}^{\ell} \sum_{q=0}^{m} \sum_{j=0}^{s-1} \left| \beta_{\ell m} \left(-\frac{1}{2s}\right)^q \binom{m}{q} \frac{(j+q)!}{j!} b_{i\ell} \right| &\le \sum_{\ell=0}^{s} \sum_{m=0}^{\ell} \sum_{q=0}^{m} \left| \beta_{\ell m} s \binom{m}{q} b_{i\ell} \right| \\ &= s \sum_{\ell=0}^{s} |b_{i\ell}| \sum_{m=0}^{\ell} |\beta_{\ell m}| 2^m. \end{aligned}$$

Chebyshev polynomials have coefficients whose signs repeat in the pattern $(+, 0, -, 0)$, and thus we can evaluate the innermost sum exactly as $|T_\ell(2\mathbf{i})|$, for $\mathbf{i} = \sqrt{-1}$. Since we bounded $|b_{i\ell}| \le \frac{2}{s}$ above, the quantity to be bounded is now $s \sum_{\ell=0}^{s} \frac{2}{s} |T_\ell(2\mathbf{i})|$. Since the explicit expression for Chebyshev polynomials yields $|T_\ell(2\mathbf{i})| = \frac{1}{2}[(2 - \sqrt{5})^\ell + (2 + \sqrt{5})^\ell]$ and since $|2 - \sqrt{5}|^\ell = (2 + \sqrt{5})^{-\ell}$ we finally bound $s \sum_{\ell=0}^{s} \frac{2}{s} |T_\ell(2\mathbf{i})| \le 1 + \sum_{\ell=-s}^{s} (2 + \sqrt{5})^\ell < 1 + \frac{2+\sqrt{5}}{2+\sqrt{5}-1} \cdot (2 + \sqrt{5})^s < 2 \cdot (2 + \sqrt{5})^s < 2 \cdot k^{0.3}$, as desired, since $s = 0.2 \log n$ and $\log(2 + \sqrt{5}) < 1.5$ and $0.2 \cdot 1.5 = 0.3$. □

We now turn to the main thrust of Lemma 6.8, showing that the scheme is $[O(\sqrt{\delta}), k]$-good, where $k = \delta n \log n$, and $\delta \ge \frac{1}{\log n}$; the following lemma, quantifying the "skinnyness" of the Chebyshev bumps is the cornerstone of this argument.

LEMMA 6.11. $|g_2(y)| \le \frac{\pi^7}{y^4 s^4}$ *for* $y \in [-\pi, \pi] \setminus (-3\pi/s, 3\pi/s)$, *and* $|g_2(y)| \le 1/2$ *everywhere.*

PROOF. Since $g_1(y) = \sum_{j=-s}^{s-1} \cos jy = \sin(sy)\cot(y/2)$, and since $\sin(\alpha+\pi) = -\sin(\alpha)$, we have the following:

$$\begin{aligned} g_2(y) &= \frac{1}{16s}\left(g_1\left(y - \frac{3\pi}{2s}\right) + 3g_1\left(y - \frac{\pi}{2s}\right) + 3g_1\left(y + \frac{\pi}{2s}\right) + g_1\left(y + \frac{3\pi}{2s}\right)\right) \\ &= \frac{1}{16s}\left(\sin(ys+\pi/2)\left(\cot\left(\frac{y}{2} - \frac{3\pi}{4s}\right) - 3\cot\left(\frac{y}{2} - \frac{\pi}{4s}\right)\right.\right. \\ &\qquad \left.\left. +3\cot\left(\frac{y}{2} + \frac{\pi}{4s}\right) - \cot\left(\frac{y}{2} + \frac{3\pi}{4s}\right)\right)\right). \end{aligned}$$

Note that $(\cot(\frac{y}{2} - \frac{3\pi}{4s}) - 3\cot(\frac{y}{2} - \frac{\pi}{4s}) + 3\cot(\frac{y}{2} + \frac{\pi}{4s}) - \cot(\frac{y}{2} + \frac{3\pi}{4s}))$ is a discrete approximation to $(\pi/2s)^3$ times the third derivative of the cotangent function evaluated at $y/2$. Thus it is bounded in magnitude by $(\pi/2s)^3$ times the maximum magnitude of $\frac{d^3}{dx^3}\cot(x)$ in the range $x \in [\frac{y}{2} - \frac{3\pi}{4s}, \frac{y}{2} + \frac{3\pi}{4s}]$. Since the magnitude of this third derivative is decreasing for $x \in (0, \pi)$, we can simply evaluate the magnitude of this derivative at $\frac{y}{2} - \frac{3\pi}{4s}$. We thus have $\frac{d^3}{dx^3}\cot(x) = \frac{-2(2+\cos(2x))}{\sin^4(x)}$, whose magnitude is at most $\frac{6}{(2x/\pi)^4}$ for $x \in (0, \pi]$. For $y \in [3\pi/s, \pi]$, we trivially have that $\frac{y}{2} - \frac{3\pi}{4s} \ge \frac{y}{4}$, and thus we have the following bound:

$$\left|\cot\left(\frac{y}{2} - \frac{3\pi}{4s}\right) - 3\cot\left(\frac{y}{2} - \frac{\pi}{4s}\right) + 3\cot\left(\frac{y}{2} + \frac{\pi}{4s}\right) - \cot\left(\frac{y}{2} + \frac{3\pi}{4s}\right)\right| \le \left(\frac{\pi}{2s}\right)^3 \frac{6}{(y/2\pi)^4} \le \frac{12\pi^7}{y^4 s^3}.$$

Since $g_2(y)$ is a symmetric function, the same bound holds for $y \in [-\pi, -3\pi/s]$. Thus $|g_2(y)| \le \frac{12\pi^7}{16s \cdot y^4 s^3} < \frac{\pi^7}{y^4 s^4}$ for $y \in [-\pi, \pi] \setminus (-3\pi/s, 3\pi/s)$. To conclude, note that $g_2(y)$ attains a global maximum at $y = 0$, with $g_2(0) = \frac{1}{16s}(6\cot(\pi/4s) - 2\cot(3\pi/4s)) \le \frac{1}{16s}\frac{24s}{\pi} < 1/2$. □

LEMMA 6.12. *The Chebyshev earthmoving scheme of Definition 6.7 is* $[O(\sqrt{\delta}), k]$*-good, where* $k = \delta n \log n$, *and* $\delta \ge \frac{1}{\log n}$.

PROOF. We split this proof into two parts: first we will consider the cost of the portion of the scheme associated with all but the first $s+1$ bumps, and then we consider the cost of the skinny bumps f_i with $i \in \{0, \ldots, s\}$.

For the first part, we consider the cost of bumps f_i for $i \ge s+1$; that is the relative earthmover cost of moving $poi(xn, i)$ mass from x to $\frac{i}{n}$, summed over $i \ge s$. By definition of relative earthmover distance, the cost of moving mass from x to $\frac{i}{n}$ is $|\log \frac{xn}{i}|$, which, since $\log y \le y - 1$, we bound by $\frac{xn}{i} - 1$ when $i < xn$ and $\frac{i}{xn} - 1$ otherwise. We thus split the sum into two parts.

For $i \ge \lceil xn \rceil$, we have $poi(xn, i)(\frac{i}{xn} - 1) = poi(xn, i-1) - poi(xn, i)$. This expression telescopes when summed over $i \ge \max\{s, \lceil xn \rceil\}$ to yield $poi(xn, \max\{s, \lceil xn \rceil\} - 1) = O(\frac{1}{\sqrt{s}})$.

For $i \le \lceil xn \rceil - 1$, we have, since $i \ge s$, that $poi(xn, i)(\frac{xn}{i} - 1) \le poi(xn, i)((1 + \frac{1}{s})\frac{xn}{i+1} - 1) = (1 + \frac{1}{s})poi(xn, i+1) - poi(xn, i)$. The $\frac{1}{s}$ term sums to at most $\frac{1}{s}$, and the rest telescopes to $poi(xn, \lceil xn \rceil) - poi(xn, s) = O(\frac{1}{\sqrt{s}})$. Thus in total, f_i for $i \ge s+1$ contributes $O(\frac{1}{\sqrt{s}})$ to the relative earthmover cost, per unit of weight moved.

We now turn to the skinny bumps $f_i(x)$ for $i \le s$. The simplest case is when x is outside the region that corresponds to the cosine of a real number—that is, when $xn \ge 4s$. It is straightforward to show that $f_i(x)$ is very small in this region. We note the general expression for Chebyshev polynomials: $T_j(x) = \frac{1}{2}[(x - \sqrt{x^2-1})^j + (x + \sqrt{x^2-1})^j]$, whose magnitude we bound by $|2x|^j$. Further, since $2x \le \frac{2}{e}e^x$, we bound this by $(\frac{2}{e})^j e^{|x|j}$, which we apply when $|x| > 1$. Recall the definition $f_i(x) = t_i(1 - \frac{xn}{2s})\sum_{j=0}^{s-1} poi(xn, j)$, where t_i is the polynomial defined so $t_i(\cos(y)) = g_3^i(y)$,

that is, t_i is a linear combination of Chebyshev polynomials of degree at most s and with coefficients summing in magnitude to at most 2, as was shown in the proof of Lemma 6.10. Since $xn > s$, we may bound $\sum_{j=0}^{s-1} poi(xn, j) \le s \cdot poi(xn, s)$. Further, since $z \le e^{z-1}$ for all z, letting $z = \frac{x}{4s}$ yields $x \le 4s \cdot e^{\frac{x}{4s}-1}$, from which we may bound $poi(xn, s) = \frac{(xn)^s e^{-xn}}{s!} \le \frac{e^{-xn}}{s!}(4s \cdot e^{\frac{xn}{4s}-1})^s = \frac{4^s s^s}{e^s \cdot e^{3xn/4} s!} \le 4^s e^{-3xn/4}$. We combine this with the above bound on the magnitude of Chebyshev polynomials, $T_j(z) \le (\frac{2}{e})^j e^{|z|j} \le (\frac{2}{e})^s e^{|z|s}$, where $z = (1 - \frac{xn}{2s})$ yields $T_j(z) \le (\frac{2}{e^2})^s e^{\frac{xn}{2}}$. Thus $f_i(x) \le poly(s) 4^s e^{-3xn/4} (\frac{2}{e^2})^s e^{\frac{xn}{2}} = poly(s)(\frac{8}{e^2})^s e^{-\frac{xn}{4}}$. Since $\frac{xn}{4} \ge s$ in this case, f_i is exponentially small in both x and s; the total cost of this earthmoving scheme, per unit of mass above $\frac{4s}{n}$ is obtained by multiplying this by the logarithmic relative distance the mass has to move, and summing over the $s + 1$ values of $i \le s$, and thus remains exponentially small, and is thus trivially bounded by $O(\frac{1}{\sqrt{s}})$.

To bound the cost in the remaining case, when $xn \le 4s$ and $i \le s$, we work with the trigonometric functions g_3^i, instead of t_i directly. For $y \in (0, \pi]$, we seek to bound the per-unit-mass relative earthmover cost of, for each $i \ge 0$, moving $g_3^i(y)$ mass from $\frac{2s}{n}(1 - \cos(y))$ to c_i. (Recall from Definition 6.7 that $c_i = \frac{2s}{n}(1 - \cos(\frac{i\pi}{s}))$ for $i \in \{1, \ldots, s\}$, and $c_0 = c_1$.) For $i \ge 1$, this contribution is at most

$$\sum_{i=1}^{s} \left| g_3^i(y)(\log(1 - \cos(y)) - \log\left(1 - \cos\left(\frac{i\pi}{s}\right)\right)\right|.$$

We analyze this expression by first showing that for any $x, x' \in (0, \pi]$,

$$|\log(1 - \cos(x)) - \log(1 - \cos(x'))| \le 2|\log x - \log x'|.$$

Indeed, this holds because the derivative of $\log(1 - cos(x))$ is positive, and strictly less than the derivative of $2 \log x$; this can be seen by noting that the respective derivatives are $\frac{\sin(y)}{1-\cos(y)}$ and $\frac{2}{y}$, and we claim that the second expression is always greater. To compare the two expressions, cross-multiply and take the difference, to yield $y \sin y - 2 + 2 \cos y$, which we show is always at most 0 by noting that it is 0 when $y = 0$ and has derivative $y \cos y - \sin y$, which is negative since $y < \tan y$. Thus we have that $|\log(1 - \cos(y)) - \log(1 - \cos(\frac{i\pi}{s}))| \le 2|\log y - \log \frac{i\pi}{s}|$; we use this bound in all but the last step of the analysis. Additionally, we ignore the $\sum_{j=0}^{s-1} poi(xn, j)$ term as it is always at most 1.

Case 1: $y \ge \frac{\pi}{s}$.

We will show that

$$\left| g_3^0(y)\left(\log y - \log \frac{\pi}{s}\right)\right| + \sum_{i=1}^{s} \left| g_3^i(y)\left(\log y - \log \frac{i\pi}{s}\right)\right| = O\left(\frac{1}{sy}\right),$$

where the first term is the contribution from f_0, c_0. For i such that $y \in (\frac{(i-3)\pi}{s}, \frac{(i+3)\pi}{s})$, by the second bounds on $|g_2|$ in the statement of Lemma 6.11, $g_3^i(y) < 1$, and for each of the at most 6 such i, $|(\log y - \log \frac{\max\{1,i\}\pi}{s})| < \frac{1}{sy}$, to yield a contribution of $O(\frac{1}{sy})$. For the contribution from i such that $y \le \frac{(i-3)\pi}{s}$ or $y \ge \frac{(i-3)\pi}{s}$, the first bound of Lemma 6.11 yields $|g_3^i(y)| = O(\frac{1}{(ys - i\pi)^4})$. Roughly, the bound will follow from noting that this sum of inverse fourth powers is dominated by the first few terms. Formally, we split up our sum over $i \in [s] \setminus [\frac{ys}{\pi} - 3, \frac{ys}{\pi} + 3]$ into two parts according to whether $i > ys/\pi$:

$$\sum_{i\geq\frac{ys}{\pi}+3}^{s}\frac{1}{(ys-i\pi)^4}\left|\left(\log y-\log\frac{i\pi}{s}\right)\right|\leq\sum_{i\geq\frac{ys}{\pi}+3}^{\infty}\frac{\pi^4}{(\frac{ys}{\pi}-i)^4}\left(\log i-\log\frac{ys}{\pi}\right)$$
$$\leq\pi^4\int_{w=\frac{ys}{\pi}+2}^{\infty}\frac{1}{(\frac{ys}{\pi}-w)^4}\left(\log w-\log\frac{ys}{\pi}\right). \tag{2}$$

Since the antiderivative of $\frac{1}{(\alpha-w)^4}(\log w-\log\alpha)$ with respect to w is

$$\frac{-2w(w^2-3w\alpha+3\alpha^2)\log w+2(w-\alpha)^3\log(w-\alpha)+\alpha(2w^2-5w\alpha+3\alpha^2+2\alpha^2\log\alpha)}{6(w-\alpha)^3\alpha^3},$$

the quantity in Equation (2) is equal to the above expression evaluated with $\alpha=\frac{ys}{\pi}$, and $w=\alpha+2$, to yield

$$O\left(\frac{1}{ys}\right)-\log\frac{ys}{\pi}+\log\left(2+\frac{ys}{\pi}\right)=O\left(\frac{1}{ys}\right).$$

A nearly identical argument applies to the portion of the sum for $i\leq\frac{ys}{\pi}+3$, yielding the same asymptotic bound of $O(\frac{1}{ys})$.

Case 2: $\frac{ys}{\pi}<1$.

The per-unit mass contribution from the 0th bump is trivially at most $|g_3^0(y)(\log\frac{ys}{\pi}-\log 1)|\leq\log\frac{ys}{\pi}$. The remaining relative earthmover cost is $\sum_{i=1}^{s}|g_3^i(y)(\log\frac{ys}{\pi}-\log i)|$. To bound this sum, we note that $\log i\geq 0$, and $\log\frac{ys}{\pi}\leq 0$ in this region, and thus split the above sum into the corresponding two parts, and bound them separately. By Lemma 6.11, we have:

$$\sum_{i=1}^{s}g_3^i(y)\log i\leq O\left(1+\sum_{i=3}^{\infty}\frac{\log i}{\pi^4(i-1)^4}\right)=O(1).$$

$$\sum_{i=1}^{s}g_3^i(y)\log\frac{ys}{\pi}\leq O\left(\log ys\right)\leq O\left(\frac{1}{ys}\right),$$

since for $ys\leq\pi$, we have $|\log ys|<4/ys$.

Having concluded the case analysis, recall that we have been using the change of variables $x=\frac{2s}{n}(1-\cos(y))$. Since $1-\cos(y)=O(y^2)$, we have $xn=O(sy^2)$. Thus the case analysis yielded a bound of $O(\frac{1}{ys})$, which we may thus express as $O(\frac{1}{\sqrt{sxn}})$.

For a distribution with histogram h, the cost of moving earth on this region, for bumps f_i where $i\leq s$ is thus

$$O\left(\sum_{x:h(x)\neq 0}h(x)\cdot x\cdot\frac{1}{\sqrt{sxn}}\right)=O\left(\frac{1}{\sqrt{sn}}\sum_{x:h(x)\neq 0}h(x)\sqrt{x}\right).$$

Since $\sum_x x\cdot h(x,y)=1$, and $\sum_x h(x)\leq n$, by the Cauchy-Schwarz inequality,

$$\sum_x\sqrt{x}h(x)=\sum_x\sqrt{x\cdot h(x)}\sqrt{h(x)}\leq\sqrt{n},$$

and hence since $k=\delta n\log n$, the contribution to the cost of these bumps is bounded by $O(\sqrt{\frac{k}{sn}})=O(\sqrt{\delta})$. As we have already bounded the relative earthmover cost for bumps f_i for $i>s$ at least this tightly, this concludes the proof. □

We are now equipped to assemble the pieces and prove Theorem 1.11.

Proof of Theorem 1.11. Let g be the generalized histogram returned by Algorithm 2, and let h be the generalized histogram constructed in Lemma 6.3—assuming the sample from the true

distribution p is "faithful," which occurs with probability $1 - e^{-n^{\Omega(1)}}$ by Lemma 6.2. Lemma 6.3 asserts that $R(p,h) = O(\frac{1}{n^{\Omega(1)}})$. Let h', g' be the generalized histograms that result from applying the Chebyshev earthmoving scheme of Definition 6.7 to h and g, respectively. By Lemma 6.8, $R(h,h') = O(\sqrt{1/c})$, and $R(g,g') = O(\sqrt{1/c})$. Our goal is to bound $R(p,g)$, which we do via the triangle inequality as

$$R(p,g) \le R(p,h) + R(h,h') + R(h',g') + R(g',g) = O(\sqrt{1/c}) + R(g',h').$$

We now show that $R(g',h') = O(\frac{1}{n^{\Omega(1)}})$, completing the proof.

Our strategy to bound this relative earthmover distance is to construct an earthmoving scheme that equates g' and h' whose cost can be related to the terms of the first constraint of the linear program. By definition, g', h' are generalized histograms supported at the bump centers c_i. Our earthmoving scheme is defined as follows: for each $i \notin [n^{\mathcal{B}}, n^{\mathcal{B}} + 2n^C]$, if $h'(c_i) > g'(c_i)$, then we move $c_i\,(h'(c_i) - g'(c_i))$ units of probability mass in h' from location c_i to location $\frac{n^{\mathcal{B}}}{n}$; analogously, if $h'(c_i) < g'(c_i)$, then we move $c_i\,(g'(c_i) - h'(c_i))$ units of probability mass in g' from location c_i to location $\frac{n^{\mathcal{B}}}{n}$. After performing this operation, the remaining discrepancy in the resulting histograms will be confined to probability range $[\frac{n^{\mathcal{B}}}{n}, \frac{n^{\mathcal{B}}+2n^C}{n}]$, and hence can be equated at an additional cost of at most

$$\log \frac{n^{\mathcal{B}} + 2n^C}{n^{\mathcal{B}}} = O(n^{C-\mathcal{B}}) = O\left(\frac{1}{n^{\Omega(1)}}\right).$$

We now analyze the relative earthmover cost of equalizing $h'(c_i)$ and $g'(c_i)$ for all $i \notin [n^{\mathcal{B}}, n^{\mathcal{B}} + 2n^C]$ by moving the discrepancy to location $\frac{n^{\mathcal{B}}}{n}$. Since all but the first $s+1$ bumps are simply the standard Poisson bumps $f_i(x) = poi(xn, i-1)$, for $i > s$ we have

$$\begin{aligned} |h'(c_i) - g'(c_i)| &= \left| \sum_{x: h(x)+g(x)\neq 0} (h(x) - g(x)) x \cdot poi(nx, i-1) \right| \\ &= \left| \sum_{x: h(x)+g(x)\neq 0} (h(x) - g(x)) poi(nx, i) \frac{i}{k} \right|. \end{aligned}$$

Recall by construction that $h(x) = g(x)$ for all $x > \frac{n^{\mathcal{B}}+n^C}{n}$. Thus by tail bounds for Poissons, the total relative earthmover cost of equalizing h' and g' for all bump centers c_i with $i > n^{\mathcal{B}} + 2n^C$ is trivially bounded by $o(\frac{\log n}{n})$.

Next, we consider the contribution of the discrepancies in the Poisson bumps with centers c_i for $i \in [s+1, n^{\mathcal{B}}]$. Since $\sum_{i \le n^{\mathcal{B}}} poi(nx, i) = o(1/n^2)$ for $x \ge \frac{n^{\mathcal{B}}+n^C}{n}$, the discrepancy in the first $n^{\mathcal{B}}$ expected fingerprints of g, h is specified, up to negligible error, by the terms in the first constraint of the linear program:

$$\begin{aligned} &\sum_{i<n^{\mathcal{B}}} \left| \sum_{x: h(x)+g(x)\neq 0} (h(x) - g(x)) poi(nx, i) \frac{i}{n} \right| \\ &\le \sum_{i<n^{\mathcal{B}}} \frac{i}{n} \cdot \frac{\sqrt{n+1}}{\sqrt{\mathcal{F}_i + 1}} \left(\left| \mathcal{F}_i - \sum_{x: g(x)\neq 0} g(x) poi(nx, i) \right| + \left| \mathcal{F}_i - \sum_{x: h(x)\neq 0} h(x) poi(nx, i) \right| \right) \\ &\le O(n^{3\mathcal{B}-1/2}) \quad = \quad O\left(\frac{1}{n^{\Omega(1)}}\right). \end{aligned}$$

Finally, we consider the contribution of the discrepancies in the first $s+1 = O(\log n)$ bump centers, corresponding to the skinny Chebyshev bumps. Note that for such centers, c_i, the corresponding bump functions $f_i(x)$ are expressible by definition as $f_i(x) = \sum_{j\geq 0} a_{ij} poi(xn, j)$, for some coefficients a_{ij}, where $\sum_{j\geq 0} a_{ij} \leq \beta$. Thus we have the following, where $\sum_x$ is shorthand for $\sum_{x:h(x)+g(x)\neq 0}$:

$$\begin{aligned}
|h'(c_i) - g'(c_i)| &= \left|\sum_x (h(x) - g(x))x f_i(x)\right| \\
&= \left|\sum_x (h(x) - g(x))x \sum_{j\geq 0} a_{ij} poi(xn, j)\right| \\
&= \left|\sum_{j\geq 0} a_{ij} \sum_x (h(x) - g(x))x poi(xn, j)\right| \\
&= \left|\sum_{j\geq 1} a_{i,j-1} \frac{j}{n} \sum_x (h(x) - g(x)) poi(xn, j)\right|.
\end{aligned}$$

Since $a_{ij} = 0$ for $j > \log n$, and since each Chebyshev bump is a linear combination of only the first $2s < \log n$ Poisson functions, the total cost of equalizing h' and g' at each of these Chebyshev bump centers is bounded as

$$\beta \left|\sum_{i=1}^{\log n} \frac{i}{n} \sum_x (h(x) - g(x)) poi(xn, j)\right| |\log c_0| \log n,$$

where the $|\log c_0|$ term, for c_0 being the first bump center, is a crude upper bound on the per-unit mass relative earthmover cost of moving the mass to probability $\frac{n^{\mathcal{B}}}{n}$, and the final factor of $\log n$ is because there are at most $s < \log n$ centers corresponding to "skinny" bumps. We bound this via the triangle inequality and an appeal to the first constraint of the linear program, as above, yielding a bound of $O(\beta n^{2\mathcal{B}} \frac{\log^3 n}{\sqrt{n}})$. Since $\beta = O(n^{0.3})$ from Lemma 6.8, this contribution is thus also $O(\frac{1}{n^{\Omega(1)}})$.

We have thus bounded all the parts of $R(g', h')$ by $O(\frac{1}{n^{\Omega(1)}})$, completing the proof. □

We note that what we actually proved applies rather more generally than to just Linear Program 4. As long as the second and third constraints are satisfied, then if the left-hand side of the first constraint, and the objective function are *somewhat* small, similar results hold.

Proposition 6.13. *For any $c > 0$, for sufficiently large k, given the fingerprint $\mathcal{F}$ from a "faithful" sample of size $n = c\frac{k}{\log k}$ from a distribution $p \in \mathcal{D}^k$, consider any vector v_x indexed by elements $x \in X := \{\frac{1}{n^2}, \frac{2}{n^2}, \frac{3}{n^2}, \ldots, \frac{n^{\mathcal{B}}+n^C}{n}\}$ such that*

- $\sum_{x\in X} x \cdot v_x + \sum_{i=n^{\mathcal{B}}+2n^C}^{n} \frac{i}{n}\mathcal{F}_i = 1$
- $\forall x \in X, v_x \geq 0$

Let $A := \sum_{x\in X} v_x$, and let $B := \sum_{i=1}^{n^{\mathcal{B}}} \frac{1}{\sqrt{\mathcal{F}_i+1}} |\mathcal{F}_i - \sum_{x\in X} poi(nx, i)v_x|$.

Appending the high-frequency portion of $\mathcal{F}$ to v as in Algorithm 2, returns a histogram g_{LP} such that

$$R(p, g_{LP}) \leq O\left(\frac{1}{\sqrt{c}} + \sqrt{\frac{A}{n\log n}} + \frac{B\log^3 n}{n^{0.2}}\right).$$

This implies, for example, that the results of Theorem 1.11 hold even when the right-hand side of the first constraint of Linear Program 4 is increased by any constant factor, and, instead of optimizing the objective function, any point with objective less than a constant multiple of k is chosen. (Of course, in practice one usually does not know k—the support size of the unknown distribution—so minimizing the objective function is a natural way to guarantee this criterion.)

REFERENCES

[1] J. Acharya, H. Das, A. Orlitsky, and S. Pan. 2011. Competitive closeness testing. In *Proceedings of the Conference on Learning Theory (COLT'11)*.

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

[3] A. Antos and I. Kontoyiannis. 2001. Convergence properties of functional estimates for discrete distributions. *Random Struct. Algor.* 19, 3–4 (2001), 163–193.

[4] Z. Bar-Yossef, R. Kumar, and D. Sivakumar. 2001. Sampling algorithms: Lower bounds and applications. In *Proceedings of the Symposium on Theory of Computing (STOC'01)*.

[5] G. P. Basharin. 1959. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory Probab. Appl.* 4, 3 (1959), 333–336.

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

[7] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. 2002. The complexity of approximating the entropy. In *Proceedings of the Symposium on Theory of Computing (STOC'02)*.

[8] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld. 2005. The complexity of approximating the entropy. *SIAM J. Comput.* 35, 1 (2005), 132–150.

[9] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White. 2000. Testing that distributions are close. In *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS'00)*.

[10] M. Brautbar and A. Samorodnitsky. 2007. Approximating entropy from sublinear samples. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA07)*.

[11] J. Bunge. Bibliography of references on the problem of estimating support size. Retrieved from http://www.stat.cornell.edu/~bunge/bibliography.html.

[12] J. Bunge and M. Fitzpatrick. 1993. Estimating the number of species: A review. *J. Am. Statist. Assoc.* 88, 421 (1993), 364–373.

[13] A. Chao and T. J. Shen. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.* 10 (2003), 429–443.

[14] M. Charikar, S. Chaudhuri, R. Motwani, and V. R. Narasayya. 2000. Towards estimation error guarantees for distinct values. In *Proceedings of the Symposium on Principles of Database Systems (PODS'00)*.

[15] B. Efron and C. Stein. 1981. The jacknife estimate of variance. *Ann. Stat.* 9 (1981), 586–596.

[16] B. Efron and R. Thisted. 1976. Estimating the number of unseen species: How many words did Shakespeare know?*Biometrika* 63, 3 (1976), 435–447.

[17] R. A. Fisher, A. Corbet, and C. B. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Br. Ecol. Soc.* 12, 1 (1943), 42–58.

[18] I. J. Good and G. H. Toulmin. 1956. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* 43 (1956), 45–63.

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

[20] S. Guha, A. McGregor, and S. Venkatasubramanian. 2006. Streaming and sublinear approximation of entropy and information distances. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA'06)*.

[21] P. J. Haas, J. F. Naughton, S. Seshadri, and A. N. Swami. 1996. Selectivity and cost estimation for joins based on random sampling. *J. Comput. System Sci.* 52, 3 (1996), 550–569.

[22] D. G. Horvitz and D. J. Thompson. 1952. A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc.* 47, 260 (1952), 663–685.

[23] J. Jiao, K. Venkat, Y. Han, and T. Weissman. 2015. Minimax estimation of functionals of discrete distributions. *IEEE Trans. Inf. Theory* 61, 5 (2015), 2835–2885.

[24] L. Kantorovich and G. Rubinstein. 1957. On a functional space and certain extremal problems. *Dokl. Akad. Nauk. SSSR (Russ.)* 115 (1957), 1058–1061.

[25] A. Keinan and A. G. Clark. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336, 6082 (2012), 740–743.

[26] D. A. McAllester and R. E. Schapire. 2000. On the convergence rate of good-Turing estimators. In *Proceedings of the Conference on Learning Theory (COLT'00)*.
[27] G. Miller. 1955. Note on the bias of information estimates. *Inf. Theory Psychol.: Probl. Methods* (1955), 95–100.
[28] M. Mitzenmacher and E. Upfal. 2005. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press.
[29] M. R. Nelson and D. Wegmann et al. 2012. An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 337, 6090 (2012), 100–104.
[30] F. Olken and D. Rotem. 1990. Random sampling from database files: A survey. In *Proceedings of the 5th International Workshop on Statistical and Scientific Data Management*.
[31] A. Orlitsky, N. Santhanam, K. Viswanathan, and J. Zhang. 2004. On modeling profiles instead of values. In *Proceedings of the Conference on Uncertainity in Artificial Intelligence (UAI'04)*. 426–435.
[32] A. Orlitsky, N. P. Santhanam, and J. Zhang. 2003. Always good turing: Asymptotically optimal probability estimation. *Science* 302, 5644 (October 2003), 427–431.
[33] A. Orlitsky, N. P. Santhanam, and J. Zhang. 2003. Always good turing: Asymptotically optimal probability estimation. In *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS'03)*.
[34] Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. 2016. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences* 113, 47 (2016), 13283–13288.
[35] L. Paninski. 2003. Estimation of entropy and mutual information. *Neur. Comput.* 15, 6 (2003), 1191–1253.
[36] S. Panzeri and A. Treves. 1996. Analytical estimates of limited sampling biases in different information measures. *Network* 7 (1996), 87–107.
[37] H. E. Robbins. 1968. Estimating the total probability of the unobserved outcomes of an experiment. *Ann. Math. Stat.* 39, 1 (1968), 256–257.
[38] J. A. Tennessen, A. W. Bigham, and T. D. O'Connor et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337, 6090 (2012), 64–69.
[39] G. Valiant and P. Valiant. 2010. A CLT and tight lower bounds for estimating entropy. (2010). Retrieved from http://www.eccc.uni-trier.de/report/2010/179/.
[40] G. Valiant and P. Valiant. 2011. Estimating the unseen: An $n/\log(n)$–sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the ACM Symposium on Theory of Computing (STOC'11)*.
[41] G. Valiant and P. Valiant. 2011. The power of linear estimators. In *Proceedings of the IEEE Symposium on Foundations of Computer Science (FOCS'11)*.
[42] G. Valiant and P. Valiant. 2013. Estimating the unseen: Improved estimators for entropy and other properties. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS'13)*.
[43] Gregory Valiant and Paul Valiant. 2016. Instance optimal learning of discrete distributions. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing (STOC'16)*. ACM, New York, NY, 142–155. DOI: http://dx.doi.org/10.1145/2897518.2897641
[44] P. Valiant. 2008. Testing symmetric properties of distributions. In *Proceedings of the Symposium on Theory of Computing (STOC'08)*.
[45] V. Q. Vu, B. Yu, and R. E. Kass. 2007. Coverage-adjusted entropy estimation. *Stat. Med.* 26, 21 (2007), 4039–4060.
[46] A. B. Wagner, P. Viswanath, and S. R. Kulkami. 2006. Strong consistency of the Good-Turing estimator. In *Proceedings of the IEEE Symposium on Information Theory*.
[47] A. B. Wagner, P. Viswanath, and S. R. Kulkami. 2007. A better good-Turing estimator for sequence probabilities. In *Proceedings of the IEEE Symposium on Information Theory*.
[48] Y. Wu and P. Yang. 2016. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Trans. Inf. Theory* 62, 6 (2016), 3702–3720.
[49] S. Zahl. 1977. Jackknifing an index of diversity. *Ecology* 58 (1977), 907–913.
[50] James Zou, Gregory Valiant, Paul Valiant, Konrad Karczewski, Siu On Chan, Kaitlin Samocha, Monkol Lek, Shamil Sunyaev, Mark Daly, and Daniel G. MacArthur. 2016. Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nat. Commun.* 7 (2016).

Received May 2015; revised July 2017; accepted July 2017