The Power of Linear Estimators

Gregory Valiant Paul Valiant

June 10, 2011

Abstract

For a broad class of practically relevant distribution properties, which includes entropy and support size, nearly all of the proposed estimators have an especially simple form. Given a set of independent samples from a discrete distribution, these estimators tally the vector of summary statistics—the number of domain elements seen once, twice, etc. in the sample—and output the dot product between these summary statistics, and a fixed vector of coefficients. We term such estimators *linear*. This historical proclivity towards linear estimators is slightly perplexing, since, despite many efforts over nearly 60 years, all proposed such estimators have significantly suboptimal convergence, compared to the bounds shown in [32, 33].

Our main result, in some sense vindicating this insistence on linear estimators, is that for any property in this broad class, there exists a near-optimal linear estimator. Additionally, we give a practical and polynomial-time algorithm for constructing such estimators for any given parameters.

While this result does not yield explicit bounds on the sample complexities of these estimation tasks, we leverage the insights provided by this result, to give explicit constructions of near-optimal linear estimators for three properties: entropy, L_1 distance to uniformity, and for pairs of distributions, L_1 distance.

Our entropy estimator, when given $O(\frac{n}{\epsilon \log n})$ independent samples from a distribution of support at most n, will estimate the entropy of the distribution to within accuracy ϵ, with probability of failure $o(1/poly(n))$. From the recent lower bounds given in [32, 33], this estimator is optimal, to constant factor, both in its dependence on n, and its dependence on ϵ. In particular, the inverse-linear convergence rate of this estimator resolves the main open question of [32, 34], which left open the possibility that the error decreased only with the square root of the number of samples.

Our distance to uniformity estimator, when given $O(\frac{m}{\epsilon^2 \log m})$ independent samples from any distribution, returns an ϵ-accurate estimate of the L_1 distance to the uniform distribution of support m. This is the first sublinear-sample estimator for this problem, and is constant-factor optimal, for constant ϵ.

Finally, our framework extends naturally to properties of pairs of distributions, including estimating the L_1 distance and KL-divergence between pairs of distributions. We give an explicit linear estimator for estimating L_1 distance to accuracy ϵ using $O(\frac{n}{\epsilon^2 \log n})$ samples from each distribution, which is constant-factor optimal, for constant ϵ.

1 Introduction

Our algorithmic toolbox is large. Given independent samples from a distribution, one might imagine a wide gamut of algorithmic strategies for recovering information about the underlying distribution. When limited by data instead of computational resources, a brute-force search through hypotheses might be the best option. More specifically, one might be guided by a Bayesian heuristic, or otherwise try to optimize "likelihood". More firmly in the realm of polynomial-time algorithms, convex programming is a powerful tool for rapidly traversing a sufficiently structured search space. At the far extreme of simplicity, are *linear estimators*. Given a vector of summary statistics of the samples, a linear estimator multiplies each entry by a fixed, position-dependent constant and returns the sum.

For the broad and practically relevant class of "symmetric" distribution properties—which includes entropy, support size, distance to uniformity, and for pairs of distributions, such distance metrics as L_1 distance and KL-divergence—despite the plethora of algorithmic options and a rich history of study by both the statistics and Computer Science communities, nearly all the proposed estimators are these algorithmically-hollow linear estimators.

Because of, or perhaps despite, their rather pedestrian nature, linear estimators have many features to recommend: they are easy to use, easy to describe, and, because of the especially transparent fashion in which they use the data, generally easy to analyze. These niceties though make it even more urgent to resolve the question: "*How good are linear estimators?*"

Despite much effort constructing linear estimators during the past century, and perhaps even more effort analyzing these estimators, for many symmetric distribution properties the best known linear estimators require many more samples than necessary to achieve a desired accuracy of estimation. Specifically, to achieve constant additive error (with high probability) for any of the following properties: entropy, distinct elements, L_1 distance and KL-divergence, existing linear estimators require $\Theta(n)$ samples, where n is a bound on the support size of the distributions being sampled, and is a natural parameterization of the sample complexities of these estimation problems. Corresponding statements hold for estimating support size and distance to uniformity, for which the sample complexities are parameterized slightly differently.[1]

Can one do any better? Yes. Recently, in a break from traditional approaches, we applied the algorithmic power of linear programming to these estimation tasks, yielding estimators for entropy and support size that require only $O(n/\log n)$ samples [33, 34]. This intriguing state of affairs provokes the question:

What richness of algorithmic machinery is needed to effectively estimate these properties?

Answers to this question could serve to guide future endeavors to construct and analyze estimators. Additionally, questions of this nature lie at the philosophical core of the theoretical approach to computing.

The main result of this paper is the near-optimality of linear estimators for additively estimating a subclass of symmetric distribution properties that includes entropy, variants of distance to uniformity, and support size (which may be viewed as a version of the distinct elements problem). Our proof is constructive, in that we give a relatively practical and polynomial-time algorithm which, on input n, k, and the property in question, outputs a linear estimator which, on input k independent samples from a distribution of support at most n, will almost surely return an ϵ-accurate approximation of the property value; this estimator is near-optimal in the sense that there exist $k' = k(1 - o(1))$, and $\epsilon' = \epsilon(1 - o(1))$ and two distributions of support at most n whose property values differ by ϵ', yet which cannot be distinguished given sets of k' samples, with any fixed probability greater than $1/2$.

[1] The problem of estimating support size is typically parameterized in terms of a lower bound, $1/n$ on the probability of any domain element. The problem of estimating the distance to the uniform distribution on m elements is parameterized by m.

1.1 Techniques

Intuitively, this result hinges on a new connection between constructing "good" lower bounds, and "good" linear estimators.

The canonical approach to creating lower bounds for property estimation consists of finding a pair of distributions, A^+, A^- with rather different property values, yet which cannot be distinguished given the specified number of samples. The condition of indistinguishability is very stringent, and requires showing that the distribution of summary statistics derived from a set of samples from A^+ is close in total variation (L_1) distance to the corresponding distribution for samples from A^-. These distributions of summary statistics are complex discrete high-dimensional distributions, which are not well understood. Recently, in [33] we showed a central limit theorem, and related tools, that help characterize these distributions in special cases. This limit theorem suggests and enables a *principled* approach to constructing lower bounds for property estimation. Here, we show the perhaps surprising result that despite the effort required to assemble the required tools, the condition of indistinguishability in this framework can be roughly expressed via an intuitive set of *linear* constraints.

Turning, for a moment, to the side of constructing linear estimators, a natural and popular approach is to represent the "characteristic function" of the property in question as a linear combination of "Poisson functions" $poi(x, i) \triangleq \frac{e^{-x}x^i}{i!}$. [13, 25, 26, 27, 30, 36] Indeed, in [26, 27], Paninski showed the existence of a sublinear-sample linear estimator for entropy via a simple nonconstructive proof that applies the Stone-Weierstrass theorem to the set of Poisson functions. We show that the task of constructing such a representation of a given accuracy can also be framed as a set of linear constraints.

Thus general techniques for proving property testing upper and lower bounds can both be *characterized* by linear constraints. One may then ask how the performance of the best such lower bound compares to the performance of the best such upper bound. Optimizing each notion of performance relative to the corresponding linear constraints can be expressed as a linear program. Amazingly, though not unexpectedly, these two linear programs—one for constructing good lower bound example pairs, and one for constructing good linear estimators, are *dual* to each other.

The fundamental complication, however, is that the range of parameters for which the lower bound program will be pertinent, and those for which the estimator program will be pertinent, are non-intersecting. Intuitively, it is clear that these parameter ranges *must* be disjoint, as one would not expect the *exact* correspondence between optimal lower bounds of this form, and optimal linear estimators, as would be implied if these programs were dual for pertinent parameters. Thus the main technical challenge is relating optimal values of the lower bound program to optimal values of the estimator program corresponding to slightly different parameters. Establishing this relation reveals some beautiful math involving the exponentials of infinite "Poisson-matrices".

1.2 Explicit Linear Estimators and Bounds on Sample Complexity

Given that the proof of near-optimality of the linear estimators is via duality, unsurprisingly, it does not yield any explicit bounds on the sample complexities of these estimation problems. Nevertheless, inspired by numerical solutions to instantiations of these linear programs, we give an explicit description of a linear estimator for entropy which, given $O(\frac{1}{\epsilon}\frac{n}{\log n})$ independent samples from a distribution of support at most n returns an ϵ-accurate estimate with probability $1 - o(\frac{1}{poly(n)})$. Given the recent lower-bounds on estimating entropy in [33], our linear estimator is optimal, up to constant factor, both in its dependence on n and its dependence on ϵ. This is the first explicit sublinear-sample linear estimator for entropy, and the inverse-linear convergence rate settles the main open question in [34],

which left the possibility that the accuracy of an optimal estimator decreases only as the square root of the number of samples.

The machinery that we develop for constructing the estimator for entropy is robust and general, and we believe it can be employed to yield near-optimal estimators for other properties. As a simple illustration of this, we give an explicit linear estimator for estimating the distance to uniformity. Given $O\left(\frac{1}{\epsilon^2} \cdot \frac{m}{\log m}\right)$ independent samples from a distribution of *any* support, our estimator will compute the L_1 distance to $Unif(m)$ to within accuracy ϵ, with high probability. The lower-bound construction in [33] can be adapted to reveal that this is tight, to constant factor, for any constant ϵ. It is worth contrasting the above bounds for approximating distance to $Unif(m)$, with the asymmetric–error problem of distinguishing a distribution that is uniform on m elements from a distribution that is at least ϵ away from $Unif(m)$, which has received some attention [10, 19]. For this asymmetric problem, $\theta(m^{1/2})$ samples suffice and are necessary, though the proper dependence on ϵ is unclear.

1.3 Further Directions

Finally, we note that our entire framework seems to apply to the setting of properties of *pairs* of distributions. Given a set of samples from A, and a set of samples from B, how close are A and B, in total variation distance (L_1 distance), or some other distance metric? This task lies at the heart of data analysis, and it is both shocking and embarrassing that we do not understand the sample complexity of this task, or how to estimate this distance near–optimally. We strongly believe but do not show in this extended abstract that our result showing the optimality of linear estimators and its proof extend naturally to this setting, and in particular, there exist (previously unknown) linear estimators that are near-optimal for estimating L_1 distance and KL-divergence. As a proof of concept, we leverage the machinery developed for the linear estimator of entropy to give an explicit linear estimator for L_1 distance, and leverage the lower bounds of [33] to show that it is constant factor–optimal for any constant accuracy, ϵ.

2 Related Work

Linear programming duality, and, more generally, SDP duality, are beloved tools for showing the optimality of algorithms. Perhaps the most clear example of this is the celebrated max-flow min-cut theorem, which reasons that any feasible flow provides a lower bound on the optimal min-cut, and vice versa. This powerful principle has spawned much work in approximation algorithms [1, 3, 17, 24, 31]. Our use of duality is slightly different—rather than having an algorithm based on a linear program then using duality to argue that on each instance, the returned *value* is near optimal, we write a linear program that searches for *algorithms* (albeit among this very restrictive class of linear estimators). We then use duality to argue that the returned algorithm is near optimal.

2.1 Property Estimation

There has been much work on estimating a variety of symmetric distribution properties, with contributions from the statistics, computer science, and information theory communities. The specific problem of estimating the support size of an unknown distribution (also referred to as the problem of estimating the number of species in a population, or the "distinct elements problem") has been much-studied and arises in many contexts (see [12] for several hundred references). Because arbitrarily many species can lie in an arbitrarily small amount of probability mass, analysis of the sample complexity of the support size problem is generally parameterized in terms of n, where elements of

the distribution are restricted to have probability mass at least $1/n$. Tight multiplicative bounds of $\Omega(n/\alpha^2)$ for approximating the entropy to a multiplicative factor of α are given in [5, 15] though they are somewhat unsatisfying as the worst-case instance is distinguishing a distribution with support size *one* from a distribution of support size α^2. The first strong lower bounds for *additively* approximating the support size were given in [29], showing that for any constant $\delta > 0$, any estimator that obtains additive error at most $(1/2 - \delta)n$ with probability at least 2/3 requires at least $n/2^{\Theta(\sqrt{\log n \cdot \log \log n})}$ samples. Recent work [33] shows a tight bound of $O(\frac{n}{\log n})$, for estimating this property to accuracy ϵn for any constant ϵ.

For the problem of entropy estimation, there has been recent work from both the computer science and statistics communities. Batu *et al.* [6, 7, 8], Guha *et al.* [20], and Valiant [35] considered the problem of multiplicatively estimating the entropy. For the problem of additively estimating entropy, recent work [34] gives an estimator that uses $O(\frac{n}{\epsilon^2 \log n})$ samples, and returns an ϵ accurate estimate. The recent lower bounds in [33] show that $O(\frac{n}{\epsilon \log n})$ samples are necessary. Thus the dependence on n is tight, though the question of whether there exists an estimator achieving an inverse-linear convergence rate—as opposed to the much slower inverse square root rate—remained.

For the problems of estimating distance to uniformity, and L_1 distance, there has been some work focusing on the asymmetric error setting: namely, distinguishing a uniform distribution from one that is far from uniform, and in the case of L_1 distance, "identity testing"—given samples from a pair of distributions, distinguishing whether the two distributions are *the same*, versus having distance .1. Algorithms for these tasks require $\theta(n^{1/2})$, and $\tilde{\theta}(n^{2/3})$ samples, respectively. [9, 10, 19]

There has been much work on estimating the support size (and the general problem of estimating frequency moments) and estimating the entropy in the setting of *streaming*, in which one has access to very little memory and can perform only a single pass over the data [2, 4, 11, 14, 21, 22, 23, 37].

2.2 Linear Estimators for Entropy

Perhaps because of the practical importance of estimating entropy, there has been a long line of research proposing and analyzing linear estimators for entropy. As an understanding of approaches to constructing such estimators will prove useful in the remainder of this paper, we briefly describe some of this work here.

Before describing some of the commonly used estimators, it will be helpful to define the *fingerprint* of a set of samples, which, intuitively, removes all the label information from the set of samples.

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

For estimating entropy, or any other property whose value is invariant to relabeling the distribution support (a "symmetric" property), the fingerprint of a sample contains all the useful information about the sample: for any estimator that uses the actual samples, there is an estimator of equal performance that takes as input only the fingerprint of the samples (see [6, 10], for an easy proof). Note that in some of the literature the fingerprint is alternately termed the *pattern*, *histogram*, or *summary statistics* of the sample.

Perhaps the three most commonly used estimators for entropy are the following [26]:

- **The 'naive' estimator:** the entropy of the empirical distribution, namely, given a fingerprint $\mathcal{F}$ derived from a set of k samples, $H^{naive}(\mathcal{F}) \triangleq \sum_i \mathcal{F}_i \frac{i}{k} |\log \frac{i}{k}|$.
- **The Miller-Madow corrected Estimator [25]:** the naive estimator H^{naive} corrected to try to account for the second derivative of the logarithm function, namely $H^{MM}(\mathcal{F}) \triangleq H^{naive}(\mathcal{F}) +$

$\frac{(\sum_i \mathcal{F}_i)-1}{2k}$, though we note that the numerator of the correction term is sometimes replaced by various other quantities, see [28].

- **The jackknifed naive estimator [16]:** $H^{JK}(\mathcal{F}) \triangleq k \cdot H^{naive}(\mathcal{F}) - \frac{k-1}{k} \sum_{j=1}^{k} H^{naive}(\mathcal{F}^{-j})$, where $\mathcal{F}^{-j}$ is the fingerprint given by removing the contribution of the jth sample.

These estimators and their many variants generally perform very well *provided that all of the elements of the support occur with large probability.* The problem with these estimators can be summarized as their inability to appropriately deal with samples from distributions where a significant portion of of the probability mass lies in domain elements not represented in the sample. For example, given $o(n)$ samples from the uniform distribution on support n, these estimators generally fail to recover an accurate estimate. In particular, these estimators make no attempt to understand the (potentially significant) contribution towards the entropy of the distribution that comes from the "unseen" portion of the distribution. The estimator we explicitly construct in Appendix A, in some sense, is specifically designed to account for this contribution.

No explicit sublinear-sample estimators were known for additively estimating entropy to within even a constant. Nevertheless, in [26, 27], Paninski proved the *existence* of a sublinear-sample estimator; the proof is non-constructive, via a direct application of the Stone-Weierstrass theorem to the set of Poisson functions. Our approach falls within this framework, though rather than employing the powerful but nonconstructive Stone-Weierstrass theorem, we explicitly construct an estimator, via a Chebyshev polynomials construction.

This framework, which is described in Section 5.2, seems well-known in the literature prior to [26], even dating back to [25] in the 1950's. The fundamental difficulty, which we overcome, essentially comes down to approximating the logarithm function via a linear combination of Poisson functions (see Appendix A). Such a representation has been attempted in the past, either explicitly or implicitly in [13, 25, 30, 36], though these works were unable to succeed in producing an accurate approximation of the logarithm function in the small-probability regime.

3 Definitions and Examples

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

Definition 2. *A* distribution *on* $[n] = \{1, \ldots, n\}$ *is a function* $p : [n] \to [0, 1]$ *satisfying* $\sum_i p(i) = 1$. *Let* $\mathcal{D}^n$ *denote the set of distributions over domain* $[n]$.

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

We now define a *linear* estimator.

Definition 3. *A* k*-sample* linear *estimator* α *is defined by a set of at least* k *coefficients,* $\alpha = (\alpha_1, \ldots, \alpha_k)$. *The estimator is defined as the dot product between the fingerprint vector* $\mathcal{F}$ *of a set of* k *samples, and the vector* α, *namely* $S_k(\mathcal{F}) \triangleq \sum_{i=1}^{k} \alpha_i \mathcal{F}_i$.

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

Definition 4. *A* property *of a distribution is a function* $\pi : \mathcal{D}^n \to \mathbb{R}$. *Additionally, a property is* symmetric *if, for all distributions* D, *and all permutations* σ, $\pi(D) = \pi(D \circ \sigma)$.

Analogous to the fingerprint of a set of samples, is what we call the *histogram of the distribution*, which captures the number of domain elements that occur with each probability value.

Definition 5. *The* histogram *of a distribution p is a mapping $h : (0,1] \to \mathbb{Z}$, where $h(x) = |\{i : p(i) = x\}|$.*

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

It is clear that any symmetric property is a function of only the histogram of a distribution. Finally, a symmetric property is *linear*, if the property value is a linear function of the histogram:

Definition 6. *A symmetric property π is* linear *if there exists some function $f_\pi : [0,1] \to \mathbb{R}$ which we term the* characteristic function *of π, such that for any distribution A with histogram h_A,*

$$\pi(A) = \sum_{x:h_A(x)\neq 0} h(x) f_\pi(x).$$

We now give several examples of symmetric linear properties:

Example 7. *The (Shannon) entropy of a discrete distribution $p \in \mathcal{D}^n$ with histogram h is given by $H(h) \triangleq \sum_{i=1}^n p(i)|\log p(i)| = \sum_{x:h(x)\neq 0} h(x)f(x)$, for the function $f(x) \triangleq x|\log x|$.*

Example 8. *The support size of a discrete distribution $p \in \mathcal{D}^n$ with histogram h is given by $\sum_{x:h(x)\neq 0} h(x)f(x)$, for the function $f(x) \triangleq 1$.*

Example 9. *The total variation distance between a discrete distribution $p \in \mathcal{D}^n$ with histogram h and a uniform distribution on s elements can be approximated to within a factor of 2 as $\sum_{x:h(x)\neq 0} h(x)f(x)$, for the function*

$$f(x) \triangleq \begin{cases} x & \text{for } x \leq \frac{1}{2s} \\ |x - \frac{1}{s}| & \text{for } x > \frac{1}{2s}. \end{cases}$$

It will also be essential to have a distance metric between distributions with respect to which the class of properties in question are continuous:

Definition 10. *For two histograms h_1, h_2, we define the* relative earthmover distance *between them, $R(h_1, h_2)$, as the minimum cost, over all schemes of moving the probability mass of the first histogram to yield the second histogram, where the cost per-unit probability of moving mass from probability x to y is $|\log(x/y)|$.*

A distribution property π is c-relative earthmover continuous *if for all distributions h_1, h_2, we have $|\pi(h_1) - \pi(h_2)| \leq c \cdot R(h_1, h_2)$.*

A linear property π with characteristic function f_π is c-relative earthmover continuous *if for all $x, y \in (0,1]$ we have $|\frac{f_\pi(x)}{x} - \frac{f_\pi(y)}{y}| \leq |\log(x/y)|$.*

We provide an example:

Example 11. *Consider a sequence of fish species, found as samples from a certain lake, $X = (trout, salmon, trout, cod, cod, whale, trout, eel, salmon)$. We have $\mathcal{F}_X = (2, 2, 1)$, indicating that two species occurred exactly once (whale and eel), two species occurred exactly twice (salmon and cod), and one species occurred exactly three times (trout).*

Suppose that the true distribution of fish is the following:

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

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

3.1 Poisson Samples

To understand the main ideas, it will be helpful to have an intuitive understanding of the distribution of the fingerprint corresponding to a set of k samples from histogram h. This distribution intimately involves the Poisson distribution. Throughout, we use $Poi(\lambda)$ to denote the Poisson distribution with expectation λ, and for a nonnegative integer j, $poi(\lambda, j) \triangleq \frac{\lambda^j e^{-\lambda}}{j!}$ denotes the probability that a random variable distributed according to $Poi(\lambda)$ takes value j. Additionally, for integers $i \geq 0$, we refer to the function $poi(x, i)$, viewed as a function of the variable x, as the ith *Poisson function*.

Given a fingerprint corresponding to a set of k samples from a distribution p, the number of occurrences of any two elements are not independent; however, if instead of taking k samples, we chose $k' \leftarrow Poi(k)$ according to a Poisson distribution with expectation k and then take k' samples from p, the number of occurrences of each domain element $i \in [n]$ will be independent random variables with distributions $Poi\left(k \cdot p(i)\right)$. This independence is quite helpful when arguing about the structure of the distribution of such fingerprints.

We provide a clarifying example:

Example 12. *Consider the uniform distribution on* $[n]$, *which has histogram* h *such that* $h(\frac{1}{n}) = n$, *and* $h(x) = 0$ *for* $x \neq \frac{1}{n}$. *Let* $k' \leftarrow Poi(5n)$ *be a Poisson-distributed random number, and let* X *be the result of drawing* k' *independent samples from the distribution. The number of occurrences of each element of* $[n]$ *will be independent, distributed according to* $Poi(5)$. *Note that* $\mathcal{F}_X(i)$ *and* $\mathcal{F}_X(j)$ *are* not *independent (since, for example, if* $\mathcal{F}_X(i) = n$ *then it must be the case that* $\mathcal{F}_X(j) = 0$, *for* $i \neq j$). *A fingerprint of a typical trial will look roughly like* $\mathcal{F}(i) \approx n \cdot poi(5, i)$.

Since $k' \leftarrow Poi(k)$ is closely concentrated around k, one might hope that in terms of most properties of interest, there is little difference between considering k-sample fingerprints and $Poi(k)$-sample fingerprints. The following easy fact, whose proof follows immediately from standard tail bounds for Poisson distributions, allows us to prove statements about k-sample fingerprints by considering the structurally more simple $Poi(k)$-sample fingerprints.

Fact 13. *For sufficiently large* k, *for any estimator that estimates entropy to within error* ϵ *with probability at least* $1 - \delta$ *when given a sample of size* $k' \leftarrow Poi(k)$, *there is an estimator that has error at most* ϵ *with probability at least* $1 - \delta - e^{-k^{.9}}$ *and takes as input a set of* $k \cdot (1 + o(1))$ *samples.*

We now consider the distribution of the ith entry of a $Poi(k)$-sample fingerprint, $\mathcal{F}(i)$. Since the number of occurrences of different domain elements are independent, $\mathcal{F}(i)$ is distributed as the sum of n independent $0, 1$ random variables $Y_1, \ldots, Y_n$, where $\Pr[Y_j = 1] = poi(k \cdot p(j), i)$ is the probability that the jth domain element occurs exactly i times in sample X. Thus

$$E[\mathcal{F}(i)] = \sum_{j \in [n]} poi(k \cdot p(j), i) = \sum_{x : h(x) \neq 0} h(x) \cdot poi(kx, i),$$

and from independence, we will have good concentration about this expectation.

4 Summary of Results

Our main theorem shows that linear estimators are near-optimal for the class of linear symmetric distribution properties, provided that they satisfy a mild continuity condition:

Theorem 1. *Let π be a symmetric linear property that is $\delta(k)$-relative earthmover continuous on distributions of support $n(k)$. If for some constant $c > 0$ and parameter $\epsilon(k) = \delta/k^{o(1)}$, any distributions of support n whose π values differ by at least ϵ are distinguishable with probability at least $\frac{1}{2}+c$ in k samples, then for each k there exists a linear estimator that estimates π on distributions of support n to within error $(1+o(1))\epsilon$ using $(1+o(1))k$ samples, and which has probability of failure $o(\frac{1}{poly(k)})$.*

To clarify, the above theorem trivially implies the following corollary:

Corollary 14. *Given a symmetric linear property π that is 1-relative earthmover continuous (such as entropy), if there exists an estimator which on input k independent samples from any distribution A of support n outputs a value v such that $|v - \pi(A)| < \epsilon$ with probability .51, then there exists a linear estimator which, given $1.01k$ samples, outputs a value v' such that $|v' - \pi(A)| \leq 2.01\epsilon$, with probability $> .9999$, provided $\epsilon \geq \frac{1}{\log^{100} k}$ and k is sufficiently large.*

While Theorem 1 does not yield bounds on the sample complexities of these estimation tasks, in Appendices A and B we leverage the insights provided by key components of the proof of Theorem 1 to give explicit constructions of linear estimators for entropy, and distance to uniformity. These estimators significantly improve upon all previously proposed estimators for these properties.

Theorem 2. *For any $\epsilon > \frac{1}{n^{0.03}}$, the estimator described in Construction 22, when given $O(\frac{n}{\epsilon \log n})$ independent samples from a distribution of support at most n will compute an estimate of the entropy of the distribution, accurate to within ϵ, with probability of failure $o(1/poly(n))$.*

We note that the performance of this estimator, up to constant factors, matches the lower bounds shown in [33, 32], both in terms of the dependence on n and the dependence on ϵ. In particular, this resolves the main open question posed in [34, 32] as to whether the sample complexity increases linearly versus quadratically with the inverse of the desired accuracy, $1/\epsilon$.

The machinery developed in the construction of our linear estimator for entropy is quite general and robust, and we show that it can easily be used to construct an explicit estimator for distance to uniformity.

Theorem 3. *For any $\epsilon > \frac{1}{4 \log m}$, the estimator described in Construction 33, when given $O\left(\frac{1}{\epsilon^2} \cdot \frac{m}{\log m}\right)$ independent samples from a distribution of* any *support, will compute the L_1 distance to $Unif(m)$ to within accuracy ϵ, with probability of failure $o(1/poly(m))$.*

This is the first $o(m)$ sample estimator for distance to uniformity, and we note that the lower bounds shown in [33, 32] imply that for any constant error ϵ, this estimator is optimal, to constant factor. This tight bound of $\Theta(m/\log m)$ on the number of samples required to yield constant error contrasts with the tight bound of $\Theta(m^{1/2})$ shown in [9, 19] for the related problem of distinguishing a uniform distribution on m samples from one that has constant distance from such a distribution.

Finally, we note that our entire approach seems to apply to the setting of properties of *pairs* of distributions. We strongly believe but do not show in this extended abstract that Theorem 1 and its proof extend naturally to this setting. As a proof of concept, we leverage the machinery developed for the linear estimator of entropy to give an explicit linear estimator for L_1 distance, and leverage the lower bounds of [33] to show that it is constant factor–optimal for any fixed accuracy.

Theorem 4. *There is a constant c such that for any $\epsilon > \frac{c}{\sqrt{\log n}}$, the estimator described in Construction 38, when given $O(\frac{n}{\epsilon^2 \log n})$ independent samples from each of two distributions of support at most n, will compute an estimate of the L_1 distance between the pair of distributions, accurate to within ϵ, with probability of failure $o(1/poly(n))$.*

Theorem 5. *For any constants $0 < a < b < \frac{1}{2}$, there exists a pair of distributions with support at most n such that distinguishing whether their L_1 distance is less than a or greater than b requires $O(\frac{n}{\log n})$ samples.*

This is the first sublinear-sample estimator for this fundamental property, and the lower bound of Theorem 5 improves upon the previous best lower bound of $n/2^{O(\sqrt{\log n})}$ shown in [35].

5 Lower Bounds and Estimators

We start by describing an intuitive approach to constructing lower bound instances for the task of estimating a given linear property, and then describe a natural and well-known approach to constructing linear estimators. It will then be immediate that these two approaches are related via linear programming duality. Finally, in Section 5.2.1 we examine the crux of the difficulty in making this correspondence precise.

5.1 Lower Bounds on Property Estimation

Given a property π, a number of samples k, and an upper bound n on the support size of distributions in question, we wish to construct lower-bounds via a principled—and in some sense mechanical—approach. Specifically, we would like to find two distributions A^+, A^- (of support at most n) which are extremal in the sense that they maximize $\delta = |\pi(A^+) - \pi(A^-)|$ while being *indistinguishable* (with high probability) given sets of k independent samples from each. Trivially, such a pair implies that *no* algorithm, on input k independent samples from a distribution of support at most n, can estimate property π to within $\pm\delta/2$ with high probability.

At least intuitively, A^+ and A^- will be difficult to distinguish, given sets of k samples, if their fingerprint expectations are very similar (relative to the size of the covariance of the distribution of fingerprints). The central limit theorem for "generalized multinomial" distributions given in [33] makes this intuition rigorous. Since these fingerprint expectations are simply *linear* functions of the histograms, this constraint that A^+ and A^- should be indistinguishable can be characterized by a set of linear constraints on the histograms of A^+ and A^-. Additionally, from Example 8, the constraint that A^+ and A^- have support size at most n is a linear constraint on the histograms: $\sum_{x:h_A(x)\neq 0} h_A(x) \leq n$. Since we are concerned with a symmetric linear property, π, which is given as $\pi(A) \triangleq \sum_{x:h_A(x)\neq 0} h_A(x) f_\pi(x)$, for some function f_π, our aim of maximizing the discrepancy in property values, $\pi(A^+) - \pi(A^-)$, is just the task of optimizing a linear function of the histograms. Thus, at least intuitively, we can represent the task of constructing an optimal lower-bound instance (A^+, A^-), as a semi-infinite linear program whose variables are $h_{A^+}(x), h_{A^-}(x)$, for $x \in (0, 1]$.

Before writing the linear program, there are a few details we should specify. Rather than solving for histogram values $h_{A^+}(x)$, it will be more convenient to solve for variables y_x^+, which are related to histogram values by $y_x^+ \triangleq h_{A^+}(x) \cdot x$. Thus y_x^+ represents the amount of probability mass accounted for by $h_{A^+}(x)$. Thus $\sum_x y_x^+ = 1$ for any distribution A^+. For reasons which will become clear, we will also restrict ourselves to the "infrequently-occurring" portion of the histogram: namely, we will only be concerned with fingerprint indices up to k^{c_1}, for a parameter $c_1 \in (0, 1)$, and will only solve for histogram entries corresponding to probabilities $x \leq \frac{1}{2}\frac{k^{c_1}}{k}$. Finally, to avoid the messiness that comes with semi-infinite linear programs, we will restrict ourselves to a finite set of variables, corresponding

to x values in some set $X \subset (0, \frac{k^{c_1}}{2k})$ that consists of a polynomially-fine mesh of points, the details of which are largely irrelevant.

Definition 15. *The* Lower Bound LP *corresponding to parameters k, c_1, c_2, X, and property π satisfying $\pi(A) \triangleq \sum_{x:h(x)\neq 0} h_A(x) f_\pi(x)$, is the following:*

Maximize:	$\sum_{x\in X} \frac{f_\pi(x)}{x}(y_x^+ - y_x^-)$	*Maximize difference in property values,*
Subject to:	$\forall i \leq k^{c_1}, \sum_x (y_x^+ - y_x^-) \cdot poi(xk, i) \leq k^{-c_2}$	*fingerprint expectations are close,*
	$\forall i \leq k^{c_1}, \sum_x (y_x^+ - y_x^-) \cdot poi(xk, i) \geq -k^{-c_2}$	*fingerprint expectations are close,*
	$\sum_{x\in X} y_x^+ + y_x^- \leq 2$	*not too much probability mass,*
	$\sum_{x\in X} \frac{y_x^+}{x} \leq n \quad$ *and* $\quad \sum_{x\in X} \frac{y_x^-}{x} \leq n$	*A and B have support at most n,*
	$\forall x \in X, y_x^+ \geq 0, y_x^- \geq 0$	*probabilities are nonnegative.*

We now argue that the intuition for the above linear program is well founded. For any reasonably well-behaved property π, given a solution to the above linear program y^+, y^-, with objective function value v, we can construct distributions A^+, A^- that are indistinguishable given k samples, and satisfy $\pi(A^+) - \pi(A^-) \geq v - \epsilon$ for some tiny ϵ. As shifting a property by a constant, $\pi \to \pi + C$ does not affect the property estimation problem, for the sake of convenience we assume that the property takes value 0 on the trivial distribution with support 1, though the following proposition remains true for rather extreme (though not unbounded) shifts away from this.

Proposition 16. *Let π be a δ-relative earthmover continuous property that takes value 0 on the trivial distribution. Given any feasible point y^+, y^- to the* Lower Bound LP *of Definition 15 that has objective function value v, then, provided $k^{c_1} \in [\log^2 k, k^{1/32}]$ and $c_2 \geq \frac{1}{2} + 6c_1$, there exists a pair of distributions A^+, A^- of support at most n such that:*

- $\pi(A^+) - \pi(A^-) > v \cdot (1 - o(1)) - O(\delta \cdot k^{-c_1} \log k)$,
- no *algorithm on $Poi(k)$-samples can distinguish A^+ from A^- with probability $1 - \Theta(1)$.*

To construct A^+, A^- from the solution y^+, y^-, there are three hurdles. First, y_x^+, y_x^- must be rounded so as to be integer multiples of $1/x$, since the corresponding histograms must be integral. Next, we must ensure that A^+, A^- have total probability mass 1. Most importantly, we must ensure that A^+, A^- are actually indistinguishable—i.e. that we can successfully apply the central limit theorem of [33]—a more stringent condition than simply having similar fingerprint expectations. These three tasks must be accomplished in a delicate fashion so as to ensure that $\pi(A^+) - \pi(A^-) \approx v$. The explicit construction, and proof of Proposition 16 are included in Appendix D.

5.2 Constructing Linear Estimators

Perhaps the most natural approach to constructing estimators for linear properties, dating back at least to the 1950's, [25] and, implicitly, far longer, is to approximate the characteristic function of the desired linear property as a linear combination of Poisson functions. To see the intuition for this, consider a property π such that $\pi(A) \triangleq \sum_{x:h_A(X)\neq 0} h_A(x) f_\pi(x)$, and assume that there exist coefficients $\beta = \beta_1, \beta_2, \ldots$ such that, for all $x \in (0, 1]$, $\sum_{i=1}^\infty \beta_i poi(xk, i) = f_\pi(x)$. Thus for a distribution with histogram h, we have

$$\sum_{x:h(x)\neq 0} h(x) f_\pi(x) = \sum_{x:h(x)\neq 0} h(x) \sum_{i\geq 1} \beta_i poi(kx, i) = \sum_{i\geq 1} \beta_i \sum_{x:h(x)\neq 0} h(x) poi(kx, i) = \sum_{i\geq 1} \beta_i E[\mathcal{F}(i)],$$

where $E[\mathcal{F}(i)]$ is the expected ith fingerprint entry derived from $Poi(k)$ independent samples. By linearity of expectation, this quantity is precisely the expected value of the linear estimator given by the coefficients β, and thus such an estimator would have *zero* bias. Additionally, since we expect the fingerprint entries to be closely concentrated about their expectations, such an estimator would also have relatively small variance, provided that the magnitudes of the coefficients $|\beta_i|$ are small relative to $1/\sqrt{k}$. (Roughly, the contribution to the variance of the estimator from the ith fingerprint entry is the product of $|\beta_i|^2$ and the variance of the ith fingerprint entry.)

For several reasons which will become apparent, instead of approximating the function $f_\pi(x)$ as $\sum_{i=1}^{\infty} \beta_i poi(kx, i)$, we instead approximate the function $\frac{f_\pi(x)}{x}$ as the 0-indexed sum $\sum_{i=0}^{\infty} z_i poi(kx, i)$. These two approaches are formally identical by setting $\beta_i = \frac{i}{k} \cdot z_{i-1}$, since $x \cdot poi(kx, i) = poi(kx, i+1)\frac{i+1}{k}$.

The following proposition formalizes this intuition, establishing the requisite relationship between the magnitudes of the coefficients, error in approximating the function $\frac{f_\pi(x)}{x}$, and the performance of the derived estimator. The relatively straightforward proof of this proposition is in Appendix E.

Proposition 17. *Let π be a linear symmetric property such that for any histogram h, we have $\pi(h) \triangleq \sum_{x:h(x)\neq 0} h(x)x \cdot r(x)$, for some function $r : (0, 1] \to \mathbb{R}$. Given integers k, n, and a set of coefficients $z_0, z_1, \ldots$ such that if we define the function $err : (0, 1] \to \mathbb{R}$ by*

$$r(x) = err(x) + \sum_{i \geq 0} z_i poi(xk, i),$$

and if for positive real numbers a, b, c the following conditions hold:

1. *$|err(x)| < a + \frac{b}{x}$,*
2. *for all $j \geq 1$ let $\beta_j = \frac{j}{k} \cdot z_{j-1}$ with $\beta_0 = 0$, then for any j, ℓ such that $|j - \ell| \leq \sqrt{j} \log k$ we have $|\beta_j - \beta_\ell| \leq c\frac{\sqrt{j}}{\sqrt{k}}$*

Then the linear estimator given by coefficients $\beta_1, \ldots, \beta_k$, when given a fingerprint derived from a set of k independent samples chosen from a distribution of support at most n will estimate the property value with error at most $a + bn + c \log k$, with probability of failure $o(1/poly(k))$.

We note that the condition on the magnitude of the error of approximation: $|err(x)| < a + \frac{b}{x}$, is designed to take into account the inevitable increase in this error as $x \to 0$. Intuitively, this increase in error is offset by the bound on support size: for a distribution of support at most n, the amount of probability mass at probability x is bounded by nx, and thus provided that the error at x is bounded by $\frac{b}{x}$, the error of the derived estimator will be at most $nx\frac{b}{x} = nb$.

The task of finding these coefficients z_i, can be expressed as the following linear program:

Definition 18 (The Linear Estimator LP)**.**

$$\begin{aligned}
\textit{Minimize:} \quad & 2z^a + n \cdot (z^{b+} + z^{b-}) + k^{-c_2} \textstyle\sum_{i=0}^{k^{c_1}} (z_i^+ + z_i^-) \\
\textit{Subject to:} \quad & \forall x \in X, \textstyle\sum_{i=0}^{k^{c_1}} poi(xk, i)(z_i^+ - z_i^-) \geq \frac{f_\pi(x)}{x} - (z^a + \frac{z^{b-}}{x}) \\
& \forall x \in X, \textstyle\sum_{i=0}^{k^{c_1}} poi(xk, i)(z_i^+ - z_i^-) \leq \frac{f_\pi(x)}{x} + z^a + \frac{z^{b+}}{x} \\
& z^a \geq 0, z^{b+}, z^{b-} \geq 0, \forall i \in [k^{c_1}],\ z_i^+ \geq 0, z_i^- \geq 0
\end{aligned}$$

To see the relation between the above definition and Proposition 17, we let the coefficients $z_i = z_i^+ - z_i^-$. The parameter a in the proposition corresponds to z^a in the LP, and the parameter b in the proposition corresponds to $\max(z^{b+}, z^{b-})$. The first two sets of constraints ensure that z^a, z^{b+}, z^{b-} capture the bias of the estimator. The objective function then minimizes this bias, while also penalizing unduly large coefficients.

5.2.1 So Close, Yet So Far

The impetus for our main result is the observation that the Lower Bound LP of Definition 15 and the Linear Estimator LP of Definition 18 are dual linear programs. The complication arises when considering the allowable settings of the parameters. Intuitively, the Lower Bound LP only begins to make sense when $c_2 > 1/2$—namely, when the discrepancy in fingerprint expectations of the implicitly described pair of distributions is less than $k^{1/2}$, since the standard deviation in fingerprint entries can never exceed this value. Conversely, the Linear Estimator LP yields reasonable estimators only when $c_2 < 1/2$, since this corresponds to coefficients at most $1/k^{1/2}$, which, coupled with the variance in fingerprint entries of up to k, would lead to an estimator having constant variance.

Thus, if we hope to leverage the duality of the two linear programs, we must understand how the solutions are affected by changing parameters. Intuitively, the quality of the best available estimator should improve as the number of samples to which one has access increases. For some number of samples, k, given a solution to the Lower Bound LP with objective function value v, we argued that there is no corresponding solution to the Linear Estimator LP with viable parameters; the above intuition suggests that by considering the Linear Estimator LP corresponding to $k' > k$ samples, the increase in the number of samples would offset the necessary decrease in c_2 and might yield a solution with objective value near v. This hope is well-founded, and we show that, even for $k' = k \cdot (1 + o(1))$, an objective value of $v \cdot (2 + o(1))$ can be attained. We discuss the high-level approach to proving this in the following section. (We note that the factor of 2 is an artifact of the duality framework itself and not of the argument that follows.)

5.3 Matrix Exponentials of Poisson Matrices

The aim of this section is to transform a solution to the Linear Estimator LP into a related estimator that: 1) has smaller coefficients; 2) takes slightly more samples; and 3) has almost unchanged bias. Intuitively, we have a vector of Poisson coefficients, z, whose magnitudes exceed $\sqrt{k}$, yet whose linear combination, the function $g : [0, \infty) \to \mathbb{R}$ defined as $g(x) = \sum_{i=0}^{\infty} z(i) \cdot poi(xk, i)$ closely approximates $\frac{f_\pi(x)}{x}$, and thus, despite its huge coefficients, the resulting function is small and well-behaved. The task is to transform this into a different linear combination that has smaller coefficients and is almost equally well-behaved. The principal tool we may leverage is the increased number of samples we have. While $poi(xk, i)$ captures the Poisson functions corresponding to taking k samples, if we instead take $\frac{k}{\alpha}$ samples for $\alpha < 1$, then the corresponding functions are $poi(\frac{xk}{\alpha}, i)$, which are "thinner" than the original Poisson functions. To phrase the intuition differently, if the target function $\frac{f_\pi(x)}{x}$ is so finely structured that approximating it with "fat" Poisson functions requires coefficients exceeding $\sqrt{k}$, we might hope that using "thinner" Poisson functions will lower the required coefficients.

We note that it is straightforward to reexpress a linear combination of Poisson functions in terms of "thinner" Poisson functions. Intuitively, this is the process of simulating a $Poi(k)$-sample estimator using $Poi(\frac{k}{\alpha})$ samples, and corresponds to subsampling. We let z_α denote the vector of coefficients induced from subsampling by α—that is, $z_\alpha(\ell) = \sum_{i=0}^{\ell} z(i) Pr[Bin(\ell, \alpha) = i]$, where $Bin(\ell, \alpha)$ represents the binomial distribution taking ℓ trials each with success probability α. The question becomes: how does the magnitude of z_α decrease with α?

We show that the square of the L_2 norm of the vector z_α is a quadratic form in z, defined by an infinite matrix M_α. We are able to analyze these norms because of the fortuitous form of its *matrix logarithm*: there exists an infinite tri-diagonal matrix A such that for all $\alpha \in (0, 1)$, $M_\alpha = \frac{1}{\alpha} e^{(1-\alpha)A}$. We show this via the Gauss relations for contiguous hypergeometric functions. Our main result then follows from the fact that the quadratic form $ze^{\alpha X}z^\intercal$ is a *log-convex* function of α, for arbitrary z and X. This approach is made rigorous in Appendix F, proving Theorem 1.

References

[1] R. Ahlswede, N. Cai, S. Li, and R. Yeung. Network information flow. volume 46, pages 1204–1216, 2000.

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

[3] S. Arora, E. Hazan, and S. Kale. $o(\sqrt{\log n})$ approximation to sparsest cut in $\tilde{O}(n^2)$ time. In *FOCS, 2004*.

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

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

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

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

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

[9] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White. Testing random variables for independence and identity. In *FOCS, 2001*.

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

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

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

[13] A. Carlton. On the bias of information estimates. *Psychological Bulletin*, 71:108–109, 1969.

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

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

[16] B. Efron and C. Stein. The jacknife estimate of variance. *Annals of Statistics*, 9:586–596, 1981.

[17] N. Garg, V. Vazirani, and M. Yannakakis. Approximate max-flow min-(multi)cut theorems and their applications. In *STOC, 1993*.

[18] P. W. Glynn. Upper bounds on Poisson tail probabilities. *Operations Research Letters*, 6(1):9–14, 1987.

[19] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. In *FOCS, 1996*.

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

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

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

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

[24] T. Leighton and S. Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. volume 46, pages 1204–1216, 1999.

[25] G. Miller. Note on the bias of information estimates. *Information Theory in Psychology II-B*, ed H. Quastler (Glencoe, IL: Free Press):pp 95–100, 1955.

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

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

[28] S. Panzeri and A. Treves. Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems*, 7:87–107, 1996.

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

[30] A. Treves and S. Panzeri. The upward bias in measures of information derived from limited data samples. *Neural Computation*, 7:399–407, 1995.

[31] L. Trevisan. Max cut and the smallest eigenvalue. In *STOC, 2009*.

[32] G. Valiant and P. Valiant. Estimating the unseen: an $n/\log(n)$–sample estimator for entropy and support size, shown optimal via new CLTs. *STOC, 2011*.

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

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

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

[36] J. Victor. Asymptotic bias in information estimates and the exponential (bell) polynomials. *Neural Computation*, 12:2797–2804, 2000.

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

A An Optimal Linear Estimator For Entropy

In this section we describe an explicit linear estimator for entropy, which, given as input $k = O\left(\frac{n}{\epsilon \log n}\right)$ samples from a distribution of support at most n will return an estimate of the entropy accurate to within ϵ, with probability of failure $o(1/poly(n))$. These bounds match the lower bounds on estimating entropy given in [33] both in terms of the dependence on n, and the dependence on the desired accuracy, ϵ, and, in particular show that the convergence rate is inverse linear in the number of samples, as opposed to the slower inverse square root which is generally expected. For clarity of exposition and because the proof of correctness is quite tedious, we first present the high-level structure of the estimator and state key lemmas, and then include the more technical proofs in the following section.

Our estimator is based on an accurate approximation of the logarithm function as a low-weight sum of the Poisson functions. The key technical insight is the strengthening and re-purposing of a Chebyshev polynomial construction which was employed in [34] as a component of an "earthmoving scheme". Here, we use this construction to turn the basis of Poisson functions into a more adroit basis of "skinny" bumps, which are, in a very rough sense, like the Poisson functions compressed by a factor of $\log k$ towards the origin. Intuitively, this superconstant factor is what allows us to construct a sublinear-sample estimator.

Perhaps the most simplistic attempt to represent the logarithm function as a sum of Poisson functions is to simply set the coefficient of $poi(xk, i)$ equal to $\log \frac{i}{k}$. This estimator is the "plug-in" estimator, and returns the entropy of the empirical distribution of the samples. Intuitively, this estimator should be good for the high-probability region. The following lemma, which we will use later, characterizes the performance of any "plug-in" estimator. The proof is contained in Section A.1.

Lemma 19. *Given a function $f : \mathbb{R} \to \mathbb{R}$ whose fourth derivative at x is bounded in magnitude by $\frac{\alpha}{x^4}$ for $x \geq 1$ and by α for $x \leq 1$, and whose third derivative at x is bounded by $\frac{\alpha}{x^3}$, then for any real x, $\sum_{i=0}^{\infty} f(i) \cdot poi(x, i)$ is within $O(\frac{\alpha}{x^2})$ of $f(x) + \frac{1}{2}xf''(x)$.*

For the plug-in estimator for entropy, this lemma implies that

$$\log x - \sum_{i=0}^{\infty} \log(i/k) poi(kx, i) = \frac{1}{2kx} + O(\frac{1}{k^2x^2}).$$

In some regimes this error is satisfactorily small, and this estimator is in fact widely used in practice. However, for $x = 1/k$ the error is constant, and for smaller x the error blows up. Thus for uniform distributions of support larger than k, the plug-in estimator performs poorly.

How can one improve this estimator? The obvious correction is to account for the second-derivative term of Lemma 19, corresponding to the term $\frac{1}{2kx}$ in the above expression for the bias for entropy. This yields the "Miller-Madow Corrected Estimator". Nevertheless, the error term is still constant for $x = 1/k$, making sublinear-sample estimation impossible. Such error is, in some sense, to be expected: the first few Poisson functions $poi(kx, i)$ have "width" $O(1/k)$.

A plug-in estimator in terms of a "skinnier" basis than the Poisson functions would make the estimate correspondingly more accurate. The crux of our estimator is to first show that we can represent "skinny bumps" as a low-weight linear combinations of Poisson functions. We then employ these skinny bumps in place of the fat Poisson functions to get correspondingly better estimators.

Perhaps the most useful building blocks for constructing functions that resemble "skinny bumps" are the trigonometric functions, $cos(nx)$, for $n = 0, 1, 2, \ldots$. Since each Poisson function $poi(x, i)$ is a degree j polynomial in x, multiplied by an exponential e^{-x}, we instead work with the polynomial equivalent of the trigonometric functions: the Chebyshev polynomials, where the jth Chebyshev polynomial T_j is defined so as to satisfy $T_j(\cos(y)) = \cos(j \cdot y)$.

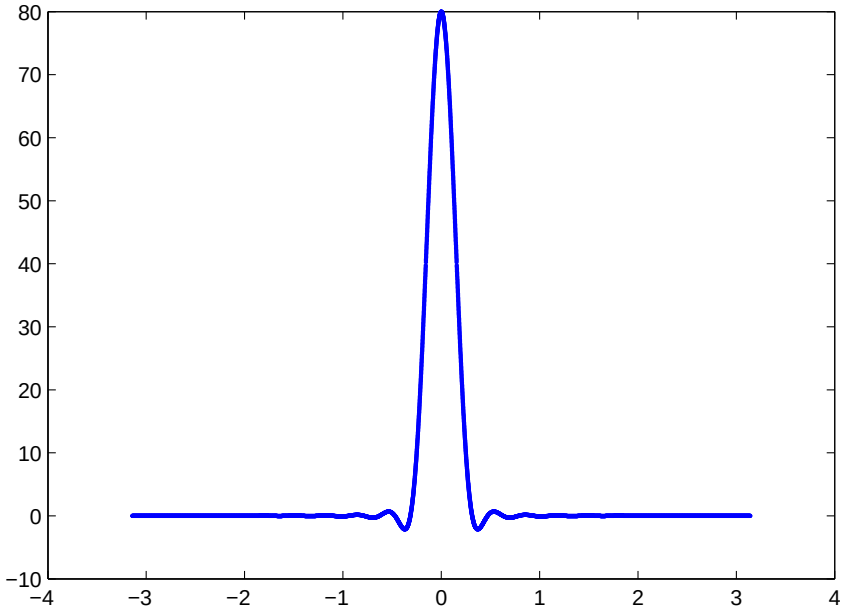

Figure 1: A plot of the "skinny" function $g_2(y)$ (without the scaling factor). This is the main ingredient in the Chebyshev bump construction of Definition 20.

Definition 20. *The* Chebyshev bump scheme *is defined in terms of k as follows. Let $s = (0.3)\log k$. Define $g_1(y) = \sum_{j=-s}^{s-1} \cos(jy)$. Define $g_2(y) = \frac{1}{16s}\left(g_1(y - \frac{3\pi}{2s}) + 3g_1(y - \frac{\pi}{2s}) + 3g_1(y + \frac{\pi}{2s}) + g_1(y + \frac{3\pi}{2s})\right)$, 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 that $t_i(\cos(y)) = g_3^i(y)$. 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 $x = \frac{2s}{k}(1 - \cos(y))$, and scaling by $\sum_{j=0}^{s-1} poi(xk, j)$. For these bumps, define $c_i = \frac{2s}{k}(1 - \cos(\frac{i\pi}{s}))$.*

See Figure 1 for a plot of $g_2(y)$, illustrating, up to coordinate transformations, a "skinny Chebyshev bump."

The following lemma shows that each of the Chebyshev bumps defined above can be expressed as a linear combination of the Poisson functions, having relatively small coefficients. The proof of this lemma is contained in Section A.1.

Lemma 21. *Each $B_i(x)$ may be expressed as $\sum_{j=0}^{\infty} a_{ij} poi(kx, j)$ for a_{ij} satisfying $\sum_{j=0}^{\infty} |a_{ij}| \leq k^{0.4}$*

This lemma will allow us to attempt to approximate the logarithm function using Poisson functions by directly approximating the logarithm function via these conveniently-skinny bumps. We note that the bound $k^{0.4}$ on the coefficients is crucial, as the coefficients of our estimator must be somewhat less than $\sqrt{k}$ in order for our k-sample estimator to have sub-constant variance. As the coefficients of Chebyshev polynomials grow exponentially in their degree, this is what limits us to the first $s = O(\log k)$ Chebyshev polynomials. Thus our approximation of the logarithm function via the Chebyshev bumps will only apply to the very low-probability region—but this is acceptable,

since above this region, the more crude approximation via the Poisson functions and Lemma 19 will suffice.

We are now prepared to define our estimator. We start by defining the coefficients $\{z_i\}$ such that $\sum_{i\geq 0} z_i \cdot poi(xk, i) \approx \log x$.

Construction 22. *As in the previous definition, let $s = (0.3)\log k$. Define the interpolation function $I : \mathbb{R} \to \mathbb{R}$ such that $I(y) = 0$ for $y \leq \frac{s}{4}$, $I(y) = 1$ for $y \geq \frac{s}{2}$, and $I(y)$ is continuous, and four-times differentiable, where for $i \in 1, \ldots, 4$, the magnitude of the ith derivative is at most c/s^i, for some fixed constant c. Such a function I can be easily constructed.*

Consider the function $f(y) \triangleq I(y)\left[\frac{1}{2y} + \log y - \log k\right]$, and provisionally set $z_i \triangleq f(i)$. Note that by Lemma 19 we have accurately represented the logarithm function via the Poisson bumps in the interval $[\frac{s}{2k}, 1]$.

We will now use the skinny Chebyshev bumps to approximate the function $v(x)$ defined as

$$v(x) \triangleq \begin{cases} \log x - I(2kx) \cdot \sum_{i=0}^{\infty} poi(xk, i) f(i) & \text{for } x \geq \frac{1}{ks} \\ \log(\frac{1}{ks}) - 1 + xsk & \text{for } x \leq \frac{1}{ks} \end{cases}$$

Thus $v(x)$ is twice differentiable for $x > 0$, $v(x) \approx 0$ for $x > \frac{s}{2k}$, $v(x) = \log x$ for $x \in (1/ks, \frac{s}{8k})$, and $v(x)$ is a linear approximation to the logarithm function for $x < 1/ks$.

Define the coefficient b_i of the ith Chebyshev bump B_i, with "center" $c_i = \frac{2s}{k}\left(1 - cos\left(\frac{i\pi}{s}\right)\right)$, to be $v(c_i)$. To conclude the construction, letting the ith Chebyshev bump B_i be represented as a sum of Poisson functions, as guaranteed by Lemma 21: $B_i(x) = \sum_j a_{i,j} poi(xk, j)$, for each $i \in \{0, \ldots, s\}$, increment z_j by $\sum_i a_{i,j} v(c_i)$.

Define the linear estimator given by coefficients $\beta_1, \ldots, \beta_k$, where $\beta_i \triangleq z_{i-1} \cdot \frac{i}{k}$.

The following theorem asserts the quality of our estimator:

Theorem (Theorem 2)**.** *There are positive constants c_1, c_2 such that for any $\epsilon > k^{-c_1}$, the estimator described in Construction 22, when given k independent samples from a distribution of support at most $n = \epsilon c_2 k \log k$, will compute an estimate of the entropy of the distribution, accurate to within ϵ, with probability of failure $o(1/poly(k))$.*

For ease of exposition we prove the theorem for $c_1 = 0.03$, though we suspect this bound can be relaxed considerably.

The proof of Theorem 2 rests on the following lemma, which is the Chebyshev-bump analog of Lemma 19, and shows that if one constructs the naive "plug-in" approximation using the skinny Chebyshev bumps, instead of the Poisson functions, the approximation is very good. While the proof of this lemma is somewhat laborious, the guiding intuition is simply that the Chebyshev bumps are reasonably symmetric and skinny. The proof of Theorem 2, and that of the following lemma are in Section A.1.

Lemma 23. *Given $\alpha \leq \beta$ and a twice-differentiable function $f(x) : [0, \frac{s}{2k}] \to \mathbb{R}$ satisfying $|f(x)| \leq \gamma$, $|f'(x)| \leq \frac{\alpha}{x}$, and $|f''(x)| \leq \frac{\beta}{x^2}$, then $f(x)$ can be approximated as $\sum_i w_i B_i(x)$ for weights $w_i = f(c_i)$ for $c_i = \frac{2s}{k}(1 - \cos\frac{i\pi}{s})$, with error of approximation at x bounded in magnitude by*

$$O(\frac{\gamma}{(xks)^{3/2}}) + O(\frac{\beta}{xks}) + O(\frac{\alpha}{(xks)^{3/2}}) + e^{-s/7}.$$

Additionally, provided $|f(0)| \geq 1 + |f(1/sk)|$, is monotonic decreasing in magnitude for $x < 100/ks$, and $|f(x)| \leq 1/\sqrt{x}$, then for $x \leq 1/ks$, the approximation will be in the range $[f(100/ks), f(0)]$.

A.1 Proofs for the Previous Section

Proof of Lemma 19. Consider the Taylor expansion of f to third order around x, $f(i) \approx a + b \cdot (i - x) + c \cdot (i-x)^2 + d \cdot (i-x)^3 + e(i)$, for $a = f(x), b = f'(x), c = \frac{1}{2}f''(x)$, and $d = \frac{1}{6}f'''(x)$, where the error, e, is a function which we will analyze later. By assumption, $d \leq \frac{\alpha}{6x^3}$. We bound $\sum_{i=\beta}^{\infty} f(i) \cdot poi(x, i)$ by thus decomposing $f(i)$. We note that we may take the lower limit of the sum to be 0, since $f(i)$ equals zero for $i < \beta$. We evaluate the first four terms by noting, respectively, that the Poisson distribution of parameter x has total probability mass 1, has mean x, has variance x, and has third moment about its mean x, leading to $\sum_{i=0}^{\infty} a \cdot poi(x, i) = a = f(x)$, $\sum_{i=0}^{\infty} b(i - x) \cdot poi(x, i) = 0$, $\sum_{i=0}^{\infty} c(i - x)^2 \cdot poi(x, i) = cx = \frac{1}{2}xf''(x)$, and $\sum_{i=0}^{\infty} d(i - x)^3 \cdot poi(x, i) = dx \leq \frac{\alpha}{6x^2}$.

We now analyze the error function $e(i)$. We note, by construction, that it and its first three derivatives are 0 at $i = x$, while its fourth derivative is everywhere equal to the fourth derivative of f, which by assumption is bounded by $\frac{\alpha}{i^4}$. Thus for $i \geq x$, the fourth derivative of $e(i)$ is bounded by $\frac{\alpha}{x^4}$ implying a bound of $|e(i)| \leq \frac{\alpha}{24x^4}(i - x)^4$ for $i \geq x$. Similarly, for $i \in [\frac{x}{2}, x]$ we have that the fourth derivative of f is bounded by $\frac{16\alpha}{x^4}$, yielding a bound of $|e(i)| \leq \frac{2\alpha}{3x^4}(i - x)^4$ for $i \in [\frac{x}{2}, x]$. For general $i < x$, we bound e by repeated integration. Since $|e''''(i)| \leq \frac{\alpha}{i^4}$ and $e'''(x) = 0$ we may integrate from i to x to yield $|e'''(i)| \leq \frac{1}{4}\alpha(\frac{1}{i^3} - \frac{1}{x^3})$, which we crudely bound by $\frac{1}{4}\frac{\alpha}{i^3}$. We repeat this process, since $e''(x) = e'(x) = 0$, to yield, successively, $|e''(i)| \leq \frac{1}{12}\frac{\alpha}{i^2}$, and $|e'(i)| \leq \frac{1}{24}\frac{\alpha}{i}$. We integrate once more, though without discarding the constant term, to yield $|e(i)| \leq \frac{1}{24}\alpha(\log x - \log i)$, again, valid for $i \leq x$. Instead of using this bound directly, we sum from 1 to x:

$$\sum_{i=1}^{x} |e(i)| \leq \frac{1}{24}\alpha \sum_{i=1}^{x} (\log x - \log i) \leq \frac{\alpha}{24} \int_0^x |\log x - \log i| \, di = \frac{\alpha}{24}x.$$

We now bound $e(0)$. If $x < 1$ then, directly, since $e'''' \leq \alpha$, we have $|e(0)| \leq \frac{\alpha}{24}x^4 \leq \frac{\alpha}{24}x$. Otherwise if $x \geq 1$, note from above that $|e(1)| \leq \frac{\alpha \log x}{24}$, $|e'(1)| \leq \frac{\alpha}{24}$, $|e''(1)| \leq \frac{\alpha}{12}$, $|e'''(1)| \leq \frac{\alpha}{4}$, and for all $i \in [0, 1]$, $e''''(i) \leq \alpha$. This immediately yields a bound that $|e(0)| \leq \alpha \left[\frac{\log x}{24} + \frac{1}{24} + \frac{1}{24} + \frac{1}{24}\right]$. Since $3 + \log x \leq 2 + x \leq 3x$ for $x \geq 1$, we have that $\sum_{i=0}^{x} |e(i)| \leq \frac{\alpha}{6}x$.

Trivially, we use this bound to bound the sum over half the domain: $\sum_{i=0}^{x/2} |e(i)| \leq \frac{\alpha}{6}x$. In sum, we will use the bound $|e(i)| \leq \frac{2\alpha}{3x^4}(i - x)^4$ for $i \geq \frac{x}{2}$, and $\sum_{i=0}^{x/2} |e(i)| \leq \frac{\alpha}{6}x$ otherwise.

To complete the proof, we note the basic fact that the Poisson distribution dies off super-polynomially fast away from its mean, relative to its standard deviation. That is, for any positive integer—we choose 6 here—there is a constant γ such that for all i, x, we have $poi(x, i) \leq \frac{\gamma}{\sqrt{x}} \left|\frac{i-x}{\sqrt{x}}\right|^{-6}$.

We thus bound $\sum_{i=0}^{\infty} e(i)poi(x, i)$ piecewise. For $i \in [x - \sqrt{x}, x + \sqrt{x}]$, we have that since $poi(x, i)$ is a distribution over i, it sums to at most 1 here; since we have the bound here that $|e(i)| \leq \frac{2\alpha}{3x^4}(i - x)^4$, we note that when $|i - x| \leq \sqrt{x}$ we have $|e(i)| \leq \frac{2\alpha}{3x^2}$, which is thus also a bound on $\left|\sum_{i=x-\sqrt{x}}^{x+\sqrt{x}} e(i)poi(x, i)\right|$. For $i > x + \sqrt{x}$ we use the bound $poi(x, i) \leq \frac{\gamma}{\sqrt{x}} \left|\frac{i-x}{\sqrt{x}}\right|^{-6}$ to see that

$$\left|\sum_{i>x+\sqrt{x}} e(i)poi(x, i)\right| \leq \sum_{i>x+\sqrt{x}} \frac{2\alpha}{3x^4}(i - x)^4 \cdot \frac{\gamma}{\sqrt{x}} \left|\frac{i-x}{\sqrt{x}}\right|^{-6} = \frac{2\alpha\gamma}{3x^{3/2}} \sum_{i>x+\sqrt{x}} \frac{1}{(i - x)^2} = O(\frac{\alpha}{x^2}).$$

The same argument yields the same bound for the sum over $i \in [\frac{x}{2}, x - \sqrt{x}]$. To bound the remaining region, when $i \leq \frac{x}{2}$, we note that for this region $poi(x, i) \leq \frac{64\gamma}{x^{7/2}}$, and since, as noted, $\sum_{i=0}^{x/2} |e(i)| \leq \frac{\alpha}{6}x$ we have that $\sum_{i=1}^{x/2} |e(i)|poi(x, i) = o(\frac{\alpha}{x^2})$. Combining all the bounds yields that

$|\sum_{i=1}^{\infty} e(i)poi(x,i)| = O(\frac{\alpha}{x^2})$, and combining this with the bounds from the power series expansion of f yields $\sum_{i=0}^{\infty} f(i)poi(x,i)$ equals $f(x) + \frac{1}{2}xf''(x)$ to within $O(\frac{\alpha}{x^2})$, as desired. □

Proof of Lemma 21. Consider decomposing $g_3^i(y)$ into a linear combination of $cos(jy)$, for $j \in \{0,\ldots,s\}$. Since $cos(-jy) = \cos(jy)$, $g_1(y)$ consists of one copy of $\cos(sy)$, two copies of $\cos(jy)$ for each j between 0 and s, and one copy of $\cos(0y)$; $g_2(y)$ consists of 8 copies of $g_1(y)$, with some shifted so as to introduce sine components, but these are canceled out in the formation of $g_3^i(y)$, which is a symmetric function for each i. Thus, together with the normalization by $1/16s$, each $g_3^i(y)$ may be regarded as a linear combination $\sum_{j=0}^{s} \cos(yj)b_{ij}$ where the sth term has coefficient at most $1/s$, and all the remaining terms have coefficients at most $2/s$.

Next, under the coordinate transformation $x = \frac{2s}{k}(1-\cos(y))$, the function $\cos(yj)$ becomes the Chebyshev polynomial $T_j(1-\frac{xk}{2s})$. We note that each term $\alpha_\ell(xk)^\ell$ from this polynomial will ultimately be multiplied by $\sum_{m=0}^{s-1} poi(xk,m)$. We reexpress this as $x^\ell \sum_{m=0}^{s-1} \frac{x^m e^{-x}}{m!} = \sum_{m=\ell}^{s+\ell-1} poi(xk,m)\frac{m!}{(m-\ell)!}$. We have thus expressed our function as a linear combination of Poisson functions. As we aim to bound the sum of the coefficients of these Poisson functions, we consider this now: $\sum_{m=\ell}^{s+\ell-1} \frac{m!}{(m-\ell)!}$ which we note equals $\frac{1}{l+1}\frac{(s+\ell)!}{s!}$ since, in general, $\sum_{m=i}^{j}\binom{m}{i} = \binom{j+1}{i+1}$. Expressing $T_j(z)$ as $\sum_{i=0}^{j}\beta_{ij}z^i$, we note that, since we evaluate Chebyshev polynomials at $1-\frac{xk}{2s}$, a term $\beta_{ij}z^i$ becomes $\beta_{ij}\sum_{\ell=0}^{i}\binom{i}{\ell}\frac{1}{(2s)^\ell}x^\ell$, which, by the previous calculation, contributes $\beta_{ij}\sum_{\ell=0}^{i}\binom{i}{\ell}\frac{1}{(2s)^\ell}\frac{1}{l+1}\frac{(s+\ell)!}{s!}$ to the total Poisson coefficients. Since $\ell \le i \le s$, we have $s+\ell \le 2s$, from which we see $\frac{1}{(2s)^\ell}\frac{(s+\ell)!}{s!} \le 1$. We thus bound $\beta_{ij}\sum_{\ell=0}^{i}\binom{i}{\ell}\frac{1}{(2s)^\ell}\frac{1}{l+1}\frac{(s+\ell)!}{s!} \le \beta_{ij}\sum_{\ell=0}^{i}\binom{i}{\ell} = \beta_{ij}2^i$.

We thus desire, for any $j \le s$, to bound $\sum_{i=0}^{j}\beta_{ij}2^i$, where β_{ij} are the coefficients of the jth Chebyshev polynomial. Chebyshev polynomials have coefficients whose signs repeat in the pattern $(+,0,-,0)$, thus we can evaluate this sum exactly as $|T_j(2\mathbf{i})|$, for $\mathbf{i} = \sqrt{-1}$. Explicitly, $|T_j(2\mathbf{i})| = \frac{1}{2}\left[(2-\sqrt{5})^j + (2+\sqrt{5})^j\right] \le (2+\sqrt{5})^j$. Since, as we showed above, in each $g_3^i(y)$ the coefficient of each $cos(jy)$ term is at most $2/s$, and thus our final bound on the sum of Poisson coefficients is $2(2+\sqrt{5})^s < k^{0.4}$ as desired. □

Lemma 24. *For any x,*

$$\sum_{i=-s}^{s-1} g_2(x + \frac{\pi i}{s}) = 1.$$

Proof. From the above definition, $g_2(y)$ 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 $\sum_{i=-s}^{s-1} g_2(x+\frac{\pi i}{s})$ 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. □

Lemma 25. $|g_2(y)| \le \frac{285}{y^4 s^4}$ *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:

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

Note that $\left(\cot(\frac{y}{2}-\frac{3\pi}{4s})-3\cot(\frac{y}{2}-\frac{\pi}{4s})+3\cot(\frac{y}{2}+\frac{\pi}{4s})-\cot(\frac{y}{2}+\frac{3\pi}{4s})\right)$ is bounded in magnitude by $(\pi/2s)^3$ times the maximum magnitude of $\frac{d^3}{dx^3}\cot(x/2)$ in the range $x\in[y-3\pi/2s, y+3\pi/2s]$. Since the magnitude of this third derivative is decreasing for $x\in(0,2\pi)$, we can simply evaluate the magnitude of this derivative at $y-3\pi/2s$. We thus have $\frac{d^3}{dx^3}\cot(x/2)=\frac{-(2+\cos(x))}{4\sin^4(x/2)}$, whose magnitude is at most $\frac{3}{4(x/\pi)^4}$ for $x\in(0,\pi]$. Thus for $y\in[3\pi/s,\pi]$, since for such y we trivially have that $y/2\leq y-3\pi/2s$, we have the following bound:

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

Since $g_2(y)$ is a symmetric function, the same bound holds for $y\in[-\pi,-3\pi/s]$. Thus $|g_2(y)|\leq\frac{3\pi^3}{16s\cdot 2y^4s^3}<\frac{285}{y^4s^4}$ for $y\in[-\pi,\pi]\backslash(-3\pi/s,3\pi/s)$. To conclude, note that $g_2(y)$ attains a global maximum at $y=0$, with $g_2(0)=\frac{1}{16s}\left(6\cot(\pi/4s)-2\cot(3\pi/4s)\right)\leq\frac{1}{16s}\frac{24s}{\pi}<1/2$. □

Lemma 26. *For $y\in[-\pi/s,\pi/s]$, sufficiently large s, and positive integers $a,b\leq s$,*

$$|\sum_{i=-a}^{b}(y+\pi i/s)\cdot g_2(y+\pi i/s)|\leq\frac{12}{s}\left(\frac{1}{a^2}+\frac{1}{b^2}\right).$$

Proof. We will first show that

$$|\sum_{i=-s}^{s-1}\left(\sin(y+\pi i/s)+\frac{\sin^3(y+\pi i/s)}{6}\right)\cdot g_2(y+\pi i/s)|=0,$$

and then will use the fact that $\sin(x)+\sin^3(x)/6\approx x$ near zero, and that $g_2(x)$ decays quickly away from zero to yield the claim. To begin, note that $g_2(x)$ is an even function, and can be written as a weighted sum of $\cos(jx)$, for integers j at most $s-1$. Since $\cos(jx)\sin(x)=\frac{1}{2}\sin((j+1)x)-\frac{1}{2}\sin((j-1)x)$, and $\sum_{i=-s}^{s-1}\sin(j(x+\frac{i\pi}{s}))=0$, for any integer $j\leq 2s-1$, we have

$$\sum_{i=-s}^{s-1}\sin(y+\pi i/s)\cdot g_2(y+\pi i/s)=0.$$

Additionally, $\sin^3(x)=\frac{3\sin(x)-\sin(3x)}{4}$, and by the above, $\cos(jx)\sin(3x)=\frac{1}{2}\sin((j+3)x)-\frac{1}{2}\sin((j-3)x)$, and thus for $s>3$, by the above,

$$\sum_{i=-s}^{s-1}\sin^3(y+\pi i/s)\cdot g_2(y+\pi i/s)=0.$$

Next, note that $|x-\sin(x)-\sin^3(x)/6|\leq 3x^5/40$, and thus from the above,

$$|\sum_{i=-s}^{s-1}(y+\pi i/s)\cdot g_2(y+\pi i/s)|\leq\sum_{i=-s}^{s-1}|g_2(y+\pi i/s)|\cdot|3(y+\pi i/s)^5/40|.$$

We now leverage the bounds on $|g_2(y)|$ from Lemma 25. For the at most 5 terms in the above sum for which $y+\pi i/s\in(-3\pi/s,3\pi/s)$, since $g_2(y)\leq 1/2$, we get a contribution of at most $\frac{5}{2}\frac{3^5\pi^5}{40s^5}\leq\frac{4700}{s^5}$. For

the remaining terms, we have $|g_2(x+\pi i/s)| \le \frac{285}{(x+\pi i/s)^4 s^4}$, and thus the contribution of the remaining terms, since $|y| < \pi/s$, is at most $2\sum_{i=2}^{s} \frac{855(\pi i/s)}{40s^4} \le \frac{43 \log s}{s^5}$. Thus for sufficiently large s,

$$|\sum_{i=-s}^{s-1} (y+\pi i/s)\cdot g_2(y+\pi i/s)| \le \frac{1}{s^4}.$$

To conclude, the claim clearly holds for $a = 1$ or 2, and for $a \ge 3$ we have

$$\begin{aligned}\sum_{i=a}^{s} |(y+\pi i/s)\cdot g_2(y+\pi i/s)| &\le \sum_{i=a}^{s} (y+\pi i/s)\frac{285}{(y+\pi i/s)^4 s^4} \\ &\le \frac{285}{\pi^3 s}\sum_{i=a-1}^{s}\frac{1}{i^3} < \frac{23}{2a^2 s}.\end{aligned}$$

□

Lemma 27. *For $y \in [-\pi/s, \pi/s]$, sufficiently large s, and positive integer $a \le s$,*

$$|\sum_{i=0}^{a} (y+\pi i/s)^2 \cdot g_2(y+\pi i/s)| \le \frac{152}{s^2}.$$

Proof. From our bounds on g_2 given in Lemma 25, have the following:

$$\begin{aligned}|\sum_{i=0}^{a} (y+\pi i/s)^2 \cdot g_2(y+\pi i/s)| &\le 3\cdot\frac{1}{2}\cdot\frac{3^2\pi^2}{s^2} + \sum_{i=3}^{a}(y+i\pi/s)^2\frac{285}{(y+i\pi/s)^4 s^4} \\ &\le \frac{27\pi^2}{2s^2} + \frac{285}{\pi^2 s^2}\sum_{i=3}^{\infty}\frac{1}{(i-1)^2} \le \frac{152}{s^2}.\end{aligned}$$

□

Lemma 28. *For $y \in [-\pi/s, \pi/s]$, sufficiently large s, and positive integers a, b, c, d such that $c \le a \le s$ and $d \le b \le s$, and a twice-differentiable function $f : [-\frac{a\pi}{s}, \frac{b\pi}{s}] \to \mathbf{R}$ satisfying $|f'(0)| \le \alpha$, $\max_{y\in[-c\pi/s, d\pi/s]} |f''(y)| \le \beta$, and $\max_y |f(y)| \le \gamma$,*

$$|\sum_{i=-a}^{b} g_2(y+\frac{i\pi}{s})f(y+\frac{i\pi}{s}) - f(0)| \le 12\gamma\left(\frac{1}{c^3}+\frac{1}{d^3}\right) + \frac{12\alpha}{s}\left(\frac{1}{c^2}+\frac{1}{d^2}\right) + \frac{304\beta}{s^2}.$$

Proof. We first bound the contribution of the terms with $i \in -a, \ldots, -c, d, \ldots, b$. Using the bounds on $|g_2|$ from Lemma 25, we have

$$|\sum_{i\in -a,\ldots,-c,d,\ldots,b} g_2(x+\frac{i\pi}{s})f(x+\frac{i\pi}{s})| \le \gamma\frac{285}{\pi^4}\left(\sum_{i=c}^{\infty}\frac{1}{(i-1)^4} + \sum_{i=d}^{\infty}\frac{1}{(i-1)^4}\right) \le 10\gamma(1/c^3 + 1/d^3).$$

We now consider $\sum_{i=-c}^{d} g_2(y+\frac{i\pi}{s})f(y+\frac{i\pi}{s})$. We express each $f(y+\frac{i\pi}{s})$ in terms of the first order Taylor expansion about 0, and note that $|f(y+\frac{i\pi}{s}) - (f(0)+(y+\frac{i\pi}{s})f'(0))| \le (y+\frac{i\pi}{s})^2\beta$. Thus we have the following:

$$\begin{aligned}|\sum_{i=-c}^{d} g_2(y+\frac{i\pi}{s})f(y+\frac{i\pi}{s}) - \sum_{i=-c}^{d} g_2(y+\frac{i\pi}{s})\left(f(0)+(y+\frac{i\pi}{s})f'(0)\right)| &\le \beta\sum_{i=-c}^{d} g_2(y+\frac{i\pi}{s})(y+\frac{i\pi}{s})^2 \\ &\le 2\beta\frac{152}{s^2} \qquad \text{from Lemma 27.}\end{aligned}$$

We now turn to analyzing the term involving the Taylor approximation:

$$\sum_{i=-c}^{d} g_2(y+\frac{i\pi}{s})\left(f(0)+(y+\frac{i\pi}{s})f'(0)\right) \;=\; \sum_{i=-c}^{d} g_2(y+\frac{i\pi}{s})f(0)+f'(0)\sum_{i=-c}^{d} g_2(y+\frac{i\pi}{s})(y+\frac{i\pi}{s}).$$

To analyze the first term above, by Lemma 24, $\sum_{i=-s}^{s} g_s(y+\frac{i\pi}{s})f(0) = f(0)$. Additionally, by Lemma 25, $\sum_{d+1}^{s-1} g_s(y+\frac{i\pi}{s}) \leq \sum_{d}^{\infty} \frac{285}{\pi^4 i^4} \leq \frac{2}{d^3}$, and analogously, $\sum_{-s}^{-c-1} g_s(x+\frac{i\pi}{s}) \leq \frac{2}{c^3}$. Thus

$$|f(0)-\sum_{i=-c}^{d} g_2(y+\frac{i\pi}{s})f(0)| \leq 2f(0)(\frac{1}{c^3}+\frac{1}{d^3}).$$

To analyze the second term, by Lemma 26,

$$|\sum_{i=-c}^{d}(y+\frac{i\pi}{s})f'(0)g_2(y+\frac{i\pi}{s})| \leq \frac{12}{s}f'(0)\left(\frac{1}{c^2}+\frac{1}{d^2}\right).$$

The desired statement now follows from adding up the above bounds. □

Lemma 29. *For $y \in [0,\pi/2]$, sufficiently large s, and twice-differentiable function f satisfying $|f(y)| \leq \gamma$, $|f'(y)| \leq \frac{\alpha}{y}$ and $|f''(y)| \leq \frac{\beta}{y^2}$,*

$$|f(y)-\sum_{i=0}^{s} g_3^i(y)f(\frac{i\pi}{s})| \leq 100000\left(\frac{\gamma}{y^3s^3}+\frac{\beta}{y^2s^2}+\frac{\alpha}{y^3s^3}\right).$$

Proof. From Lemma 25, we have $g_3^0(y)f(0)+g_3^s(y)f(\pi) \leq 2\gamma\frac{285}{y^4s^4}$.

Next, define $i_y \triangleq \lfloor \frac{ys}{\pi} \rfloor$, and let $\delta_y := y-\frac{i_y\pi}{s}$. Thus $\delta_y \in [0,\pi/s]$. For any $j \in -i_y+1,\ldots,s-i_y-1$, we have

$$\begin{aligned} g_3^{i_y+j}(y) &= g_2(y-\frac{(i_y+j)\pi}{s})+g_2(y+\frac{(i_y+j)\pi}{s}) \\ &= g_2(\delta_y-\frac{j\pi}{s})+g_2(\delta_y+\frac{(2i_y+j)\pi}{s}). \end{aligned}$$

Defining the function $r_y(w) = f(y-w)$, we have the following:

$$\begin{aligned} \sum_{i=1}^{s-1} g_3^i(y)f(\frac{i\pi}{s}) &= \sum_{i=1}^{s-1}\left(g_2(y-\frac{i\pi}{s})+g_2(y+\frac{i\pi}{s})\right)r_y\left(-(\frac{i\pi}{s}-y)\right) \\ &= \sum_{i=1}^{s-1}\left(g_2(\delta_y+\frac{(i_y-i)\pi}{s})+g_2(y+\frac{i\pi}{s})\right)r_y\left(\delta_y+\frac{(i_y-i)\pi}{s}\right) \\ &= \sum_{j=-s+i_y+1}^{i_y-1} g_2(\delta_y+\frac{j\pi}{s})r_y(\delta_y+\frac{j\pi}{s})+\sum_{i=1}^{s-1} g_2(y+\frac{i\pi}{s})r_y\left(\delta_y+\frac{(i_y-i)\pi}{s}\right). \end{aligned}$$

The idea now is that Lemma 28 guarantees that the first term above is roughly $r_y(0) = f(y)$, and it is easy to show that the second term above will be very small. We start by bounding the magnitude of the second term, using the bound on g_2 given in Lemma 25:

$$\sum_{i=1}^{s-1} g_2(y+\frac{i\pi}{s})r_y\left(\delta+\frac{(i_y-i)\pi}{s}\right) \;\leq\; \gamma/i_y^3.$$

We now consider the first term above, and apply Lemma 28 with $a = c = |-s+i_y+1|$, $d = \lfloor i_y/2 \rfloor$, and $b = i_y - 1$ to yield:

$$|f(y) - \sum_{j=-s+i_y+1}^{i_y-1} g_2(\delta_y + \frac{j\pi}{s}) r_y(\delta_y + \frac{j\pi}{s})| \quad \leq \frac{24\gamma}{i_y^3} + \frac{16 \cdot 304}{s^2}\frac{\beta}{y^2} + \frac{2 \cdot 12}{s i_y^2}\frac{\alpha}{y},$$

from which the desired claim follows. □

Fact 30. *(From [18]) For $\lambda > 0$, and an integer $n \geq 0$, if $n \leq \lambda$,*

$$\sum_{i=0}^{n} poi(\lambda, i) \leq \frac{poi(\lambda, n)}{1 - n/\lambda},$$

and for $n \geq \lambda$,

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

Lemma 31. *For $x \leq \frac{s}{2k}$, $1 - \sum_{i=0}^{s-1} poi(xk, i) \leq e^{-s/6}$.*

Proof. This discrepancy is maximized at $x = \frac{s}{2k}$, and by tail bounds of Poissons and Fact 30, this is at most

$$\sum_{i=s}^{\infty} poi(s/2, i) \leq 2poi(s/2, s) \leq e^{-s/6}.$$

□

Proof of Lemma 23. Recall from Definition 20 that $B_i(x)$ is related to $g_3^i(y)$ by the coordinate transformation $x = \frac{2s}{k}(1 - \cos(y))$, and scaling by $\sum_{j=0}^{s-1} poi(xk, j)$. By Lemma 31 we can ignore the scaling factor for $x \leq \frac{s}{2k}$ and lose only $s \cdot (1 + \log ks) \cdot e^{-s/6} < e^{-s/7}$ in approximation, since there are s skinny bumps, and in Construction 22 each skinny bump has a coefficient of magnitude at most $\max_x |v(x)| = 1 + \log ks + e^{-s/7}$. To represent $f(x)$ as a linear combination of $B_i(x)$'s, we will represent $r(y)$ as a linear combination of $g_3^i(y)$'s, where r is chosen so that $r(y) = f(\frac{2s}{k}(1 - \cos(y)))$. Note that

$$\begin{aligned} |r'(y)| &\leq |f'(\frac{2s}{k}(1 - \cos(y)))\frac{2s}{k}\sin(y)| \\ &\leq \frac{\alpha k}{2s(y^2/3)}\frac{2s}{k}y \qquad \text{since for } y \in [0, \pi/2], 1 - \cos(y) \geq y^2/3, \text{ and } \sin(y) \leq y. \\ &= \frac{3\alpha}{y}. \end{aligned}$$

Similarly,

$$\begin{aligned} |r''(y)| &\leq |f''(\frac{2s}{k}(1 - \cos(y)))(\frac{2s}{k}\sin(y))^2 + f'(\frac{2s}{k}(1 - \cos(y)))(\frac{2s}{k}\cos(y))| \\ &\leq \frac{9\beta}{y^2} + \alpha \leq \frac{30\beta}{y^2}. \end{aligned}$$

Thus by Lemma 29, we can approximate $r(y)$ as a linear combination of $g_3^i(y)$ to within error $O(\frac{\gamma}{y^3 s^3}) + O(\frac{\beta}{y^2 s^2}) + O(\frac{\alpha}{y^3 s^3}) + e^{-s/7}$. For $y \in [0, \pi/2]$, note that $(1 - \cos(y)) \in [y^2/3, y^2/2]$ and thus

the error in the corresponding approximation of $f(x)$ via the linear combination of $B_i(x)$'s will have error at most

$$O(\frac{\gamma}{(\sqrt{\frac{3xk}{2s}})^3 s^3}) + O(\frac{\beta}{(\sqrt{\frac{3xk}{2s}})^2 s^2}) + O(\frac{\alpha}{(\sqrt{\frac{3xk}{2s}})^3 s^3}) + e^{-s/7},$$

$$= \quad O(\frac{\gamma}{(xks)^{3/2}} + O(\frac{\beta}{xks}) + O(\frac{\alpha}{(xks)^{3/2}}) + e^{-s/7},$$

as desired.

We now turn to bounding the approximation of $f(x)$ for small $x \leq 1/ks$, which thus equates to bounding the approximation of $r(y)$ via the $g_3^i(y)$'s for $y < 2/s$. The desired lemma now follows from noting that the approximation of $r(y)$ for such values of y is a convex combination of $r(i\pi/s)$ for $i \in 0, 1, 2, \ldots$, where the weight on $r(0)$ is trivially seen to be at least .1, and the contribution to the approximation from g_3^j for $j \geq 100$ is bounded by $\sum_{j \geq 101} g_3^j(y) r(i\pi/s) \leq .1$, from Lemma 25 and the assumption that $|f(x)| \leq 1/x^{.5}$. □

Proof of Theorem 2. Consider the function $f(x) \triangleq I(x) \left[\log x - \log k + \frac{1}{2x}\right]$, and note that it satisfies the conditions of Lemma 19, with $\alpha = O(1)$, and thus

$$|\sum_{i=0}^{\infty} f(i) \cdot poi(x, i) - \left(f(x) + \frac{1}{2} x f''(x)\right)| \leq O(1/x^2).$$

For $x > s/2$, we have $I(x) = 1$ and thus for such x

$$f(x) + \frac{1}{2} x f''(x) = \log x - \log k + O(\frac{1}{x^2}).$$

Thus via the change of variables $y = \frac{x}{k}$, we have that for $y \in [\frac{s}{2k}, \infty]$,

$$|\log y - \sum_{i=0}^{\infty} poi(yk, i) f(i)| \leq O(\frac{1}{k^2 y^2}).$$

Thus we have accurately represented the logarithm function via the Poisson bumps in the interval $[\frac{s}{2k}, 1]$.

We now consider the Chebyshev-bump approximation of the function $v(y)$ defined above as

$$v(y) \triangleq \begin{cases} \log y - I(2ky) \cdot \sum_{i=0}^{\infty} poi(yk, i) f(i) & \text{for } y \geq \frac{1}{ks} \\ \log(\frac{1}{ks}) - 1 + ysk & \text{for } y \leq \frac{1}{ks} \end{cases}$$

Note that $v(y)$ satisfies the conditions of Lemma 29 with $\gamma < \log(sk) + 2$ and $\alpha, \beta = O(1)$. Thus $v(y)$ can be accurately represented by $\sum_i B_i(y) v(c_i)$, yielding that for sufficiently large k,

$$|\sum_{i=1}^{s} B_i(y) v(c_i) + \sum_{i=1}^{\infty} poi(yk, i) f(i) - \log(y)| \leq \begin{cases} \log(y) + \log(ks) + O(1) & \text{for } y \leq \frac{1}{ks} \\ O(\frac{1}{yks}) + e^{-s/7} & \text{for } y \in (\frac{1}{ks}, \frac{s}{2k}) \\ O(\frac{1}{k^2 y^2}) & \text{for } y \geq \frac{s}{2k}. \end{cases}$$

We will now apply Proposition 17 with $a = O(\epsilon), b = O(\epsilon/n)$, and $c = k^{-0.09}$. Note that by Lemma 21, the coefficients are sufficiently small and vary sufficiently slowly, satisfying the second condition of Proposition 17. For the first condition of Proposition 17, it suffices to show that $err(y) \leq$

ϵ for $y \geq \frac{1}{n}$, and $err(y) \leq \frac{\epsilon}{yn}$ for $y \leq \frac{1}{n}$. To show this, consider setting $n = \epsilon ks$. For $y \leq 1/ks$, since $y < 1/n$, we have

$$\begin{aligned}(ny)\left(\log(y) + \log(ks) + O(1)\right) &\leq (\epsilon ksy)\left(\log(ksy) + O(1)\right) \\ &\leq \epsilon\left(\log(ksy)ksy + O(1)ksy\right) \\ &\leq \epsilon O(1),\end{aligned}$$

and thus the error in this region is good enough to yield an $O(\epsilon)$ estimator. For $y \in (\frac{1}{ks}, \frac{s}{2k})$, $e^{-s/7} = O(k^{-.04}) = o(\epsilon)$, and for $y > 1/n$, we have error of approximation of the logarithm function at most $O(n/ks) = O(\epsilon)$, and if $y < 1/n = 1/\epsilon ks$, we have $ny \cdot O(1/yks) = O(\epsilon)$, which is sufficient to yield an $O(\epsilon)$ estimator. Finally, in the region $y \geq \frac{s}{2k}$, if $y > 1/n$, which implies that $\epsilon > 1/yks$, we have error $O(1/k^2y^2) = O(1/yks) \cdot \frac{s}{yk}$. Because of our bound on y, $s/yk \leq 2$, and thus this error is $O(1/yks) = O(\epsilon)$. In the case that $y \leq 1/n$, we have $ny \cdot O(1/k^2y^2) \leq \epsilon ksO(1/k^2y) = O(\epsilon s/ky) \leq O(\epsilon)$, again because of our bound on y. Thus the above approximation scheme of the logarithm function is sufficiently accurate to yield $O(\epsilon)$-error estimators of entropy for distributions of support at most $O(\epsilon k \log k)$. □

B Testing Uniformity

The machinery developed in the previous section for constructing our estimator for entropy is quite robust. Here, we leverage this machinery—in particular, the "Chebyshev Bump" scheme—to yield a linear estimator for *distance to uniformity*. While distance to uniformity is not a linear property, there exists a linear property that is easily seen to be a 2-approximation:

Fact 32. *The total variation distance between a discrete distribution $p \in \mathcal{D}^n$ with histogram h and a uniform distribution on m elements, denoted by $D(h, Unif(m))$, can be approximated to within a factor of 2 as $\sum_{x:h(x)\neq 0} h(x) f_u(x)$, for the function*

$$f_u(x) \triangleq \begin{cases} x & \text{for } x \leq \frac{1}{2m} \\ |x - \frac{1}{m}| & \text{for } x > \frac{1}{2m}. \end{cases}$$

To see the intuition behind this, note that to calculate the distance between a distribution p and the uniform distribution on m elements, one takes the m elements p_i of h that have the highest probability, and computes the cost of changing each of their probability masses to $\frac{1}{m}$, namely $|p_i - \frac{1}{m}|$, and then adds to this the cost of changing every other mass p_i to 0, namely $|p_i|$. This is lower-bounded and 2-approximated by the cost of sending every element that is below $\frac{1}{2m}$ down to 0, and sending every element above $\frac{1}{2m}$ to $\frac{1}{m}$, as defined in Fact 32.

As for estimating entropy, we will use the Chebyshev bump construction of Definition 20 to help approximate the function $\frac{f_u(x)}{x}$ via a sum of Poisson functions, then apply Proposition 17.

Construction 33. *Let $s = (0.3)\log k$. Define the interpolation function $I : \mathbb{R} \to \mathbb{R}$ such that $I(x) = 0$ for $x \leq \frac{s}{4}$, $I(x) = 1$ for $x \geq \frac{s}{2}$, and $I(x)$ is continuous, and four-times differentiable, where for $i \in 1, \ldots, 4$, the magnitude of the ith derivative is at most c/s^i, for some fixed constant c. Such a function I can be easily constructed.*

Consider the function $g(x) \triangleq I(kx)\frac{f_u(x)}{x}$, and provisionally set $z_i \triangleq g(\frac{i}{k})$. We will now use the skinny Chebyshev bumps to approximate the function $v(x) = (1 - I(kx))\frac{f_u(x)}{x}$.

Define the coefficient of the ith Chebyshev bump B_i, with "center" $c_i = \frac{2s}{k}\left(1 - cos\left(\frac{i\pi}{s}\right)\right)$, to be $v(c_i)$. To conclude the construction, letting the ith Chebyshev bump B_i be represented as a sum of

Poisson functions, as guaranteed by Lemma 21: $B_i(x) = \sum_j a_{i,j} poi(xk, j)$, *for each* $i \in \{0, \ldots, s-1\}$, *increment* z_j *by* $\sum_i a_{i,j} v(c_i)$.

Define the linear estimator given by coefficients $\beta_1, \ldots, \beta_k$, *where* $\beta_i \triangleq z_{i-1} \cdot \frac{i}{k}$.

The following theorem asserts the quality of our estimator:

Theorem (Theorem 3)**.** *For any* $\epsilon > \frac{1}{4 \log m}$, *the estimator described in Construction 33, when given* $O\left(\frac{1}{\epsilon^2} \cdot \frac{m}{\log m}\right)$ *independent samples from a distribution of* any *support, will compute the* L_1 *distance to* $Unif(m)$ *to within accuracy* ϵ, *with probability of failure* $o(1/poly(m))$.

The lower-bound construction in [33] shows that this is tight for constant ϵ; in particular, for any constant $\epsilon > 0$, there is a constant c such that for sufficiently large m, there exist two distributions A, A' such that $D(A, Unif(m)) < \epsilon$, $D(A', Unif(m)) > 0.49$, but distributions A, A' are indistinguishable (with probability at least 2/3,) given $c\frac{m}{\log m}$ samples.

We note that, in contrast to the estimator for entropy, the estimator for distance to uniformity does not need any assumption on the support size of the distribution being sampled. Additionally, the convergence rate is as the inverse of the square root of the number of samples, as opposed to the much faster inverse linear relationship of the estimator for entropy. Intuitively, this is because the function $f_u(y)$ has a kink at probability $y = 1/m$, as opposed to the smooth logarithm function.

The proof of Theorem 3 is considerably easier than for our estimator of entropy:

Proof of Theorem 3. Consider setting $m = \epsilon^2 k \log k$, for some $\epsilon > \frac{4}{\log m}$, and thus the portion of $f_u(x)/x$ approximated exclusively by the Poisson bumps (ie $x > \frac{s}{2k}$) corresponds to $x > \frac{1}{m}$, and in this range $f_u(x)/x = 1 - \frac{1}{xm}$. In particular, the function $\frac{f_u(x/k)}{x/k}$ has jth derivative bounded in magnitude by $O(\frac{k}{mx^{j+1}})$, for constant j, and thus satisfies the conditions of Lemma 19 with $\alpha = O(\frac{k}{ms})$, and thus the approximation in this regime is accurate to $O(\frac{k}{msx^2}) + O(\frac{x}{2}\frac{k}{mx^3}) = O(\frac{k}{mx^2})$, which is maximized by minimizing x, in which case the error is $O(\frac{k}{ms^2}) = O(\frac{1}{\epsilon^2 \log^3 k})$, which is at most $O(\epsilon)$, as in the case that $\epsilon = 1/\log k$.

We now consider the error in approximation from the skinny bumps (ie for $x < \frac{s}{2k}$. In this regime, the function $f_u(x)/x$ is $O(\frac{1}{mx^2})$-Lipschitz for $x > 1/2m$. By Lemma 25 (arguing that the functions g_3^i decay super quadratically), Lemma 31, and the change of coordinates, the width of the Chebyshev bumps centered at x are $O(\frac{\sqrt{xk \log k}}{k \log k})$; thus the error of approximation is the product of this width and the Lipschitz constant, yielding $O(\frac{1}{\epsilon^2 (xk \log k)^{3/2}})$. This is maximized by minimizing x, and thus taking $x = O(1/m)$ yields error $O(\epsilon)$, as desired. Since $f_u(x)/x = 1$ is constant for $x < 1/2m$, the error in this small regime is $o(\epsilon)$. Thus the error of approximating the function $f_u(x)/x$ is $O(\epsilon)$. To conclude, since the coefficients of the approximation are sufficiently small (bounded via Lemma 21 as in the estimator for entropy), we may now apply Proposition 17 to yield the claim. □

C Properties of Pairs of Distributions

Perhaps unsurprisingly, our very general framework for constructing constant-factor optimal estimators for symmetric properties of distributions can also be extended to yield constant-factor optimal estimators for symmetric (label-permutation invariant) properties of pairs of distributions, which includes properties such as variational distance (L_1 distance), and Kullback–Leibler divergence between two distributions.

For these properties, one is given as input a set of k_1 samples from distribution A, and k_2 samples from distribution B. In analogy with the case for properties of a single distribution, we note that for

both lower and upper bounds, up to constant factor it suffices to consider the Poissonized setting in which one takes $Poi(k_1)$ samples from A, and $Poi(k_2)$ samples from distribution B.

Definition 34. *The* fingerprint $\mathcal{F}$ *of a set of k_1 samples from A and k_2 samples from B is a $k_1 \times k_2$ matrix, whose entry $\mathcal{F}(i,j)$ is given by the number of domain elements that are seen exactly i times in the samples from A and exactly j times in the samples from B.*

Definition 35. *The* histogram $h_{A,B} : [0,1]^2 \to \mathbb{N} \cup 0$ *of a pair of distributions A, B is defined by letting $h_{A,B}(x,y)$ be the number of domain elements that occur with probability x in distribution A and probability y in distribution B.*

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 full set of k_1 samples from A and k_2 samples from B, there is an estimator of equivalent performance that takes as input the fingerprint $\mathcal{F}$ derived from such a set of samples.

Definition 36. *A property π is a* linear property *of a pair of distributions if there exists a function $f : [0,1]^2 \to \mathbb{R}$ such that*

$$\pi(A,B) = \sum_{(x,y):h_{A,B}(x,y)\neq 0} h(x,y) f(x,y).$$

For properties of single distributions, Proposition 17 shows that if one can accurately approximate $f(x)/x$ as a sum of poisson functions $poi(xk,i)$ with well-behaved coefficients, then the corresponding estimator will be accurate, when given a set of $Poi(k)$ samples. Here, we show that if the function $\frac{f(x,y)}{x+c\cdot y}$ can be accurately approximated by $\sum_{i,j\geq 0} \beta_{i,j} \cdot poi(kx,i) \cdot poi(cky,j)$, then the corresponding property can be accurately approximated given k samples from A and ck samples from B. For clarity of exposition, we only state the proposition for $c = 1$, and we give the analog of Proposition 17.

Proposition 37. *Let π be a linear symmetric property of pairs of distributions such that for any histogram $h_{A,B}$ corresponding to a pair of distributions A, B, we have $\pi(A,B) \triangleq \sum_{(x,y):h_{A,B}(x,y)\neq 0} h_{A,B} f_\pi(x,y)$, for some function $f_\pi : [0,1]^2 \setminus \{(0,0)\} \to \mathbb{R}$. Given integers k, n and a set of coefficients $z_{i,j}$ for all integers $i,j \geq 0$, we define the function $err : [0,1]^2 \setminus \{(0,0)\} \to \mathbb{R}$ by*

$$\frac{f_\pi(x,y)}{x+y} = err(x,y) + \sum_{i,j\geq 0} z_{i,j} poi(kx,i) poi(ky,j).$$

For all integers $i,j \leq k$ with $(i,j) \neq (0,0)$, define $\beta_{i,j} \triangleq z_{i-1,j}\frac{i}{k} + z_{i,j-1}\frac{j}{k}$. If, for positive real numbers a, b, c, the following conditions hold:

- $|err(x,y)| < a + \frac{b}{x+y}$,
- *for any $(i,j), (i',j')$ such that $|i-i'| \leq \sqrt{i}\log k$, and $|j-j'| \leq \sqrt{j}\log k$, we have $|\beta_{i,j} - \beta_{i',j'}| \leq c\frac{\sqrt{i}+\sqrt{j}}{\sqrt{k}}$,*

then the linear estimator given by coefficients $\beta_{i,j}$ when given a fingerprint derived from a set of k independent samples from A and k independent samples from B, that are distributions of support at most n, will estimate the property value with error at most $a + bn + c\log k$, with probability of failure $o(1/poly(k))$.

Proof.

$$\begin{aligned}\pi(A,B) &= \sum_{(x,y):h(x,y)\neq 0} h(x,y) f_\pi(x,y) \\ &= \sum_{(x,y):h(x,y)\neq 0} h(x,y)(x+y)\cdot err(x,y) + \sum_{(x,y):h(x,y)\neq 0}\sum_{i,j\geq 0} \beta_{i,j} h(x,y) poi(kx,i)\cdot poi(ky,j)(x+y).\end{aligned}$$

We first analyze the second term in the above expression:

$$\begin{aligned}&\sum_{(x,y):h(x,y)\neq 0}\sum_{i,j\geq 0} \beta_{i,j} h(x,y) poi(kx,i)\cdot poi(ky,j)(x+y) \\ &= \sum_{(x,y):h(x,y)\neq 0}\sum_{i,j\geq 0} \beta_{i,j}\left(h(x,y) x poi(kx,i) poi(ky,j) + h(x,y) poi(kx,i)\cdot y\cdot poi(ky,j)\right) \\ &= \sum_{i,j\geq 0} \beta_{i,j} h(x,y)\left(\frac{i+1}{k} poi(kx,i+1) poi(ky,j) + \frac{j+1}{k} poi(kx,i) poi(ky,j+1)\right) \\ &= \sum_{i\geq 1,j\geq 0} \beta_{i-1,j} E[\mathcal{F}(i,j)]\frac{i}{k} + \sum_{i\geq 0,j\geq 1} \beta_{i,j-1} E[\mathcal{F}(i,j)]\frac{j}{k}.\end{aligned}$$

Thus this term is simply the expected value of our linear estimator. Note that we started with an approximation that used all the products of the Poisson functions, even $poi(xk,0)\cdot poi(yk,0)$, and yielded an expression which is the sum of all fingerprint expectations *except* the $(0,0)$th–specifically, the sum involves exactly those fingerprint entries which we have access to.

This expected value of the estimator, together with the bounds on the magnitude of the the error term, derived from the first conditions of the proposition as was done in Proposition 17, yields that the estimator has bias at most $a+bn$. Additionally, exactly as in the proof of Proposition 17, the bounds on the variation in the coefficients given by the second condition yield the desired bounds on the variance of the estimator. □

We now describe an explicit linear estimator for L_1 distance.

Construction 38. *Let $s = \frac{1}{15}\log k$, and for $i,j \leq s$, let $B_{i,j} \triangleq \sum_{i',j'} a_{i,i'} a_{j,j'} poi(xk,i') poi(yk,j')$, be our "2-dimensional Chebyshev bumps," where $a_{i,i'}$ is defined to be the coefficient of $poi(xk,i')$ in the expression of B_i as a sum of $poi(xk,j)$, for B_i as defined in Definition 20.*

Let $f_{L_1}(x,y) \triangleq |x-y|$, be the characteristic function of L_1 distance, and let $r(x,y) \triangleq \frac{f_{L_1}(x,y)}{x+y}$. Define the interpolation function $I(x,y)$ such that $I(x,y)=0$ for $x+y \leq \frac{s}{4k}$, and $I(x,y)=1$ for $x+y \geq \frac{s}{3k}$, where I has partial derivatives bounded in magnitude by $O(\frac{k}{s})$. Note that such a function clearly exists.

We define the coefficients $z_{i,j} := I(\frac{i}{k},\frac{j}{k})\cdot r(\frac{i}{k},\frac{j}{k})$, for i,j satisfying $i+j \geq s/4$.

We now use the skinny bumps $B_{i,j}$ to represent $(1-I(x,y))\cdot r(x,y)$. Define $c_i = \frac{2s}{k}(1-\cos(i\pi/s))$, and let $\gamma_{i,j} = (1-I(c_i,c_j))\cdot r(c_i,c_j)$, be the "Chebyshev bump" coefficients. To conclude the construction, by Lemma 21 we can represent this linear combination of Chebyshev bumps as a linear combination of the Poisson bumps.

Theorem (Theorem 4)**.** *There is a constant c such that for any $\epsilon > \frac{c}{\sqrt{\log n}}$, the estimator described in Construction 38, when given $O(\frac{n}{\epsilon^2 \log n})$ independent samples from each of two distributions of support at most n, will compute an estimate of the L_1 distance between the pair of distributions, accurate to within ϵ, with probability of failure $o(1/poly(n))$.*

The proof of this theorem is nearly identical to that of Theorem 3; given $k = \frac{n}{\epsilon^2 \log n}$, we leverage the skinniness of the Chebyshev bumps, as guaranteed by Lemma 25, and use the fact that $\frac{f_{L_1}(x,y)}{x+y}$ is bounded by 1 everywhere, and has partial derivatives at (x, y) bounded in magnitude by $\frac{1}{x+y}$. Note that $s = \frac{1}{15} \log k$, which by the proof of Lemma 21 guarantees that the magnitude of the sum of the coefficients in the representation of $B_{i,j}$ as a sum of $poi(x, i') \cdot poi(y, j')$ is bounded by at most $(2(2+\sqrt{5})^s)^2 \leq k^{0.2}$, and thus we may apply Proposition 37 with $a = O(\epsilon), b = O(\epsilon/n)$, and $c = k^{-.1}$, yielding the theorem.

C.1 Lower Bounds for L_1 Distance

Theorem (Theorem 5)**.** *For any constants $0 < a < b < \frac{1}{2}$, distinguishing for a pair of distributions with support at most n whether their L_1 distance is less than a or greater than b requires $O(\frac{n}{\log n})$ samples.*

Proof. In [33, 32] we exhibited, for any sufficiently small ϕ, a pair of distributions, $p^n, p^{n/2}$ on support n such that are ϕ-close in the relative earthmover—and hence the L_1 sense—to uniform distributions, respectively, on n and $\frac{n}{2}$ elements, yet are indistinguishable in k samples. Construct such distributions for $\phi < \min\{a, \frac{1}{2} - b\}$. Consider now the task of distinguishing, for random permutations σ_1, σ_2, the *pair* of distributions $(\sigma_1(p^n), \sigma_2(p^n))$ from the pair $(\sigma_1(p^{n/2}), \sigma_2(p^n))$, where we consider that the application of a permutation relabels its elements.

Assume for the sake of contradiction that these pairs are distinguishable given k samples from each. We could thus construct a tester that distinguishes p^n from $p^{n/2}$ by, on being given k samples from one or the other, simulating the application of this hypothetical algorithm on these k samples, and k samples constructed ad hoc from a random permutation of p^n, and returning "p^n" if the L_1 distance is small, "$p^{n/2}$" otherwise.

Thus we have the desired contradiction and no such tester can exist. □

D Proof of Proposition 16

In this section we show that a solution y^+, y^- to the Lower Bound LP of Definition 15, for appropriate parameters, corresponds to a pair of distributions p^+, p^- of support n whose property values differ by roughly the objective value of the linear program and which are indistinguishable to property testers. We demonstrate indistinguishability by following the outline of the corresponding result in[33]. In fact, the Lower Bound LP can be seen as mechanizing the approach of [33]. Whereas in that paper there was a single explicit pair of distributions for which the goal was an indistinguishability result, the analysis there, while quite involved, essentially relied on nothing beyond the conditions of our Lower Bound LP. In the current paper, we thus outline how this analysis can be applied in this generality. We refer the reader to [33] for the full details and motivation for the approach. We omit several details here.

The essential approach is: 1) round y^+ and y^- to distributions p^+, p^- where the constraints of the linear program imply that p^+ and p^- will have almost identical expected fingerprints; 2) invoke Lemma 20 of [33] to see that the fingerprint distributions will thus also have almost identical covariances; 3) invoke the central limit theorem (Theorem 4) of [33] to conclude that the distributions of fingerprints are essentially multivariate Gaussian distributions of almost matching expectation and covariance, and hence indistinguishable. We briefly review each step in turn.

The bounds of the central limit theorem work well when the distributions have variance that is "large in every direction". (In other words, the covariance matrix has a large minimum eigenvalue.)

To ensure this, we add a "fattening" step to the rounding procedure of step 1–see Definition 17 of [33] and the related discussion. Essentially, all that is needed is to add a positive integer ϕ to each of the histogram entries $\frac{i}{k}$ for $i \in [k^{c_1}]$. Lemma 19 of [33], adapted to our context by replacing $\log k$ there with k^{c_1}, yields that if all these entries were 1, the minimum variance of the first k^{c_1} fingerprint entries would be at least $\frac{1}{6}k^{-9c_1/2}$; thus with ϕ entries the minimum variance is at least $\frac{\phi}{6 \cdot k^{9c_1/2}}$. The extra details of rounding histogram entries to integer values, ensuring that each distribution has at most n entries, and has total probability mass 1, are straightforward.

The second step, of leveraging Lemma 20 from [33], relies on the following trivial corollary of that result, implicit in that paper:

Corollary 39. *Given two distributions p^+, p^- such that when taking $Poi(k)$ samples from p^+, p^- respectively the expectations of the fingerprints match to within $k\epsilon$, element-by-element, for some $\epsilon > 0$, then the covariance matrices of the fingerprints match to within $O(k\sqrt{\epsilon}|\log \epsilon|)$, element-by-element.*

Finally, apply the central limit theorem, which states that the distribution of the first k^{c_1} fingerprints may be approximated by a Gaussian of corresponding mean and covariance, up to statistical distance $O(\frac{k^{4c_1/3}}{\sigma^{1/3}} \log n)$, where σ^2 is the minimum variance of the fingerprint distribution, in any direction, which we have bounded above as $\frac{\phi}{6 \cdot k^{9c_1/2}}$. Thus, for example, letting $\phi = \omega(k^{25c_1/2} \log^6 n)$ yields $o(1)$–closeness. Details are below. (We will actually use a somewhat larger ϕ in the proof of the following result so as to ensure further properties not mentioned in this introduction.)

Proposition (Proposition 16)**.** *Let π be a δ-relative earthmover continuous property that takes value 0 on the trivial distribution. Given any feasible point y^+, y^- to the* Lower Bound LP *of Definition 15 that has objective function value v, then, provided $k^{c_1} \in [\log^2 k, k^{1/32}]$ and $c_2 \geq \frac{1}{2} + 6c_1$, there exists a pair of distributions p^+, p^- of support at most n such that:*

- $\pi(p^+) - \pi(p^-) > v \cdot (1 - o(1)) - O(\delta \cdot k^{-c_1} \log k)$,
- no *algorithm on $Poi(k)$-samples can distinguish p^+ from p^- with probability* $1 - \Theta(1)$.

Proof. We prove the lemma for the case $\delta = 1$, as otherwise, we may divide the property by δ, and only the objective of the linear program will be affected, and thus both sides of the first claim of the proposition are proportional to δ, and nothing else is affected.

We note that 1-relative earthmover continuity implies that $|\frac{f_\pi(x)}{x}| \leq |\log x|$ for any x. Further, for the range under consideration, $x \in X = (0, \frac{k^{c_1}}{2k})$, this implies $|f_\pi(x)| \leq x|\log x| < \frac{k^{c_1}}{2k} \log k$.

For the case when $n < k^{1-2c_1}$, we thus have the LP constraint $\sum_{x \in X} \frac{y_x^+}{x} \leq n$ implies that the corresponding portion of the objective function is bounded as $\left|\sum_{x \in X} \frac{f_\pi(x)}{x} y_x^+\right| \leq n \frac{k^{c_1}}{2k} \log k \leq \frac{1}{2}k^{-c_1} \log k$, implying that the objective value of the LP is at most twice this, and thus that the proposition may be trivially satisfied by the pair consisting of any distribution and itself.

The other trivial case is when (for $n \geq k^{\log k}$) there exists some $x \geq \frac{1}{n}$ for which $|\frac{f_\pi(x)}{x}| \geq \log^2 k$. Let x^+ be the number in the interval $[\frac{1}{n}, \frac{1}{k^3}]$ that maximizes $\frac{f_\pi(x)}{x}$, and let x^- be the number that minimizes this. It is straightforward to see that relative earthmover continuity implies that, for the optimum (y_x^+, y_x^-) of the linear program, $\sum_{x \in X} \frac{f_\pi(x)}{x} y_x^+ \leq \frac{f_\pi(x^+)}{x^+} + 3 \log k$ and $\sum_{x \in X} \frac{f_\pi(x)}{x} y_x^- \geq \frac{f_\pi(x^-)}{x^-} - 3 \log k$, implying that $\frac{f_\pi(x^+)}{x^+} - \frac{f_\pi(x^-)}{x^-} \geq v \cdot (1 - o(1))$. Thus the uniform distributions on, respectively, $1/x^+$ and $1/x^-$ elements will have property values that differ by $v \cdot (1 - o(1))$, and further, will have indistinguishable fingerprint distributions (statistical distance $O(1/k)$ from each other), as in either case, no element will be seen more than once in $Poi(k)$ samples, except with $O(1/k)$ probability.

Otherwise, if neither of the above two cases apply, then we derive the distributions p^+, p^- directly from the linear program solution (y^+, y^-), via "fattening and rounding", applying Corollary 39 and then the central limit theorem from [33] to prove indistinguishability.

We first analyze what corresponds to "total probability mass" in each of y^+, y^-. Note that for any positive λ, $\sum_{i=0}^{\infty} poi(\lambda, i) = 1$. Consider combining the first two LP constraints into, for each $i \in \{0, \ldots k^{c_1}\}$, $\left|\sum_{x \in X}(y_x^+ - y_x^-) \cdot poi(xk, i)\right| \leq k^{-c_2}$, and then summing over $i < k^{c_1}$ to yield $\left|\sum_{x \in X}(y_x^+ - y_x^-)\left(\sum_{i<k^{c_1}} poi(xk, i)\right)\right| \leq k^{c_1}k^{-c_2}$. Since X consists only of elements less than $\frac{k^{c_1}}{2k}$, and by assumption, $k^{c_1} \geq \log^2 k$, Poisson tail inequalities yield that for any such x, we have $1 > \sum_{i<k^{c_1}} poi(xk, i) > 1 - o(\frac{1}{poly(k)})$. Thus $\sum_{x \in X} y_x^+$ and $\sum_{x \in X} y_x^-$ are at most $2k^{c_1}k^{-c_2} + o(\frac{1}{poly(k)})$ apart. Our first modification to y^+, y^- is to take whichever one has the higher sum and decrease its entries arbitrarily until the two sums are equal. Since $poi(xk, i) \leq 1$ in general, this will affect each constraint by at most $2k^{c_1}k^{-c_2} + o(\frac{1}{poly(k)})$, and will affect the objective function by at most $O(k^{c_1}k^{-c_2}\log^2 k)$. Next, multiply each of the entries in y^+, y^- by the largest number less than 1 that would make $\sum_{x \in X} y_x^+ \leq 1 - k^{-2c_1}$ and $\sum_{x \in X} \frac{y_x^+}{x} \leq n - k^{1-3c_1} - 1$, along with the corresponding statements for y^-. We note that the LP constraints imply this scaling is by $1 - o(1)$. Since before this scaling we had for each $i \leq k^{c_1}$ that $|\sum_x (y_x^+ - y_x^-) \cdot poi(xk, i)| \leq 3k^{c_1}k^{-c_2}$, after scaling both y^+, y^- by the same number less than 1, this will remain true. The final steps of the transformation are to round each of y^+, y^- into histograms h^+, h^- with integral entries, though which will not have total probability mass 1; fatten: for each $i \in [k^{c_1}]$ increment $h^+_{i/k}$ and $h^-_{i/k}$ by $\phi = k^{1-4c_1}$; to make each histogram have total probability mass 1, let m be the probability mass that must be added to each (which will be the same for each, by construction), and increment both h^+_m and h^-_m by 1. (There are some details involved in rounding appropriately, but the analysis is straightforward, and neither the objective value term nor the constraint terms corresponding to the difference in expected fingerprints will be affected by more than $o(k^{c_1}k^{-c_2})$.)

Thus h^+, h^- are now histograms of distributions, each having support at most n. Since $poi(xk, i) = \frac{i+1}{xk} \cdot poi(xk, i+1)$, we have, since h^+_x, h^-_x correspond to rounded versions of $\frac{y^+_x}{x}, \frac{y^-_x}{x}$, that the LP constraints for a certain i yield bounds on the $i+1$st fingerprint entries, specifically, the fact that $|\sum_x (y_x^+ - y_x^-) \cdot poi(xk, i)| \leq 3k^{c_1}k^{-c_2}$ implies that that the expected fingerprint entries up to k^{c_1} must match to within $3k^{1+c_1-c_2}$. Corollary 39 yields that the fingerprint covariances must thus match to within $O(k^{1-(c_2-c_1)/2}\log k)$. Further, since there are $\phi = k^{1-4c_1}$ elements in each distribution at each probability $\frac{i}{k}$ for $i < k^{c_1}$, Lemma 19 of [33] implies that the minimum covariance of either fingerprint distribution, in any direction, is at least $\Omega(k^{1-17c_1/2})$. Thus the central limit theorem yields that the statistical distance of each fingerprint distribution from the Gaussian of corresponding mean and covariance is $O(\frac{k^{4c_1/3}}{k^{(1-17c_1/2)/6}}\log n) < O(\frac{k^{3c_1}}{\sqrt[6]{k}}\log n)$. We note that, while we cannot bound n directly, we note that distribution h^+ is indistinguishable (statistical distance $O(\frac{1}{k})$ from a distribution obtained by modifying it so that no probabilities lie below $\frac{1}{k^3}$. Thus if we modify both h^+, h^- in this fashion before applying the central limit theorem, we have effectively made $n \leq k^3$, and thus, for $c_1 \leq \frac{1}{20}$ we have $O(\frac{k^{3c_1}}{\sqrt[6]{k}}\log k^3) = o(1)$.

We have thus shown that h^+, h^- are indistinguishable from Gaussians of corresponding mean and covariance. Comparing multivariate Gaussians is straightforward, and we appeal to Proposition 32 of[33] to see that two Gaussians are indistinguishable when the smallest covariance in any direction is $\omega(1)$ times larger than both the square of the distance between their means, and the product of the dimension (k^{c_1}) and the largest pairwise discrepancy between any entries of the two covariance matrices. The smallest covariance has been bounded by $\Omega(k^{1-17c_1/2})$; the element-wise difference between the means is at most $O(k^{1+c_1-c_2})$ implying that the square of their Euclidean distances is at most $O(k^{2+3c_1-2c_2})$. To ensure that the squared distance between the means is $o(1)$ times the smallest covariance, it is enough to let $c_2 \geq \frac{1}{2} + 6c_1$. Finally, the pairwise discrepancy between the

two covariance matrices was bounded above by $O(k^{1-(c_2-c_1)/2}\log k)$, which, plugging in our bound for c_2 yields $O(k^{3/4-5c_1/2}\log k$; the condition that this times the dimension (k^{c_1}) is $o(1)$ times the minimum covariance in any direction yields that, since by assumption $k^{c_1} \geq \log^2 k$, it is enough to set $c_1 \leq \frac{1}{32}$, yielding the desired indistinguishability. □

Many of the steps in the analysis here can be tightened considerably, as most of the tools we use were developed in a different paper for a different context[33]. However, for our present purposes of deriving asymptotic bounds without much concern for the constants involved, they serve well.(???promote this to somewhere in the body of the paper?)

E Proof of Proposition 17

Proposition (Proposition 17)**.** *Let π be a linear symmetric property such that for any histogram h, we have $\pi(h) \triangleq \sum_{x:h(x)\neq 0} h(x)x \cdot r(x)$, for some function $r : (0,1] \to \mathbb{R}$. Given integers k, n, and a set of coefficients $z_0, z_1, \ldots$ such that if we define the function $err : (0,1] \to \mathbb{R}$ by*

$$r(x) = err(x) + \sum_{i\geq 0} z_i poi(xk, i),$$

and if for positive real numbers a, b, c the following conditions hold:

1. *$|err(x)| < a + \frac{b}{x}$,*
2. *for all $j \geq 1$ let $\beta_j = \frac{j}{k} \cdot z_{j-1}$ with $\beta_0 = 0$, then for any j, ℓ such that $|j - \ell| \leq \sqrt{j}\log k$ we have $|\beta_j - \beta_\ell| \leq c\frac{\sqrt{j}}{\sqrt{k}}$*

Then the linear estimator given by coefficients $\beta_1, \ldots, \beta_k$, when given a fingerprint derived from a set of k independent samples chosen from a distribution of support at most n will estimate the property value with error at most $a + bn + c\log k$, with probability of failure $o(1/poly(k))$.

Proof of Proposition 17. To start, consider that instead of k samples, we are given $k' \leftarrow Poi(k)$ samples from the distribution. Trivially, if we prove the proposition in this setting, then, because $k' = k$ with probability at least $\frac{1}{O(\sqrt{k})}$, and our probability of failure is $o(1/poly(k))$, the conditional probability of failure given exactly k samples must also be $o(1/poly(k))$. Thus, for the remainder of the proof, assume we are given $k' \leftarrow Poi(k)$ samples.

The proof consists of two parts, we first argue that the first condition above guarantees that the expected value of the estimator is within $a + bn$ of the true property value—thus the resulting estimator has small *bias*. We then argue that the second conditions above implies, via tail bounds, that the value of the estimator will be very closely concentrated about its expectation.

For a histogram h corresponding to a distribution of support at most n, we have the following:

$$\begin{aligned} r(h) &= \sum_{x:h(x)\neq 0} h(x)x \cdot r(x) \\ &= \sum_{x:h(x)\neq 0} h(x)x \left(err(x) + \sum_{i\geq 0} z_i poi(xk, i) \right) \\ &= \left(\sum_{i\geq 0} \beta_{i+1} \frac{k}{i+1} \sum_{x:h(x)\neq 0} h(x)x \cdot poi(xk, i) \right) + \sum_{x:h(x)\neq 0} h(x) \cdot x \cdot err(x) \ . \end{aligned} \tag{1}$$

We start by bounding the magnitude of the second term (the error term). Since $\sum_x h(x) \leq n$, and $\sum_x h(x)x = 1$, we have

$$\sum_{x:h(x)\neq 0} h(x)x \cdot err(x) \quad \leq \quad \sum_{x:h(x)\neq 0} h(x)x \cdot a + \sum_{x:h(x)\neq 0} h(x)x \cdot \frac{b}{x} \leq a + nb.$$

We now turn to the first term in Equation 1. Observe that

$$x \cdot poi(xk, i) = x\frac{(xk)^i e^{-xk}}{i!} = \frac{(xk)^{i+1}e^{-xk}}{(i+1)!}\frac{i+1}{k} = poi(xk, i+1)\frac{i+1}{k}.$$

Additionally, $\sum_{x:h(x)\neq 0} h(x)poi(xk, j)$ is simply $E[\mathcal{F}_j]$, the expected jth fingerprint entry given $Poi(k)$ samples from h. Thus the first term in Equation 1 becomes:

$$\sum_{i\geq 0} \beta_{i+1}\frac{k}{i+1} \sum_{x:h(x)\neq 0} h(x)x \cdot poi(xk, i) \quad = \quad \sum_{i\geq 0} \beta_{i+1} \sum_{x:h(x)\neq 0} h(x)poi(xk, i+1) = \sum_{i\geq 1} \beta_i E[\mathcal{F}_i],$$

which is simply the expected value of our estimator. Thus the bias of the estimator is at most $a+bn$, as desired.

We now argue that with high probability the error will be tightly concentrated about this bias. We note that for $\lambda \geq 1$, the probability of a Poisson distribution $Poi(\lambda)$ taking a value outside the range $\lambda \pm \sqrt{\lambda}\log k$ decays super-polynomially fast with k. Thus letting $j = \lfloor\lambda\rfloor$, we thus also have that $Poi(\lambda)$ will lie outside $j \pm \sqrt{j}\log k$ with $o(1/poly(k))$ probability. Thus, with all but $o(1/poly(k))$ probability, each element in the support of the distribution such that $k \cdot p(i) \geq 1$ will be sampled a number of times that lies in the interval $j \pm \sqrt{j}\log k$, for $j = \lfloor k \cdot p(i)\rfloor$. Thus from Condition 2 of the proposition, each such element will contribute to the property estimate a number in an interval of radius $c\frac{\sqrt{j}}{\sqrt{k}} \leq c\frac{\sqrt{k\cdot p(i)}}{\sqrt{k}} = c\sqrt{p(i)}$ and hence diameter at most $2c\sqrt{p(i)}$. With a view towards applying Hoeffding's inequality, we bound the sum of the squares of the diameters of these intervals: $\sum_{i:p(i)\geq 1/k} 4c^2 \cdot p(i) \leq 4c^2$. Thus Hoeffding's inequality says that the contribution of the elements of probability at least $1/k$ to the estimate will be within $\sqrt{4c^2\frac{\log k}{4}} = \frac{c\log k}{2}$ of its expectation, except with $2 \cdot e^{-\frac{\log^2 k}{8}} = o(1/poly(k))$ probability.

Next we consider those elements for which $p(i) < \frac{1}{k}$. We note that for $\lambda < 1$ and $\ell \geq 1$ we have $poi(\lambda, \ell) = \frac{\lambda^\ell e^{-\lambda}}{\ell!} \leq \frac{\lambda}{\ell!}$. Thus the total probability that any element of probability less than $1/k$ appears more than $\log k$ times is at most $\left(\sum_{\ell>\log k} \frac{1}{\ell!}\right)\sum_i k \cdot p(i)$. The first term is $o(1/poly(k))$, and the second term equals k, leading to a total bound of $o(1/poly(k))$. Similar to above, we may use the bound from Condition 2 of the proposition, for $j = 1$ to say that, except with this negligible probability, each such element with $p(i) < \frac{1}{k}$ contributes to the property estimate a value in an interval of radius $\frac{c}{\sqrt{k}}$. We further bound the variance of each such contribution: since an element of probability $p(i) < \frac{1}{k}$ will likely be seen 0 times, and in fact will be seen a nonzero number of times only with probability less than $k \cdot p(i)$, the variance of each such contribution will be at most $k \cdot p(i) \cdot (2\frac{c}{\sqrt{k}})^2 = 4c^2 \cdot p(i)$, which must thus sum to at most $4c^2$. Thus we have a sum of independent random variables each in an interval of diameter $\frac{2c}{\sqrt{k}}$ and having total variance at most $4c^2$. Bennett's inequality says that in such a case, with a sum of independent random variables of total variance σ^2, each bounded to be within m of its mean, then the probability that the sum is more than t away from its mean is at most $2\exp(-\frac{\sigma^2}{m^2} \cdot \phi(\frac{mt}{\sigma^2}))$ where the function ϕ is defined as $\phi(x) = (1+x)\log(1+x) - x$. In our present case, we consider the probability that the contribution to the estimate from the small distribution elements deviates from its mean by more than $\frac{c\log k}{2}$, yielding

a bound of $2\exp(-k\cdot\phi(\frac{\log k}{4\sqrt{k}})$. Since for $x\le 1$, $\phi(x)>\frac{x^2}{3}$, our bound becomes $2\exp(-\frac{\log^2 k}{48})$, which is negligible.

Thus in either case, the probability of deviating from the expectation by more than $\frac{c\log k}{2}$ is negligible in k, so thus the total estimate will never deviate by more than $c\log k$ from its expectation, except with negligible probability. Thus the error of our estimator is at most $a+bn+c\log k$, with $o(1/poly(k))$ probability of failure.

□

F Matrix Exponentials of Poisson Matrices

Given a vector of Poisson coefficients, z, indexed from 0 through ∞, we may associate it with the real function $g:[0,\infty)\to\mathbb{R}$ defined as $g(x)=\sum_{i=0}^{\infty}z(i)\cdot poi(x,i)$. As we use it in this paper, the input of the function g is typically scaled by the number of samples, as in $g(xk)$. Consider the task we call "resampling", that is, given coefficients z and a constant α, finding a vector z_α that yields a corresponding g_α such that $g(\alpha x)=g_\alpha(x)$ for all $x\ge 0$. That is, if z is the vector of coefficients for a k-sample estimator, z_α will be a vector of coefficients for a $\frac{k}{\alpha}$ sample estimator that has *identical* expected estimates. (See Proposition 17.) Constructing such an estimator for $\alpha<1$ is straightforward—intuitively, taking more samples can never hurt. More specifically, given a Poisson process $Poi(\frac{x}{\alpha})$ that returns an integer ℓ, namely, "ℓ Poisson events have occurred", we may simulate a Poisson process $Poi(x)$ by, for each "event", accepting it with probability α and otherwise ignoring it; that is, when the Poisson process $Poi(\frac{x}{\alpha})$ returns ℓ, our simulation of $Poi(x)$ returns $i\le\ell$ with probability $\alpha^i(1-\alpha)^{\ell-i}\binom{\ell}{i}$, that is, the probability that a binomial distribution with parameter α returns i heads out of ℓ samples. Symbolically, $poi(x,i)=\sum_{\ell=i}^{\infty}poi(\frac{x}{\alpha},\ell)\alpha^i(1-\alpha)^{\ell-i}\binom{\ell}{i}$. To ensure $\sum_{i=0}^{\infty}z(i)\cdot poi(x,i)=\sum_{\ell=0}^{\infty}z_\alpha(\ell)\cdot poi(\frac{x}{\alpha},\ell)$ for all x, we expand and then change the order of summation:

$$\begin{aligned}\sum_{i=0}^{\infty}z(i)\cdot poi(x,i)&=\sum_{i=0}^{\infty}\sum_{\ell=i}^{\infty}z(i)poi(\frac{x}{\alpha},\ell)\alpha^i(1-\alpha)^{\ell-i}\binom{\ell}{i}\\&=\sum_{\ell=0}^{\infty}poi(\frac{x}{\alpha},\ell)\sum_{i=0}^{\ell}z(i)\alpha^i(1-\alpha)^{\ell-i}\binom{\ell}{i}\end{aligned}$$

which implies that we should set $z_\alpha(\ell)=\sum_{i=0}^{\ell}z(i)\alpha^i(1-\alpha)^{\ell-i}\binom{\ell}{i}$, as we do in the following construction.

Construction 40 (Resampling)**.** *Given a vector z, indexed from 0 through ∞, let z_α be the* resampled *version of z, defined as $z_\alpha(\ell)=\sum_{i=0}^{\ell}z(i)\alpha^i(1-\alpha)^{\ell-i}\binom{\ell}{i}$. We define $z_1\triangleq z$.*

Lemma 41. *Resampling a vector z by factor α to yield z_α satisfies $\sum_{i=0}^{\infty}z(i)\cdot poi(x,i)=\sum_{\ell=0}^{\infty}z_\alpha(\ell)\cdot poi(\frac{x}{\alpha},\ell)$ for all $x\ge 0$.*

To bound the size of the coefficients as α decreases, we prove the following general structural result, which is central to this section.

Proposition 42. *For arbitrary vector z of finite support and for $\alpha\in(0,1]$, let z_α be the α-resampled version of z, and let $||\cdot||_2$ denote the L_2 norm. Then $\sqrt{\alpha}||z_\alpha||_2$ is log-convex in α. Further, letting g denote the function represented by z, that is, $g(x)=\sum_{i=0}^{\infty}z(i)\cdot poi(x,i)$, then the limit as α approaches 0 of $\sqrt{\alpha}||z_\alpha||_2$ equals the L_2 norm of g.*

We first set up some preliminaries that will help us characterizes the behavior of $||z_\alpha||_2$.

Definition 43. *Define the matrix M_α for $\alpha \in (0,1)$ by $M_\alpha(i,j) = \sum_{\ell=0}^{\infty} \binom{\ell}{i}\binom{\ell}{j}\alpha^{i+j}(1-\alpha)^{2\ell-i-j}$, and the matrix A such that $A(i,i) = 1-2i, A(i,i+1) = A(i+1,i) = i+1$ for all $i \geq 0$ with zero entries otherwise, where both matrices are indexed by the nonnegative integers.*

We note that M_α is chosen so that, trivially, $||z_\alpha||_2^2 = zM_\alpha z^\intercal$. We relate M_α to the much simpler matrix A by the following lemma, in terms of *matrix exponentiation*.

Lemma 44. $M_\alpha = \frac{1}{\alpha}e^{(1-\alpha)A}$.

Proof. Note that $\frac{d}{d\alpha}\frac{1}{\alpha}e^{(1-\alpha)A} = -A\frac{1}{\alpha}e^{(1-\alpha)A} - \frac{1}{\alpha^2}e^{(1-\alpha)A}$, so we prove the result by showing that $\frac{d}{d\alpha}M_\alpha = -AM_\alpha - \frac{1}{\alpha}M_\alpha$, and noting that when $\alpha = 1$ we have that $\frac{1}{\alpha}e^{(1-\alpha)A}$ equals the identity matrix, which is easily seen to equal $\lim_{\alpha\to 1} M_\alpha$. We treat this as our initial condition.

We first evaluate $M_\alpha(i,j)$. Assume for the moment that $i \leq j$. Thus the sum that defines $M_\alpha(i,j)$ only has nonzero terms for $\ell \geq j$, so we may substitute $m = \ell - j$ and sum over m going from 0 to infinity instead. We aim to represent the terms using *rising factorial* notation, namely, for a number x, let $(x)_m$ denote $x(x+1)(x+2)\cdot\ldots\cdot(x+m-1)$. Further, aiming to use only an argument of m in the rising factorial notation for the mth component of the sum, we note that $\binom{\ell}{j} = \binom{m+j}{j} = \frac{(j+1)_m}{m!}$ and $\binom{\ell}{i} = \binom{m+j}{i} = \frac{(i+1)_{m+j-i}}{(m+j-i)!} = \frac{(j+1)_m}{(j-i+1)_m}\binom{j}{i}$. Thus $M_\alpha(i,j) = \alpha^{i+j}(1-\alpha)^{j-i}\binom{j}{i}\sum_{m=0}^{\infty}\frac{(j+1)_m(j+1)_m}{(j-i+1)_m m!}(1-\alpha)^{2m}$, where we may immediately read off the sum as the hypergeometric function ${}_2F_1(j+1,j+1,j-i+1;(1-\alpha)^2)$. Thus for $i \leq j$,

$$M_\alpha(i,j) = \alpha^{i+j}(1-\alpha)^{j-i}\binom{j}{i}{}_2F_1(j+1,j+1,j-i+1;(1-\alpha)^2).$$

We now turn to the claim, that $AM_\alpha + \frac{1}{\alpha}M_\alpha + \frac{d}{d\alpha}M_\alpha = 0$. Because of the structure of A, the (i,j)th entry of AM_α equals $iM_\alpha(i-1,j) + (1-2i)M_\alpha(i,j) + (i+1)M_\alpha(i+1,j)$. Further, to evaluate the derivative of M_α, we note that in general, we have the Gauss relation $\frac{d}{dt}{}_2F_1(x,y,z;t) = \frac{z-1}{t}\left({}_2F_1(x,y,z-1;t) - {}_2F_1(x,y,z;t)\right)$. Combining everything yields a linear combination of the hypergeometric functions ${}_2F_1(j+1,j+1,j-i;(1-\alpha)^2), {}_2F_1(j+1,j+1,j-i+1;(1-\alpha)^2)$, and ${}_2F_1(j+1,j+1,j-i+2;(1-\alpha)^2)$ which equals zero because of the corresponding Gauss relation between these three contiguous hypergeometric functions. (A slightly different linear combination arises for the border case where $i = j$, but again, the Gauss relations are sufficient.) □

We now prove our main proposition.

Proof of Proposition 42. Since by construction, $||z_\alpha||_2^2 = zM_\alpha z^\intercal$, and by Lemma 44 $M_\alpha = \frac{1}{\alpha}e^{(1-\alpha)A}$, we have that $(\sqrt{\alpha}||z_\alpha||_2)^2 = ze^{(1-\alpha)A}z^\intercal$. Substituting $1-\alpha \to \alpha$ yields that this is a log-convex function of α provided $ze^{\alpha A}z^\intercal$ is. Denote $f(\alpha) = ze^{\alpha A}z^\intercal$. We note that since the second derivative of the logarithm of a positive function f equals $\frac{f''\cdot f - f'^2}{f^2}$, we have that f is log-convex provided $f\cdot f'' \geq f'^2$. Since the vectors z are constant, we may differentiate $e^{\alpha A}$ and post- and pre-multiply by z. By definition, $\frac{d}{d\alpha}e^{\alpha A} = Ae^{\alpha A}$, and thus further $\frac{d^2}{d\alpha^2}e^{\alpha A} = A^2e^{\alpha A}$. We note that the power series representation $e^X \triangleq \sum_{i=0}^{\infty}\frac{X^i}{i!}$ implies, since A is symmetric, that A commutes with $e^{\alpha A}$. Since the square of $e^{\frac{1}{2}\alpha A}$ equals $e^{\alpha A}$, we may thus reexpress the first derivative of f as $ze^{\frac{1}{2}\alpha A}Ae^{\frac{1}{2}\alpha A}z^\intercal$, and the second derivative as $ze^{\frac{1}{2}\alpha A}A^2e^{\frac{1}{2}\alpha A}z^\intercal$. Letting $v_\alpha \triangleq ze^{\frac{1}{2}\alpha A}$, since all the matrices are symmetric, we thus have that $f(\alpha) = vv^\intercal$, $f'(\alpha) = vAv^\intercal$, and $f''(\alpha) = vA^2v^\intercal$, and the desired relation $f\cdot f'' \geq f'^2$ is the Cauchy-Schwarz inequality: $f'(\alpha)^2 = (vAv^\intercal)^2 \leq |vA|^2|v|^2 = (vAAv^\intercal)(vv^\intercal) = f''(\alpha)\cdot f(\alpha)$.

Finally, we show that for $g(x) = \sum_{i=0}^{\infty} z(i)\cdot poi(x,i)$, we have $\lim_{\alpha\to 0}\sqrt{\alpha}||z_\alpha||_2 = ||g||_2$. Note that $z_\alpha(\ell) = \sum_{i=0}^{\ell} z(i)Bin(\ell,\alpha,i)$, where $Bin(\ell,\alpha,i)$ denotes the probability that a binomial distribution with parameter α will draw i heads from ℓ trials. Recall that as α approaches 0, the binomial

distribution becomes very well approximated by the Poisson process of parameter $\alpha\ell$, yielding $z_\alpha(\ell) \approx \sum_i z(i) poi(\alpha\ell, i) = g(\alpha\ell)$. Thus $\lim_{\alpha\to 0} \alpha \cdot \sum_\ell z_\alpha(\ell)^2 = \int g(x)^2 dx = ||g||_2^2$, yielding the claim. □

We have thus shown that $\sqrt{\alpha}||z_\alpha||_2$ varies log-concavely with α; to complete the analysis of its behavior for $\alpha \in (0,1)$ we need to understand its behavior at the endpoints. The Linear Estimator LP provides us, in rough form, with bounds on both the size of the coefficients $z(i)$, and the size of the function the coefficients represent, $g(x) = \sum_{i=0}^\infty z(i) \cdot poi(x,i)$—that is, intuitively, bounds for the $\alpha = 1$ and $\alpha = 0$ cases respectively. However, we must eliminate one odd possibility before proceeding: for very small x, the linear program essentially bounds the linear combination of poisson functions as a multiple of $1/x$. The function $1/x$, however, has *infinite* L_2 norm, so a $1/x$ blowup would in fact be unworkable. Fortunately, this kind of blowup is in fact overly pessimistic: a linear combination of Poisson functions with bounded coefficients cannot "blowup" like $1/x$ at the origin; the following lemma characterizes this.

Lemma 45. *Given a vector z of coefficients that induces a function $g(x) = \sum_{i=0}^\infty z(i) \cdot poi(x,i)$, where for each i, $|z(i)|$ is at most some bound b, and $|g(x)| \le \frac{1}{x}$, then the L_2 norm of g is $O(\log b)$.*

Proof. We note that $\int_1^\infty g(x)^2 dx \le 1$, so we need only bound the blowup as x approaches 0. We reexpress $g(x)^2$ as a sum of "thin" Poisson functions, $g(x)^2 = \sum_{\ell=0}^\infty \omega(\ell) poi(2x,\ell)$ via $poi(x,i) \cdot poi(x,j) = poi(2x, i+j) 2^{-(i+j)} \binom{i+j}{i}$, and note that the new coefficients are bounded by b^2 since for any index ℓ, we have $\omega(\ell) = \sum_{i=0}^\ell 2^{-\ell}\binom{\ell}{i} z(i) z(\ell - i)$, and $\sum_{i=0}^\ell 2^{-\ell}\binom{\ell}{i} = 1$.

We may further alter $g(x)^2$ so that it is still expressible by Poisson functions as: $g(x)^2 e^{-2x} = \sum_{\ell=0}^\infty \omega(\ell) poi(2x,\ell) e^{-2x} = \sum_{\ell=0}^\infty \omega(\ell) 2^{-\ell} poi(4x,\ell)$. Since $|\omega(\ell)| \le b^2$, we may cut off this sum at $\ell = 2\log_2 b$ without altering its value by more than 1. Define $h(x) = \sum_{\ell=0}^{2\log_2 b} \omega(\ell) 2^{-\ell} poi(4x,\ell)$. We note that the integral of h differs from the integral of $g(x)^2 e^{-2x}$ by less than 1, since $\int_0^\infty poi(4x,\ell)dx = \frac{1}{4}$, and thus the integral of the ℓth term of the sum is bounded by $\frac{1}{4} b^2 2^{-\ell}$, so the terms beyond $2\log_2 b$ will contribute at most $\frac{1}{4}$ to the integral.

We express h as $e^{-4x} P(x)$ where P is some polynomial of degree $2\log_2 b$. We may thus approximate $\int_0^1 h(x)^2$ to within factor e^4 by $\int_0^1 P(x)dx$. Gauss-Legendre quadrature trivially implies that if a polynomial of degree d is bounded on the interval $[\frac{1}{d^2}, 1]$, then its integral over $[0,1]$ is bounded identically. Since by assumption, $|h(x)| \le e^{-2x}\frac{1}{x^2} + 1$, where the final 1 captures the error from truncating at $2\log_2 b$, setting $d = 2\log_2 b$ yields the desired result. □

Finally, we assemble the pieces to transform a solution to the linear program into a near-optimal estimator, using Proposition 17 for the final step. The following construction will yield a vector of "Poisson coefficients," in terms of a parameter α and a solution to the linear estimator LP, that will yield, under Proposition 17, a $\frac{k}{\alpha}$-sample estimator whose performance—when α converges to 1 suitably—will be good enough to yield Theorem 1.

Construction 46. *Given a solution $z = z^+ - z^-$ to the linear estimator LP for a property represented as f_π, letting $\epsilon = \frac{2\log k}{k^{c_1}}$, for parameter $\alpha \in (0,1)$, construct the* α-scaled estimator *as follows: Attenuate the coefficients, defining $\tilde{z}(i) \triangleq z(i) \cdot (1-\epsilon)^i$. Resample $\tilde{z}$ by α to yield $\tilde{z}_\alpha$, as in Construction 40. Finally, construct the Poisson coefficients $z_E(i) \triangleq \tilde{z}_\alpha(i) + (1 - e^{-\epsilon\alpha i}) f_\pi(\frac{(i+1)\alpha}{k}) \frac{k}{(i+1)\alpha}$ for $i \le k$.*

For the next proposition, it will simplify the analysis to scale the property π under consideration so that it is 1-relative earthmover continuous, and shift it so that it takes value 0 on the trivial distribution with support 1: $\pi(\text{"1"}) = 0$. Clearly such a transform will not affect the behavior of linear estimators that are correspondingly transformed.

Proposition 47. *Let $z = z^+ - z^-$ be a solution to the linear estimator LP that has objective value v for a property π that is 1-relative earthmover continuous and takes value 0 on the trivial distribution, where $k^{c_1} \in [\log^2 k, k^{1/4}]$ and $c_2 < 1$. Then Proposition 17 when applied to the results of Construction 46 for $\alpha \in (\frac{1}{2}, 1)$ will yield a $\frac{k}{\alpha}$-sample estimator with error $v \cdot (1 + o(1)) + O(k^{\alpha c_2 + (3/2-\alpha)c_1 - 1/2} \log^4 k + k^{-c_1/2} \log^2 k)$ and probability of failure $o(\frac{1}{poly(k)})$ provided $v \leq \log^2 k$; if $v > \log^2 k$ then the "estimator" that returns 0 always will have error at most $v \cdot (1 + o(1))$.*

Proof. Defining the linear combination of Poissons $g(x) \triangleq \sum_{i=0}^{k^{c_1}} poi(xk, i) z(i)$, we first note that if we attenuate the coefficients, as in the construction, letting $\tilde{z}(i) \triangleq z(i) \cdot (1-\epsilon)^i$ and consider the corresponding linear combination of Poissons, $\tilde{g}(x)$, then g and $\tilde{g}$ are related as $\tilde{g}(x) \triangleq \sum_{i=0}^{k^{c_1}} z(i)(1-\epsilon)^i \frac{e^{-xk}(xk)^i}{i!} = g(x \cdot (1-\epsilon))e^{-\epsilon k x}$. We then resample this vector by α to yield $\tilde{z}_\alpha$. Our first task is to bound the coefficients here. We do this using the log-convexity of the resampling operation, as shown by Proposition 42. Explicitly, Proposition 42 implies $\sqrt{\alpha}||\tilde{z}_\alpha||_2 \leq ||\tilde{z}||_2^\alpha \cdot ||\tilde{g}||_2^{1-\alpha}$. We must bound each term on the right hand side. For the first term, we note that because each term in the objective function of the linear program is non-negative, the objective value v thus bounds the portion of the objective function $k^{-c_2} \sum_{i=0}^{k^{c_1}} |z(i)|$. Thus the L_1 norm of z is at most $v \cdot k^{c_2}$, which hence also bounds the L_1 norm of the attenuated coefficients, $\tilde{z}$; further, the L_1 norm of a vector bounds its L_2 norm, so we have $||\tilde{z}||_2 \leq v \cdot k^{c_2}$.

Bounding the second term, $||\tilde{g}||_2$ takes a bit more work. Consider the characteristic function of the property, f_π. By assumption, $f(1) = 0$. Further, relative-earthmover continuity imposes the condition $|f_\pi(x)/x - f_\pi(y)/y| \leq |\log \frac{x}{y}|$; letting $y = 1$ yields $|f_\pi(x)|/x \leq |\log x|$. We note that for the range of x considered in the linear program, $x \in (0, \frac{k^{c_1}}{2k})$, we may crudely bound $|\log x| < \frac{k^{c_1}}{kx} \log k$. For each such x, the linear program bounds the positive error of the Poisson approximation by $z^a + \frac{z^{b+}}{x}$ and the negative error by $z^a + \frac{z^{b-}}{x}$, where the objective function penalizes large z^a, z^{b+}, z^{b-} via the term $2z^a + n \cdot (z^{b+} + z^{b-})$. We consider two cases. For $n < k^{1-c_1}$ we note that if we replace the triple (z^a, z^{b+}, z^{b-}) by $(0, z^{b+} + \frac{z^a}{n}, z^{b-} + \frac{z^a}{n})$ then the objective function remains unchanged, and further, each of the linear program constraints becomes looser, as, since $x < \frac{1}{n}$, we have $z^a + \frac{z^{b+}}{x} \leq 0 + \frac{z^{b+} + z^a/n}{x}$ with the corresponding statement for z^{b-}. Thus at optimum, we may assume $z^a = 0$. Since as noted above, $|\frac{f_\pi(x)}{x}| < \frac{k^{c_1}}{kx} \log k$, we have that letting $z^{b+} = z^{b-} = \frac{k^{c_1}}{k} \log k$ and all the other variables being 0 is a feasible point of the linear program with objective value $n(z^{b+} + z^{b-})$ and thus the since all variables of the linear program are restricted to be nonnegative, the sum $z^{b+} + z^{b-} = 2\frac{k^{c_1}}{k} \log k$ bounds both z^{b+} and z^{b-} at the optimum of the linear program. Thus at optimum, the bound in each constraint of the linear program may be bounded as $z^a + \frac{z^{b\pm}}{x} \leq 2\frac{k^{c_1}}{xk} \log k$. We analyze this in a moment.

For the other case, when $n \geq k^{1-c_1}$, we note that the bound in each constraint of the linear program may be bounded as $z^a + \frac{z^{b\pm}}{x} \leq 2z^a + \frac{z^{b+} + z^{b-}}{x} \frac{n}{k^{1-c_1}} \leq \frac{2z^a + n \cdot (z^{b+} + z^{b-})}{xk^{1-c_1}} \leq \frac{v}{xk^{1-c_1}}$. Thus for both cases we have the bound $z^a + \frac{z^{b\pm}}{x} \leq \frac{k^{c_1}}{xk} \max\{2\log k, v\}$. Adding this to the above bound $|\frac{f_\pi}{x}| \leq \frac{k_1^c}{xk} \log k$ yields a bound on the right hand sides of each constraint in the linear program, namely a bound on g, the left hand side of the linear program constraints, of $|g(x)| \leq \frac{k^{c_1}}{xk}(v + 3\log k)$ for $x \in (0, \frac{k^{c_1}}{2k})$. To bound $|g(x)|$ for $x \geq \frac{k^{c_1}}{2k}$ we note that g is a linear combination of Poissons with coefficients as high as $v \cdot k^{c_2}$, and may thus reach as high as $v \cdot k^{c_2}$. We note, however, that we are dealing with the *attenuated* version of g, namely, as derived above, $\tilde{g}(x) = g(x \cdot (1-\epsilon))e^{-\epsilon k x}$ where $\epsilon = \frac{2\log k}{k_1^c}$. Thus at $x = \frac{k^{c_1}}{2k}$ the attenuation is already $e^{-\log k} = \frac{1}{k}$, and will clearly decay at least as fast as $\frac{1}{x}$ beyond this. Thus, for all x, we have $\tilde{g}(x) \leq 2\frac{k^{c_1}}{xk}(v + 3\log k)$, where the 2 is a crude bound on $\frac{1}{1-\epsilon}$. Thus if we scale $\tilde{g}$ by $\frac{1}{2k^{c_1}(v+3\log k)}$ so that it is bounded by $\frac{1}{kx}$ and apply Lemma 45 to $\frac{\tilde{g}(xk)}{2k^{c_1}(v+3\log k)}$, we thus have a bound on the L_2 norm of $\tilde{g}$ of $||\tilde{g}||_2 = O(2k^{c_1}(v + 3\log k)\log(v \cdot k^{c_2})) =$

$O(k^{c_1}\log^3 k)$ for $v < \log^2 k$.

Thus, as discussed at the beginning of the proof, we may combine this bound and the bound $||\tilde{z}||_2 \leq v \cdot k^{c_2}$ via log-convexity to yield a bound on the L_2 norm of the resampled coefficients: $\sqrt{\alpha}||\tilde{z}_\alpha||_2 = O(k^{\alpha c_2+(1-\alpha)c_1} \cdot \log^3 k)$. We will consider cases where $\alpha \in (\frac{1}{2}, 1)$, so we may drop the $\sqrt{\alpha}$ term from the left hand side while preserving the asymptotic expression.

As each element of $\tilde{z}_\alpha$ must be at most the L_2 norm of the whole, we have the element-by element bound of $|\tilde{z}_\alpha(i)| = O(k^{\alpha c_2+(1-\alpha)c_1}\log^3 k)$. We are now in a position to analyze the application of Proposition 17 to the coefficients $z_E(i) = \tilde{z}_\alpha(i) + (1-e^{\alpha\epsilon i})f_\pi(\frac{(i+1)\alpha}{k})\frac{k}{(i+1)\alpha}$, where for $i > k$ we extend this definition by letting $z_E(i) = z_E(k)$.

We first analyze the second condition of Proposition 17, where we separately bound the contributions from $\tilde{z}_\alpha$ and from the remaining term. We have just derived the bound $|\tilde{z}_\alpha(i)| = O(k^{\alpha c_2+(1-\alpha)c_1}\log^3 k)$, and we use this for $i \leq 2k^{c_1}$. Our aim is to find a bound c such that for all j, ℓ between 1 and $2k^{c_1}$ such that $|j-\ell| \leq \sqrt{j}\log k$ we have $c > \frac{\sqrt{k}}{\sqrt{j}}|\frac{j}{k}\tilde{z}_\alpha(j-1) - \frac{\ell}{k}\tilde{z}_\alpha(\ell-1)|$. We note that $\frac{j}{\sqrt{j}} = O(k^{c_1/2})$, and that $\frac{\ell}{\sqrt{j}} \leq \frac{j+\sqrt{j}\log k}{\sqrt{j}} = O(k^{c_1/2})$, which implies that we may set c to be $O(k^{c_1/2})$ times our just-derived bound on $|\tilde{z}_\alpha(i)|$, namely, $c = O(k^{\alpha c_2+(3/2-\alpha)c_1-1/2}\log^3 k)$.

For the case where one of j, ℓ is greater than $2k^{c_1}$ we now derive a bound on how $\tilde{z}_\alpha(i)$ decays for large i. As each original coefficient $z(i)$ is bounded by $v \cdot k^{c_2}$, each attenuated coefficient is bounded as $|\tilde{z}(i)| \leq v \cdot k^{c_2}(1-\epsilon)^i$. Assume for the moment that each coefficient equals exactly this. The corresponding linear combination of Poissons is hence $\tilde{g}(x) = v \cdot k^{c_2}e^{-\epsilon k x}$; resampling by α factor replaces x with αx, which has the effect of replacing ϵ by $\alpha\epsilon$, yielding coefficients $v \cdot k^{c_2}(1-\alpha\epsilon)^i$. Since resampling involves a *positive* linear combination of the coefficients, we thus have the bound $|\tilde{z}_\alpha(i)| \leq v \cdot k^{c_2}(1-\alpha\epsilon)^i$. As $v < \log^2 k$ and $(1-\alpha\epsilon)^i < e^{-\alpha\epsilon i} = e^{-\frac{2\alpha i \log k}{k^{c_1}}}$, then for $\alpha \geq \frac{1}{2}$, $c_2 < 1$, and $i > k^{c_1}$ we have $|\tilde{z}_\alpha(i)| < \log^2 k$ and decaying by another factor of k for each addition of k^{c_1} to i. Thus, trivially, the c from above applies to this region.

We now examine the contribution to the second condition of Proposition 17 from the remaining term of z_E, namely $(1-e^{-\alpha\epsilon i})f_\pi(\frac{(i+1)\alpha}{k})\frac{k}{(i+1)\alpha}$. As above, we desire a bound $c' > \frac{\sqrt{k}}{\alpha\sqrt{j}}|(1-e^{-\alpha\epsilon(j-1)})f_\pi(\frac{j\alpha}{k}) - (1-e^{-\alpha\epsilon(\ell-1)})f_\pi(\frac{\ell\alpha}{k})|$ for pairs $j, \ell \geq 1$ such that $|j-\ell| \leq \sqrt{j}\log k$. For the case that $j \leq \sqrt{k}$, we use the bound $f_\pi(x) \leq x|\log x|$, the trivial bound $(1-e^y) < 1$ for any y, and the triangle inequality to yield a bound of $c' = O(k^{-1/4}\log k)$. For $j > \sqrt{k}$, we note that $e^{-\alpha\epsilon(j-1)}$ and $e^{-\alpha\epsilon(\ell-1)}$ are both negligible in k, and thus it is sufficient to bound $\frac{\sqrt{k}}{\alpha\sqrt{j}}|f_\pi(\frac{j\alpha}{k}) - f_\pi(\frac{\ell\alpha}{k})|$. To bound this change in f_π, recall that for general x, y we have $|f_\pi(x)/x - f_\pi(y)/y| \leq |\log\frac{x}{y}|$, yielding $|f_\pi(\frac{j\alpha}{k}) - \frac{j}{\ell}f_\pi(\frac{\ell\alpha}{k})| = O(\frac{j\alpha}{k}|\log\frac{\ell}{j}| = O(\frac{\sqrt{j}\log k}{k})$. We add this to the bound $|\frac{\ell-j}{\ell}f_\pi(\frac{\ell\alpha}{k})| = O(\frac{\sqrt{j}\log^2 k}{k})$. Combining, yields a bound of $c' = \frac{\sqrt{k}}{\alpha\sqrt{j}}O(\frac{\sqrt{j}\log^2 k}{k}) = O(\frac{\log^2 k}{\sqrt{k}})$. We note that, since $\alpha \in (\frac{1}{2}, 1)$ and $c_2 > \frac{1}{2}$, the bound derived earlier of $c = O(k^{\alpha c_2+(3/2-\alpha)c_1-1/2}\log^3 k)$ is at least $O(k^{-1/4}\log^3 k)$, which thus subsumes the two just derived bounds of respectively $O(k^{-1/4}\log k)$ and $O(\frac{\log^2 k}{\sqrt{k}})$. Thus we take $c = O(k^{\alpha c_2+(3/2-\alpha)c_1-1/2}\log^3 k)$ for the bound on the second condition of Proposition 17.

We now turn to the first condition of Proposition 17, essentially examining the bias of the estimator. We must compare $\frac{f_\pi(x)}{x}$ to the linear combination of Poissons $\sum_{i\geq 0} z_E(i) \cdot poi(\frac{xk}{\alpha}, i)$. We consider each of the two terms of z_E separately, and start by comparing the fraction of our target $(1-e^{-\epsilon k x})\frac{f_\pi(x)}{x}$ to the combination of Poissons corresponding to the second term of z_E, namely

$\sum_{i\geq 0}(1-e^{-\alpha\epsilon i})f_\pi(\frac{(i+1)\alpha}{k})\frac{k}{(i+1)\alpha}\cdot poi(\frac{xk}{\alpha},i)$. Since $\sum_{i\geq 0} poi(\frac{xk}{\alpha},i)=1$, we may thus bound

$$\sum_{i\geq 0}\left|(1-e^{-\alpha\epsilon i})f_\pi\left(\frac{(i+1)\alpha}{k}\right)\frac{k}{(i+1)\alpha}-(1-e^{-\epsilon kx})\frac{f_\pi(x)}{x}\right|\cdot poi(\frac{xk}{\alpha},i)$$

$$\leq\left(\sum_{i\geq 0}\left|e^{-\alpha\epsilon i}-e^{-\epsilon kx}\right|\frac{f_\pi(x)}{x}\cdot poi(\frac{xk}{\alpha},i)\right)+\left(\sum_{i\geq 0}(1-e^{-\alpha\epsilon i})poi(\frac{xk}{\alpha},i)\left|f_\pi\left(\frac{(i+1)\alpha}{k}\right)\frac{k}{(i+1)\alpha}-\frac{f_\pi(x)}{x}\right|\right)$$

We bound each of the sums separately, noting throughout that $\alpha\in(\frac{1}{2},1)$. Recalling that $\epsilon=\frac{2\log k}{k^{c_1}}$, we bound the first sum for $x\leq\frac{1}{\epsilon k}$ by noting that since e^{-y} has derivative at most 1 for positive inputs, we have $|e^{-\alpha\epsilon i}-e^{-\epsilon kx}|\leq\alpha\epsilon|i-\frac{kx}{\alpha}|$. Since $\left|\frac{f_\pi(x)}{x}\right|\leq|\log x|$, the first sum is thus bounded by $\alpha\epsilon|\log x|$ times the expected distance of $Poi(\frac{xk}{\alpha})$ from its mean, which is bounded by the square root of its variance, namely $\sqrt{\frac{kx}{\alpha}}$, yielding a bound on the first sum of $O(\epsilon\sqrt{kx}|\log x|)$. We apply this bound for $x\leq\frac{1}{\epsilon k}$; since $|x\log x|$ is an increasing function of x for $x<e^{-1}$, we evaluate this bound by plugging in $x=\frac{1}{\epsilon k}$ to yield $O(k^{-c_1/2}\log^{3/2}k)$. For $x>\frac{1}{\epsilon k}$, we note that $poi(\frac{xk}{\alpha},i)$ is negligible unless i is within a factor of 2 of $\frac{xk}{\alpha}$. Thus for $\epsilon i\geq\frac{\epsilon kx}{2a}$ we bound $|e^{-\alpha\epsilon i}-e^{-\epsilon kx}|\leq\alpha\epsilon|i-\frac{kx}{\alpha}|e^{-\epsilon kx/2}$, and thus, corresponding to the above bound on the first sum, we now have a bound of $O(\epsilon\sqrt{kx}|\log x|e^{-\epsilon kx/2})$. Because of the exponential term, this expression is maximized for $x=O(\frac{1}{\epsilon k})$, and as above we may bound the first sum as $O(k^{-c_1/2}\log^{3/2}k)$.

For the second sum, consider $x>\frac{1}{k\sqrt{\epsilon}}$. We note that $|\frac{f_\pi(y)}{y}-\frac{f_\pi(x)}{x}|\leq|\log\frac{x}{y}|$, which, when y is within a factor of two of x is bounded as $2\frac{|y-x|}{x}$. Since with all but negligible probability, when i is drawn from $Poi(\frac{xk}{\alpha})$ we will have $\frac{(i+1)\alpha}{k}$ within a factor of 2 of x, we have a bound for this case of $\left|f_\pi\left(\frac{(i+1)\alpha}{k}\right)\frac{k}{(i+1)\alpha}-\frac{f_\pi(x)}{x}\right|\leq 2\frac{|(i+1)\alpha/k-x|}{x}=2\frac{|(i+1)-xk/\alpha|}{xk/\alpha}$. Further, $1-e^{-\alpha\epsilon i}\leq\alpha\epsilon i\leq\frac{2xk}{\alpha}$, and is also at most 1. Thus we can bound the second term by $O(\frac{\min\{1,\epsilon kx\}}{xk}$ times the expected distance of $Poi(\frac{xk}{\alpha})$ from its mean; this latter quantity is bounded by $O(\sqrt{xk})$, yielding a bound on the second sum of $O(\frac{\min\{1,\epsilon kx\}}{\sqrt{xk}})$. The expression inside the asymptotic notation is maximized when $x=\frac{1}{\epsilon k}$, yielding a bound on the second sum of $O(\sqrt{\epsilon})=O(k^{-c_1/2}\log^{1/2}k)$ for $x>\frac{1}{k\sqrt{\epsilon}}$. Otherwise, for $x\leq\frac{1}{k\sqrt{\epsilon}}$ we analyze the second sum in two parts, noting that, since $i\geq 0$, we have $\frac{(i+1)\alpha}{k}\geq\frac{\alpha}{k}$, yielding that $\left|f_\pi\left(\frac{(i+1)\alpha}{k}\right)\frac{k}{(i+1)\alpha}\right|\leq|\log\frac{\alpha}{k}|<1+\log k$. Since $(1-e^{-\alpha\epsilon i})\leq\alpha\epsilon i$, we have $\sum_{i\geq 0}(1-e^{-\alpha\epsilon i})\left|f_\pi\left(\frac{(i+1)\alpha}{k}\right)\frac{k}{(i+1)\alpha}\right|\cdot poi(\frac{xk}{\alpha},i)\leq(1+\log k)\alpha\epsilon\cdot\mathrm{E}[Poi(\frac{xk}{\alpha})]=\epsilon xk(1+\log k)$. For $x<\frac{1}{k\sqrt{\epsilon}}$ this is $O(\sqrt{\epsilon}\log k)=O(k^{-c_1/2}\log^{3/2}k)$. The remaining part of the second sum we easily bound as $\sum_{i\geq 0}(1-e^{-\alpha\epsilon i})\left|\frac{f_\pi(x)}{x}\right|\cdot poi(\frac{xk}{\alpha},i)\leq\alpha\epsilon\left|\frac{f_\pi(x)}{x}\right|\sum_{i\geq 0}i\cdot poi(\frac{xk}{\alpha},i)=\epsilon xk\left|\frac{f_\pi(x)}{x}\right|\leq\epsilon xk|\log x|$. This last expression is increasing in x, and hence we have a bound for $x\leq\frac{1}{k\sqrt{\epsilon}}$ of $O(\sqrt{\epsilon}\log k)=O(k^{-c_1/2}\log^{3/2}k)$. Thus we have shown that the portion of z_E other than $\tilde{z}_\alpha$ contributes to the linear combination of Poissons a function that is within $O(k^{-c_1/2}\log^{3/2}k)$ of $(1-e^{-\epsilon kx})\frac{f_\pi(x)}{x}$.

It remains to compare the remaining portion of z_E with the remaining fraction of $\frac{f_\pi(x)}{x}$, namely, compare $\sum_{i\geq 0}\tilde{z}_\alpha(i)\cdot poi(\frac{xk}{\alpha},i)$ to $e^{-\epsilon kx}\frac{f_\pi(x)}{x}$. We start the analysis by considering the vector z returned by the linear program, which, for positive numbers a,b satisfies $\left|\frac{f_\pi(x)}{x}-\sum_{i\geq 0}z(i)\cdot poi(xk,i)\right|\leq a+\frac{b}{x}$, for $x\in[0,\frac{k^{c_1}}{2k}]$, where the objective value of the linear program, v, is guaranteed by the linear program to be at least as large as $a+bn$.

As argued above, attenuating z to form $\tilde{z}(i) \triangleq z(i) \cdot (1-\epsilon)^i$ transforms the linear combination of Poissons $g(x) \triangleq \sum_{i=0}^{k^{c_1}} poi(xk, i)z(i)$ into $\tilde{g}(x) = g(x \cdot (1-\epsilon))e^{-\epsilon kx}$. Thus $\tilde{g}(x)$ is within $a + \frac{b}{x(1-\epsilon)}$ of $e^{-\epsilon kx}\frac{f_\pi(x(1-\epsilon))}{x(1-\epsilon)}$, where $\frac{f_\pi(x(1-\epsilon))}{x(1-\epsilon)}$ is within $|\log(1-\epsilon)|$ of $\frac{f_\pi(x)}{x}$. By the triangle inequality, $\tilde{g}(x)$ is thus within $a + O(\epsilon) + \frac{b\cdot(1+O(\epsilon))}{x}$ of $e^{-\epsilon kx}\frac{f_\pi(x)}{x}$, provided $x(1-\epsilon) \in [0, \frac{k^{c_1}}{2k}]$. Otherwise, we have $x > \frac{k^{c_1}}{2k}$, implying $e^{-\epsilon kx} \leq e^{-\log k} = \frac{1}{k}$, which is small enough to wipe out any discrepancy that may occur in this region. Specifically: since the Poisson coefficients sum to at most $k^{c_2}v \leq k\log^2 k$, and since any Poisson distribution of parameter λ has each probability bounded by $O(\frac{1}{\sqrt{\lambda}})$, we have that for $x > \frac{k^{c_1}}{2k}$, the linear combination of Poissons $g(x)$ must be at most $O(k^{1-c_1/2}\log^2 k)$, implying $\tilde{g}(x) = O(k^{-c_1/2}\log^2 k)$ in this range. Trivially, $e^{-\epsilon kx}\frac{f_\pi(x)}{x} = O(\frac{\log k}{x})$. Thus for arbitrary positive x we have that $\tilde{g}(x)$ is within $a + O(k^{-c_1/2}\log^2 k) + \frac{b\cdot(1+O(k^{-c_1}\log k))}{x}$ of $e^{-\epsilon kx}\frac{f_\pi(x)}{x}$. Resampling $\tilde{z}$ to $\tilde{z}_\alpha$ is exact, with $\tilde{g}(x) = \sum_{i\geq 0}\tilde{z}_\alpha(i)\cdot poi(\frac{xk}{\alpha}, i)$, so thus these bounds apply to $\sum_{i\geq 0}\tilde{z}_\alpha(i)\cdot poi(\frac{xk}{\alpha}, i)$ as well, as desired.

We thus invoke Proposition 17. For the first condition, we have shown that $\sum_{i\geq 0} z_E(i)\cdot poi(\frac{xk}{\alpha}, i)$ approximates $\frac{f_\pi(x)}{x}$ to within $a + O(k^{-c_1/2}\log^2 k) + \frac{b\cdot(1+O(k^{-c_1}\log k))}{x}$, where $a + bn \leq v$. We have shown that the second condition applies for $c = O(k^{\alpha c_2 + (3/2-\alpha)c_1 - 1/2}\log^3 k)$. Thus Proposition 17 yields that: the linear estimator z_E estimates the property π to within error $v\cdot(1+o(1)) + O(k^{\alpha c_2+(3/2-\alpha)c_1-1/2}\log^4 k + k^{-c_1/2}\log^2 k)$ using $\frac{k}{\alpha}$ samples, with probability of failure $o(\frac{1}{poly(k)})$, provided $v \leq \log^2 k$.

The proof will be complete upon analyzing the unusual but essentially trivial case of $v > \log^2 k$. Note that any distribution of support at most n must have relative earthmover distance from the trivial distribution (support on 1 element) at most $\log n$, and thus property value between $\pm\log n$. Thus if $n < v\cdot k^3$ then the "estimator" that always returns 0 will always have error at most $\log v + 3\log k = v\cdot(1+o(1))$. We consider the case when $n \geq v\cdot k^3$. Let π^+, π^- denote respectively the maximum and minimum value of $\frac{f_\pi(x)}{x}$ for $x \in [\frac{k}{n}, \frac{1}{vk^2}]$, with x^+, x^- denoting respectively the values at which π^+, π^- are attained. For this range of x, the Poisson functions take very limited values: $poi(xk, 0) = e^{-xk} \in [1-\frac{1}{vk}, 1]$, and thus the remaining Poissons sum up to at most $\frac{1}{vk}$. Thus since the coefficients of the vector z are at most $v\cdot k^{c_2} \leq vk$, we may use the triangle inequality to bound the difference between the expected estimates returned in the "$+$" and "$-$" case: $\sum_{i\geq 0} z(i)\cdot|poi(x^+k, i) - poi(x^+k, i)| \leq 4$. Letting e^+ be the expected estimate returned in the "$+$" case, we consider the constraints corresponding to x^+ and x^- from the linear program: $|\pi^+ - e^+| \leq z^a + \frac{\max\{z^{b+}, z^{b-}\}}{x^+}$ and $|\pi^- - e^+| \leq 4 + z^a + \frac{\max\{z^{b+}, z^{b-}\}}{x^-}$. Since $v = 2z^a + n\cdot(z^{b+} + z^{b-})$, we note that $x \geq \frac{k}{n}$ implies $\frac{z^{b+}+z^{b-}}{x^\pm} \leq \frac{v}{k}$, and we have, letting $\pi^\pm$ denote either π^+ or π^- that $|\pi^\pm - e^+| \leq \frac{v}{2} + \frac{v}{k} + 4$. Thus by the triangle inequality we have $|\pi^+ - \pi^-| \leq v + \frac{2v}{k} + 8$. Consider the relative earthmover cost of taking an arbitrary distribution of support at most n, and making all its probabilities lie in the interval $[\frac{k}{n}, \frac{1}{vk^2}]$. We note that trivially, this is at most $\max\{\log k, \log vk^2\} = \log v + 2\log k$. Thus the interval encompassing *all* possible values π might take has diameter at most $v + \frac{2v}{k} + 8 + 2(\log v + 2\log k)$ and contains 0. Hence the "estimator" that always returns 0, without looking at any samples, will be accurate to within $v\cdot(1+o(1))$ for $v = \omega(\log k)$, as desired. □

Theorem (Theorem 1)**.** *Let π be a symmetric linear property that is $\delta(k)$-relative earthmover continuous on distributions of support $n(k)$. If for some constant $c > 0$ and parameter $\epsilon(k) = \delta/k^{o(1)}$, any distributions of support n whose π values differ by at least ϵ are distinguishable with probability at least $\frac{1}{2} + c$ in k samples, then for each k there exists a linear estimator that estimates π on distributions of support n to within error $(1+o(1))\epsilon$ using $(1+o(1))k$ samples, and which has probability of failure $o(\frac{1}{poly(k)})$.*

Proof of Theorem 1. Without loss of generality, we assume $\delta = 1$, as we may replace π, ϵ, δ by $\frac{\pi}{\delta}, \frac{\epsilon}{\delta}, 1$ respectively, and scaling the property by $\frac{1}{\delta}$ simply scales the estimation error correspondingly. Further, without loss of generality, we assume that the property has value 0 on the trivial distribution of support 1, as the property estimation problem is unaffected by constant shifts.

Let c_1, as a function of k, be such that it converges to 0 as k increases, yet large enough that $k^{-c_1/2}\log^2 k = o(\min\{\epsilon, 1\})$. Let $c_2 = \frac{1}{2} + 6c_1$. Consider k large enough so that $c_1 \leq \frac{1}{32}$. Proposition 16 implies that, for these parameters, any solution to the Lower Bound LP with objective value v induces a pair of indistinguishable distributions whose property values differ by at least $v \cdot (1 - o(1)) - O(k^{-c_1}\log k)$, which must thus be smaller than than ϵ, as defined by the theorem. Thus $v \leq \epsilon \cdot (1 + o(1))$.

We then apply Proposition 47 to conclude that, for any $\alpha \in (\frac{1}{2}, 1)$ there exists a $\frac{k}{\alpha}$-sample estimator that has $o(\frac{1}{poly(k)})$ probability of failure, and error at most $v\cdot(1+o(1))+O(k^{\alpha c_2+(3/2-\alpha)c_1-1/2}\log^4 k+ k^{-c_1/2}\log^2 k)$. As already noted, $v \leq \epsilon \cdot (1 + o(1))$, and by assumption, $k^{-c_1/2}\log^2 k = o(\epsilon)$. For the remaining (middle) term, we note that since $c_2 = \frac{1}{2} + 6c_1$ we have $\alpha c_2 + (3/2 - \alpha)c_1 - 1/2 \leq \frac{1}{2}(\alpha - 1) + \frac{13}{2}c_1$. Setting $\alpha = 1 - 15c_1$ yields that this expression is at most $-c_1$, yielding that $k^{\alpha c_2+(3/2-\alpha)c_1-1/2}\log^4 k \leq k^{-c_1}\log^4 k$. By assumption, this is $o(\min\{\epsilon, 1\}^2) = o(\epsilon)$. Thus, the estimator guaranteed by Proposition 47 has total error at most $\epsilon\cdot(1+o(1))$, as desired. Since $\alpha = 1-o(1)$, the estimator uses at most $k \cdot (1 + o(1))$ samples. □