The Power of Linear Estimators

Gregory Valiant
UC Berkeley
gregory.valiant@gmail.com

Paul Valiant
UC Berkeley
pvaliant@gmail.com

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 [26], [27].

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 additive accuracy ϵ, with probability of failure $o(1/poly(n))$. From the recent lower bounds given in [26], [27], 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 [26], [28], 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 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 additive accuracy ϵ using $O(\frac{n}{\epsilon^2 \log n})$ samples from each distribution, which is constant-factor optimal, for constant ϵ. This is the first sublinear-sample estimator for this fundamental property.

Keywords-Property Testing, Entropy Estimation, L1 Estimation, Duality

Gregory Valiant's work is supported by a National Science Foundation Graduate Research Fellowship.

Paul Valiant's work is supported by a National Science Foundation Mathematical Sciences Postdoctoral Research Fellowship.

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 parametrization 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 [27], [28]. 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 with high probability 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 $1/2$.

1.1. 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 [27] 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" $poi(x, i) \triangleq \frac{e^{-x}x^i}{i!}$; see [11], [20], [21], [22], [25], [30]. Indeed, in [21], [22], Paninski showed the existence of a sublinear-sample linear estimator for entropy via a simple nonconstructive proof that applies the Stone-Weierstrass theorem to approximate the logarithm function (the characteristic function of entropy) via 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 in retrospect 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 traverses 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

[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.

estimating entropy in [27], 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 [28], 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, in the full version of the paper we give an explicit constant-factor optimal linear estimator for estimating the distance to uniformity.

Our entire framework extends 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. In the full version of this paper we give an explicit linear estimator for L_1 distance, and show it is constant factor optimal for any constant accuracy ϵ by giving a simple construction that leverages the lower bounds of [27].

2. Related Work

Linear programming duality is a beloved tool for showing the optimality of algorithms. Perhaps the clearest 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. 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 problems of estimating the support size (see [10] for several hundred references), and estimating the entropy have, perhaps, received the most attention, both in the setting of multiplicative approximations, and additive approximations.

Tight multiplicative bounds of $\Omega(n/\alpha^2)$ for approximating the support size to a multiplicative factor of α (where elements of the distribution are restricted to have probability mass at least $1/n$) are given in [3], [13] 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 [24], showing that for any constant $\epsilon \in (0, \frac{1}{2})$, any estimator that obtains additive error at most ϵn with probability at least $2/3$ requires at least $n/2^{\Theta(\sqrt{\log n \cdot \log\log n})}$ samples. Recent work [26], [27] improves this to a tight bound of $\Omega(\frac{n}{\log n})$.

For entropy estimation, Batu *et al.* [4], [5], [6], Brautbar *et al.* [9], Guha *et al.* [16], and Valiant [29] considered the problem of *multiplicative* approximation. For the problem of additively estimating entropy, recent work [26], [28] gives an estimator that uses $O(\frac{n}{\epsilon^2 \log n})$ samples, and returns an ϵ accurate estimate. The recent lower bounds in [26], [27] 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. [7], [8], [15]

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 [1], [2], [12], [17], [18], [19].

2.2. Linear Estimators for Entropy

There has been a long line of research proposing and analyzing linear estimators for entropy. 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 superfluous 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 [4], [8], 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 [21]:

- **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 [20]:** 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 [23].
- **The jackknifed naive estimator [14]:** $H^{JK}(\mathcal{F}) \triangleq k \cdot H^{naive}(\mathcal{F}) - \frac{k-1}{k} \sum_{j=1}^{k} H^{naive}(\mathcal{F}^{-j})$, 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. The estimator we construct in Section 6, 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 [21], [22], 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 [21], even dating back to [20] 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 Section 6). Such a representation has been attempted in the past, either explicitly or implicitly in [11], [20], [25], [30], though these works were unable to succeed in producing an accurate approximation of the logarithm function in the small-probability regime.

2.3. Comparison with [26], [27], [28]

This work relies heavily on techniques developed recently in [26], [27], [28], which, for the problems of estimating support size and entropy, provided the first upper bounds and the first lower bounds that are constant factor optimal, for constant additive error. The main result of this current paper comes from, in a sense, mechanizing the lower bound approach of [26] (a full version of which may be found in [27]) and realizing that, once suitably abstracted, the search for the best lower bound in this framework is *dual* to the search for the best linear estimator in the classic framework mentioned above, of trying to approximate the characteristic function of a linear property as a linear combination of Poisson functions (see Definition 6, and Examples 7, 8 and 9 for examples of characteristic functions of several properties).

The upper bound techniques introduced in [26] on the surface are very different from those of the current paper. The estimators of [26] use the data samples to construct a linear program, whose solution yields the property estimate. While this current paper uses linear programming to construct estimators, each of the estimators is linear, and thus involves the computation of a single dot product.

The explicit linear estimators we construct—for entropy, and in the full version, distance to uniformity, and L_1 distance—rely on a Chebyshev polynomial construction that was first used *nonconstructively* in [26] as a proof technique to demonstrate the performance of the linear programming estimators constructed there. While the approach of [26] can also yield sublinear-sample (non-linear) estimators for distance to uniformity, and, perhaps with some additional work, L_1 distance, the tighter correspondence between the estimators of this paper and the lower bound constructions can yield better estimators: the error of our linear estimator for entropy decreases inverse linearly with the number of samples, whereas the estimator of [26] has an inverse square-root relationship. (See [28] for the details and proof of correctness of the construction of the estimators of [26].)

3. Definitions

We state some definitions that will be used throughout.

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* $\pi : \mathcal{D}^n \to \mathbb{R}$. *A property is* symmetric *if, for all distributions* D*, and all permutations of the support,* σ, $\pi(D) = \pi(D \circ \sigma)$, *where* $D \circ \sigma$ *denotes the distribution obtained from* D *by permuting the support according to* σ.

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*,*

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

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* $|\frac{f_\pi(x)}{x} - \frac{f_\pi(y)}{y}| \leq |\log(x/y)|$.

3.1. Poisson Samples

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 invaluable when arguing about the structure of the distribution of such fingerprints. Since $k' \leftarrow Poi(k)$ is closely concentrated around k, we may often easily replace k-sample testing with $Poi(k)$-sample testing and benefit from this independence.

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, the variances of fingerprint entries are also easy to work with, and for example are clearly seen to sum to at most k.

4. Summary of Results

Our main theorem shows that linear estimators are near-optimal for additively estimating 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 $o(\frac{1}{poly(k)})$. Additionally, such a linear estimator is given as the solution to a polynomial-sized linear program.

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

Corollary. *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, we leverage the insights provided by key components of the proof of Theorem 1 to give explicit constructions of near-optimal linear estimators for entropy (Section 6), distance to uniformity (see full version), and L_1 distance between pairs of distributions (see full version). 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 17, 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 [26], [27], both in terms of the dependence on n and the dependence on ϵ. In particular, this resolves the main open question posed in [26], [28] as to whether the sample complexity increases linearly versus quadratically with the inverse of the desired accuracy, $1/\epsilon$.

Theorem 3. *For any $\epsilon > \frac{1}{4 \log m}$, there is an explicit linear estimator that, 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 linear estimator for distance to uniformity, and we note that the lower bounds shown in [26], [27] 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 [7], [15] for the related problem of distinguishing a uniform distribution on m samples from one that has constant distance from such a distribution.

Theorem 4. *There is an explicit linear estimator for L_1 distance and a constant c such that for any $\epsilon > \frac{c}{\sqrt{\log n}}$, the estimator, 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 $o(1/poly(n))$. Further, this number of samples has optimal dependence on n, as, 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 with probability $\frac{2}{3}$ requires $\Omega(\frac{n}{\log n})$ samples.*

This is the first sublinear-sample estimator for this fundamental property, and the lower bound (which follows easily from [27]) improves upon the previous best lower bound of $n/2^{O(\sqrt{\log n})}$ shown in [29].

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 employing this correspondence to our ends.

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 $\pm\delta/2$.

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 [27] 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, 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 11. *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:*

$$\begin{array}{ll}
\textit{Maximize:} & \sum_{x\in X} \frac{f_\pi(x)}{x}\left(y_x^+ - y_x^-\right) \\
\textit{Subject to:} & \\
\forall i \leq k^{c_1}, & \sum_x \left(y_x^+ - y_x^-\right) \cdot poi(xk, i) \leq k^{-c_2} \\
\forall i \leq k^{c_1}, & \sum_x \left(y_x^+ - y_x^-\right) \cdot poi(xk, i) \geq -k^{-c_2} \\
 & \sum_{x\in X} y_x^+ + y_x^- \leq 2 \\
 & \sum_{x\in X} \frac{y_x^+}{x} \leq n \quad \textit{and} \quad \sum_{x\in X} \frac{y_x^-}{x} \leq n \\
 & \forall x \in X,\ y_x^+ \geq 0, y_x^- \geq 0
\end{array}$$

In words, this linear program maximizes the discrepancy in property values of the distributions corresponding to y^+ and y^- subject to the following conditions: the first two constraints ensure that the fingerprint expectations of the two distributions are similar, the third condition ensures that y^+ and y^- together represent at most 2 units of probability mass, the fourth condition ensures that the two distributions have support at most n, and the last condition ensures that all elements of the support are assigned nonnegative probability values.

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^- that has objective function value v, we will 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 12. *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 11 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 [27]—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 12 are included in the full version of the paper.

5.2. Constructing Linear Estimators

Perhaps the most natural approach to constructing estimators for linear properties, dating back at least to the 1950's, [20] 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

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

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, while the total variance of all the fingerprint entries is roughly k.)

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.

Proposition 13. *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 14 (The Linear Estimator LP)**.**
Minimize: $2z^a + n \cdot (z^{b+} + z^{b-}) + k^{-c_2} \sum_{i=0}^{k^{c_1}} (z_i^+ + z_i^-)$
Subject to:

$$\forall x \in X, \quad \sum_{i=0}^{k^{c_1}} poi(xk, i)(z_i^+ - z_i^-) \geq \frac{f_\pi(x)}{x} - (z^a + \frac{z^{b-}}{x})$$

$$\forall x \in X, \quad \sum_{i=0}^{k^{c_1}} poi(xk, i)(z_i^+ - z_i^-) \leq \frac{f_\pi(x)}{x} + z^a + \frac{z^{b+}}{x}$$

$$\forall i \in [k^{c_1}], \quad z_i^+ \geq 0, z_i^- \geq 0, \quad z^a \geq 0, z^{b+} \geq 0, z^{b-} \geq 0.$$

To see the relation between the above definition and Proposition 13, 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 11 and the Linear Estimator LP of Definition 14 are dual linear programs. Complications arise, however, when one considers 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.

As our goal is to find a linear estimator of near-optimal performance, we start with a solution to the Lower Bound LP with objective value v, which, provided $c_2 > \frac{1}{2}$ is suitably chosen, yields a lower bound of $\approx \frac{v}{2}$, on the accuracy of estimating (via *any* algorithm) the desired property given k samples. We invoke duality to yield a k-sample linear estimator with coefficients described by the vector z, and with objective value also v in the Linear Estimator LP, with parameter $c_2 > \frac{1}{2}$ as above. The issue is that the entries of z may be unsuitably large, as the only bound we have on them is that of the objective function of the Linear Estimator LP, which yields that their sum is at most $v \cdot k^{c_2}$. Since $c_2 > \frac{1}{2}$, the entries may be bigger than $\sqrt{k}$, which corresponds to an estimator with inadmissibly super-constant variance.

5.3. Matrix Exponentials of Poisson Matrices

The aim of this section is to transform a solution to the Linear Estimator LP with $c_2 > 1/2$ 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, Theorem 1, then follows from the fact that the quadratic form $||z_\alpha||_2^2 = z e^{\alpha X} z^\intercal$ is a *log-convex* function of α, for arbitrary z and X, and thus we can bound the size of the entries of the coefficient vector z_α, for α in the interval $(0, 1)$, by interpolating between the values of its L_2 norm at the endpoints. Details are given in the full version of the paper.

6. An Optimal Linear Estimator For Entropy

In this section we describe an explicit linear estimator for entropy, which, given as input $k = \Omega\left(\frac{n}{\epsilon \log n}\right)$ 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 [27] 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.

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 [28] 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 natural primitives 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 $T_j(\cos(y)) = \cos(j \cdot y)$.

Definition 15. *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 $c_i = \frac{2s}{k}(1 - \cos(\frac{i\pi}{s}))$.*

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—and thus eventually leading to an estimator with small variance.

Lemma 16. *Each $B_i(x)$ may be expressed as $\sum_{j=0}^{\infty} a_{ij} poi(kx, j)$ for a_{ij} satisfying $\sum_{j=0}^{\infty} |a_{ij}| \leq k^{0.4}$*

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 17. *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.*

Let $f(y) \triangleq I(y)\left[\frac{1}{2y} + \log y - \log k\right]$, and provisionally set $z_i \triangleq f(i)$. Note that $\sum_{i=0}^{\infty} z_i \cdot poi(xk, i)$ accurately represents the logarithm function via the Poisson bumps in the interval $[\frac{s}{2k}, 1]$; the $\frac{1}{2y}$ term corrects for errors due to to the concavity of the logarithm function.

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

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

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 $\log x$ 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 16: $B_i(x) = \sum_j a_{i,j} poi(xk, j)$, 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 $\beta_i \triangleq z_{i-1} \cdot \frac{i}{k}$.

In the full version of this paper, we explicate this construction and prove Theorem 2.

REFERENCES

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

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

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

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

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

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

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

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

[9] M. Brautbar and A. Samorodnitsky, "Approximating entropy from sublinear samples," in *SODA, 2007*.

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

[11] A. Carlton, "On the bias of information estimates," *Psychological Bulletin*, vol. 71, pp. 108–109, 1969.

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

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

[14] B. Efron and C. Stein, "The jacknife estimate of variance," *Annals of Statistics*, vol. 9, pp. 586–596, 1981.

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

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

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

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

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

[20] G. Miller, "Note on the bias of information estimates," *Information Theory in Psychology. II-B*, pp. 95–100, 1955.

[21] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[22] L. Paninski, "Estimating entropy on m bins given fewer than m samples," *IEEE Trans. on Information Theory*, vol. 50, no. 9, pp. 2200–2203, 2004.

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

[24] 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.*, vol. 39, no. 3, pp. 813–842, 2009.

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

[26] 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*.

[27] 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.

[28] 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.

[29] P. Valiant, "Testing symmetric properties of distributions," in *STOC, 2008*.

[30] J. Victor, "Asymptotic bias in information estimates and the exponential (Bell) polynomials," *Neural Computation*, vol. 12, pp. 2797–2804, 2000.