Electronic Colloquium on Computational Complexity, Report No. 180 (2010)

Estimating the unseen: A sublinear-sample canonical estimator of distributions

Gregory Valiant   Paul Valiant

November 17, 2010

Abstract

We introduce a new approach to characterizing the unobserved portion of a distribution, which provides sublinear-sample additive estimators for a class of properties that includes entropy and distribution support size. Together with the lower bounds proven in the companion paper [29], this settles the longstanding question of the sample complexities of these estimation problems (up to constant factors). Our algorithm estimates these properties up to an arbitrarily small additive constant, using $O(n/\log n)$ samples; [29] shows that no algorithm on $o(n/\log n)$ samples can achieve this (where n is a bound on the support size, or in the case of estimating the support size, $1/n$ is a lower bound the probability of any element of the domain). Previously, no explicit sublinear-sample algorithms for either of these problems were known.

Additionally, our algorithm runs in time *linear* in the number of samples used.

ISSN 1433-8092

Think not, because no man sees,
Such things will remain unseen.
–Henry Wadsworth Longellow, from *"The Builders"*.

1 Introduction

Given samples from an unknown discrete distribution, what can we infer about the distribution? The empirical distribution of the samples roughly captures the portion of the distribution which we have observed, but *what can we say about the unobserved portion of the distribution*? Answers to this question are, at least implicitly, central to many estimation problems fundamental to statistics. Despite much research from both the statistics and computer science communities (originating, coincidentally, in independent work of Fisher [14], and Turing [16]—arguably the founding fathers of modern statistics and computer science), this question is still poorly understood. For the two important problems of estimating the support size, and estimating the entropy of a distribution, basic questions, such as the sample complexity of these tasks, have not been resolved. And this is not solely a theoretical question: in contrast to many tasks for which existing algorithms or heuristics perform well in practice (in some cases despite poor worst-case performance), for these two problems, there seems to be no approach that is fully embraced by practitioners [10]. Despite this, much of the recent theoretical work on these problems analyzes properties of existing heuristics. A new, practical algorithm for these tasks may, potentially, have widespread immediate application in the many fields for which these problems arise, including Biology, Ecology, Genetics, Linguistics, Neuroscience, and Physics (see the discussion and bibliographies in [9, 26]).

We introduce a new approach to characterizing the unobserved portion of a distribution, which provides sublinear-sample additive estimators for a class of properties that includes entropy and distribution support size. Together with the lower bounds proven in the companion paper [29], this settles the longstanding open question of the sample complexities of these estimation problems (up to constant factors). Our algorithm estimates these properties up to an arbitrarily small additive constant, using $O(n/\log n)$ samples. We show in [29] that no algorithm on $o(n/\log n)$ samples can achieve this. Here, n is a bound on the support size.[1] Previously, no explicit sublinear-sample algorithms for either of these problems were known.[2] Finally, we note that our algorithm runs in time *linear* in the number of samples used.

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

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

[2] See [26] for a nonconstructive proof of the existence of a $o(n)$-sample entropy estimator. Prior to [29], the previous lower-bounds were $n/2^{\Theta(\sqrt{\log n})}$, from [30].

1.1 Historical Background

The problem of estimating an unknown discrete distribution from "too few" samples has a very rich history of study in both statistics and computer science, with early contributions from both R.A Fisher, and Alan Turing. In the early 1940's, R. A. Fisher was approached by a naturalist, Corbet, who had just returned from two years of collecting butterflies in the Malay peninsula. Corbet presented Fisher with data on his butterfly collections—specifically, he indicated the number of species for which he had only seen a single specimen (118 species), the number of species for which he had two specimens (74 species), three specimens (44 species), and so on. Corbet hoped that from this data, the great statistician Fisher would be able to deduce some properties of the true distribution of butterflies in Malay, and in particular, he wanted an estimate of the number of new species he might find if he were to return to the Malay jungle for another 2 years. Using basic properties of the Poisson distribution, Fisher provided a partial answer to these questions in [14].

At roughly the same time, at the height of WWII, Alan Turing and I.J. Good were working on a similar problem in the rather different context of the pivotal British war-effort to analyze the statistics of the German Enigma Machine ciphers. After the war, the results of their work, the *Good-Turing frequency estimation* scheme were published [16]. In addition to many practical applications of the Good-Turing estimates, there has been considerable recent work from the computer science community analyzing variants of these estimation schemes [22, 24, 25, 31, 32]. While the high-level goals of these estimators are related to our own, the analysis typically fixes a distribution and considers the behavior as the number of samples taken approaches infinity, and thus is somewhat orthogonal to the questions considered here.

The specific problem of estimating the support size of an unknown distribution (also referred to as the problem of estimating the number of species in a population) has a very long history of study and arises in many contexts (see [9] for several hundred references). Because arbitrarily many species can lie in an arbitrarily small amount of probability mass, analysis of the sample complexity of this problem is generally parameterized in terms of n, where elements of the distribution are restricted to have probability mass at least $1/n$. Tight multiplicative bounds of $\Omega(n/\alpha^2)$ for approximating this problem to a multiplicative factor of α are given in [4, 12] 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 [28], showing that for any constant $\delta > 0$, any estimator that obtains additive error at most $(1/2 - \delta)n$ with probability at least 2/3 requires at least $n^{1-o(1)}$ samples. To the best of our knowledge, there were no improvements upon the trivial $\Omega(n)$ upper bound for this problem.

For the problem of entropy estimation, there has been recent work from both the computer science and statistics communities. Batu *et al.* [5, 6, 13], Guha *et al.* [17], and Valiant [30] considered the problem of multiplicatively estimating the entropy; in all these works, the estimation algorithm has the following basic form: given a set of samples, discard the species that occur infrequently and return the entropy of the empirical distribution of the frequently-occurring elements, adjusted by some function of the amount of missing probability mass. In particular, no attempt is made to understand the portion of the true distribution consisting of infrequently occurring elements. In [26, 27], Paninski proved the existence of a sublinear sample estimator for additively approximating the entropy to within a constant; the proof is via a direct application of the Stone-Weierstrass theorem to the set of Poisson functions. Prior to [29], the best lower bound was $n/2^{\Theta(\sqrt{\log n})}$, given in [30].

Additionally, there has been much work on estimating the support size (and the general problem of estimating frequency moments) and estimating the entropy in the setting of *streaming*, in which one has access to very little memory and can perform only a single pass over the data [2, 3, 8, 11, 18, 19, 20, 33].

Teleologically, perhaps the work most similar to our own is Orlitsky *et al.*'s investigation into what they term "pattern maximum likelihood" [1, 23]. Their work is prompted by the following natural question: given a set of samples, what distribution maximizes the likelihood of seeing the observed species frequencies, that is, the number of species observed once, twice, etc? (What Orlitsky *et al.* term the *pattern* of a sample, we call the *fingerprint*, as in Definition 3.) While it seems unclear how to prove that such a likelihood maximizing distribution would, necessarily, have similar property values to the true distribution, at least intuitively one might hope that this is true. From a computational standpoint, while Orlitsky *et al.* show that such likelihood maximizing distributions can be found in some specific settings, the problem of finding or approximating such distributions in the general setting seems daunting.

1.2 Definitions and Examples

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

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

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

We now define the notion of a *symmetric property*

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

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

Throughout, we will be dealing exclusively with symmetric properties. For such properties, the fingerprint of a sample contains all the useful information about the sample: for any estimator that uses the actual samples, there is an estimator of equal performance that takes as input only the fingerprint of the samples (see [5, 7], 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.

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. Any symmetric property is clearly a function of the histogram of the distribution.

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

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

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

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

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

The structure of the distribution of fingerprints intimately involves the Poisson distribution. Throughout, we use $Poi(\lambda)$ to denote the Poisson distribution with expectation λ, and for a non-negative integer j, $poi(\lambda, j) := \frac{\lambda^j e^{-\lambda}}{j!}$, denotes the probability that a random variable distributed according to $Poi(\lambda)$ takes value j. Additionally, for integers $i \geq 0$, we refer to the function $poi(x, i)$, viewed as a function of the variable x, as the jth *Poisson function*.

We now provide two clarifying examples of the above definitions:

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

Suppose that the true distribution of fish is the following:

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

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

Example 7. *Consider the uniform distribution on $[n]$, which has histogram h such that $h(\frac{1}{n}) = n$, and $h(x) = 0$ for $x \neq \frac{1}{n}$. Let $k \leftarrow Poi(5n)$ be a Poisson-distributed random number, and let X be the result of drawing k independent samples from the distribution. The number of occurrences of each element of $[n]$ will be independent, distributed according to $Poi(5)$. Note that $\mathcal{F}_X(i)$ and $\mathcal{F}_X(j)$ are* not *independent (since, for example, if $\mathcal{F}_X(i) = n$ then it must be the case that $\mathcal{F}_X(j) = 0$, for $i \neq j$). A fingerprint of a typical trial will look roughly like $\mathcal{F}(i) \approx n \cdot poi(5, i)$.*

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

Definition 8. *A symmetric distribution property π is* (ϵ, δ)-continuous *if for all distributions D_1, D_2 with respective histograms h_1, h_2 satisfying $R(h_1, h_2) \leq \delta$ it follows that $|\pi(D_1) - \pi(D_2)| \leq \epsilon$.*

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

Fact 9. *For a distribution $p \in \mathcal{D}^m$, and $\delta > 0$*

- *The entropy, $H(p) := -\sum_i p(i) \cdot \log p(i)$ is (δ, δ)-continuous, with respect to the relative earthmover distance.*

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

2 Results and Outline

We view the main contribution of this work to be the introduction of a novel approach to creating estimators for symmetric distribution properties. We hope (and believe) that variants of our proposed estimator will prove useful in practice.

Our main technical result is a canonical estimator for relative-earthmover continuous properties. We stress that our estimator is truly canonical in that it is agnostic to the choice of property that one is trying to estimate. In particular, the estimator works by first constructing a distribution completely independently of the property in question, and then simply returning the evaluation of the property on this distribution. Even if the property in question is computationally intractable to evaluate, the first stage of our estimator still runs in time linear in the number of samples, returning a distribution capturing the value of the property.

Theorem 1. *For sufficiently large n, and any constant $c > 1$, given $c\frac{n}{\log n}$ independent samples from $D \in \mathcal{D}^n$, with probability at least $1 - e^{-n^{.03}}$ over the random samples, our algorithm returns a distribution D', representable as an $O(c\frac{n}{\log n})$-length vector, such that the relative-earthmover distance between D and D' satisfies*

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

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

We suspect that the $\log c$ term is an artifact of our analysis, rather than a property of the algorithm. For entropy and support size, the following corollaries follow immediately from Theorem 1 together with Fact 9:

Corollary 10. *There exists a function $f : \mathbb{R}^+ \to \mathbb{R}^+$, with $f(x) = O(x^{2+\epsilon})$ for all $\epsilon > 0$, such that for sufficiently large n and any constant $\alpha > 0$, given $f(\alpha)\frac{n}{\log n}$ independent samples from $D \in \mathcal{D}^n$, our estimator runs in time $f(\alpha)O(\frac{n}{\log n})$ and with probability at least $1 - e^{-n^{.03}}$ returns a value ϕ such that*

$$|\phi - H(D)| < \frac{1}{\alpha},$$

where $H(D)$ is the entropy of distribution D.

Corollary 11. *There exists a function $f : \mathbb{R}^+ \to \mathbb{R}^+$, with $f(x) = O(x^{2+\epsilon})$ for all $\epsilon > 0$, such that for sufficiently large n and any constant $\alpha > 0$, given $f(\alpha)\frac{n}{\log n}$ independent samples from $D \in \mathcal{D}^n$, with $\min_{i \in [n]: p(i) > 0} p(i) \geq 1/n$, our estimator runs in time $f(\alpha)O(\frac{n}{\log n})$ and with probability at least $1 - e^{-n^{.03}}$ returns a value ϕ such that*

$$|\phi - S(D)| < \frac{n}{\alpha},$$

where $S(D) := |\{i : p_i > 0\}|$ is the support size of distribution D.

2.1 Poisson Samples

Before describing the linear program and motivating intuitions behind the proof of Theorem 1, 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. In such a set of samples, the number of occurrences of any two elements are not independent; however, as indicated in Example 7, 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, the number of occurrences of each 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.

Since 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 proposition (whose proof is included in the Appendix, see Proposition 21) formalizes this intuition, and allows us to prove statements about k-sample fingerprints by considering the structurally more simple $Poi(k)$-sample fingerprints.

Proposition. *Given $k > 30$, and any set of fingerprints A, let $\overline{A}$ be the set of fingerprints that can be obtained by adding or removing at most $k^{.6}$ samples from some fingerprint in set A. Let $\mathcal{F}$ denote a random k-sample fingerprint, and let $\mathcal{F}'$ denote a fingerprint obtained from choosing $k' \leftarrow Poi(k)$, random samples. Then*

$$\Pr[\mathcal{F} \in A] \leq \Pr[\mathcal{F}' \in \overline{A}] + e^{-k^{.1}/2}.$$

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 $\Pr[Y_j = 1] = poi(kp_j, i)$ 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 have good concentration about this expectation.

2.2 Outline

We informally describe the linear program, solutions to which can be regarded as histograms. Given a fingerprint $\mathcal{F}$, the linear program is constructed to recover a histogram h' with the property that if we were given a set of samples X from h', for each i, $E[\mathcal{F}_X(i)] \approx \mathcal{F}(i)$.

The proof of correctness of our linear program has two parts: in the first part we show that with high probability, if the linear program is created from a set of at least $\Omega(\frac{n}{\log n})$ samples then it has a feasible point that is "close" to the true distribution. In the second, more involved part of the proof, we argue that any pair of solutions are "close"; the core of this argument is an earthmoving construction in which we leverage properties of Chebyshev polynomials.

Given these two parts, by the triangle inequality, with high probability any solution must be "close" to the true distribution; thus, by the relative-earthmover continuity of the properties in question, the recovered distribution will have a similar property value to that of the true unknown distribution.

In the remainder of this extended abstract, we summarize the intuition and general structure of the components of the proof. *For clarity, we include complete proofs in the context of a full write-up in the appendix.*

3 The Linear Program

Intuitively, given the fingerprint $\mathcal{F}$ of a set of samples from an unknown distribution with histogram h, we wish to reconstruct a distribution h' that is similar to h. For the frequently-occurring elements, say elements whose probabilities are at least k^{-1+a}, for some small constant $a \in (0, 1)$, we can simply let h' agree with the empirical distribution, namely setting $h'(j/k) = \mathcal{F}(j)$. For the portion of h'

below probability k^{-1+a}, we would like the fingerprint expectations for samples from h' to roughly agree with the observed fingerprints $\mathcal{F}(j)$ in this regime (roughly, for $j \leq k \cdot k^{-1+a} = k^a$).

To see why this makes sense, consider a high-probability element i (with $p(i) > k^{-1+a}$), and note that we expect to see it roughly $k^a \gg \log k$ times in our sample, and thus we can expect good concentration around this expectation. In contrast, for the portion of h below probability k^{-1+a}, we instead rely on the concentration of each fingerprint entry, $\mathcal{F}(j)$, about its expectation.

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

Given a k-sample fingerprint $\mathcal{F}$, choose $c \in [1,2]$ such that the total "mass" in $\mathcal{F}$ between frequencies ck^a and $ck^a + 4k^{.6a}$ is at most $4k^{-.4a}$. Namely,

$$\sum_{j=\lceil ck^a \rceil}^{\lceil ck^a + 4k^{.6a} \rceil} j\mathcal{F}(j) \leq 4k^{1-.4a}.$$

Note that such a choice of c can be found, for otherwise the total number of samples accounted for by fingerprint entries in the interval $[k^a, 2k^a]$ would exceed k.

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

Definition 12. *Given a k-sample fingerprint $\mathcal{F}$, bounds $A := ck^{-1+a}$, $B := 4k^{-1+.6a}$, and real number $\gamma := k^{-3/2}$, the linear program consists of variables $v_x \geq 0$ for all $x \leq A + B/2$ in the set $X := \{\gamma, 2^2\gamma, 3^2\gamma, 4^2\gamma, \ldots, A + B/2\}$, subject to the following three conditions:*

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

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

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

We remark that we carefully chose the set X of probabilities for which we solve. If we instead take the set X to be a very fine mesh—for example $\{\frac{1}{k^2}, \frac{2}{k^2}, \ldots, 1\}$—several of the proofs would simplify, but then the computation time to solve the resulting linear program would be $O(k^7)$. We instead opt to take a coarse quadratically-spaced mesh so as to minimize the number of variables for which we solve. Perhaps coincidentally, while our approach seems to require at least $k^{1/4}$ variables in the LP, we use $|X| = \Theta(k^{\frac{1}{4}+a}) \leq k^{1/3.5}$ variables, and thus the LP can be solved in time linear in k, the number of samples[21].

Given a solution to the linear program v, the definition below extends v to yield the histogram h^v, which we refer to as the *histogram associated to the solution* v. Roughly, to obtain h^v, we start with v and first adjoin the empirical distribution for probabilities above $A + B$, then round each value down to the nearest integer. Finally, to compensate for the decrease in mass resulting from the rounding, we scale the support by a factor of $1+\epsilon$ (while keeping the values of the histogram fixed) thereby increasing the total mass in the histogram by a factor of $(1+\epsilon)$, where ϵ is chosen so as to make the total probability mass equal 1 after the rounding. We formalize this process below:

Definition 13. *Let $X = \{\gamma, 2^2\gamma, 3^2\gamma, 4^2\gamma, \ldots, A + B/2\}$ be the set of probabilities for which the linear program solves. Given a k-fingerprint $\mathcal{F}$ and a solution v to the associated linear program, the corresponding histogram h^v is derived from v according to the following process in which generalized histogram h' is constructed, then rounded to create h^v.*

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

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

Algorithm:. The Estimator
`Given a set of` k `samples having fingerprint` $\mathcal{F}$`:`

- `Construct the linear program of Definition 12 corresponding to` $\mathcal{F}$`.`
- `Find a solution` v `to the the linear program.  If no solution exists, output FAIL.`
- `Output histogram` h^v `associated to solution` v`, as defined in Definition 13.`

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

Theorem 2. *For a constant $\delta \in (0,1]$, consider a sample consisting of k independent samples from a distribution h of support size at most $\delta k \log k$. With probability at least $1 - e^{-k^{.04}}$, the linear program of Definition 12 has a solution and furthermore, for any solution to the linear program, v, the associated histogram h^v constructed from v in Definition 13 satisfies*

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

The proof of Theorem 2 has two parts. In the first part, we show that, with the claimed probability, the linear program has a feasible point v, whose associated histogram h^v is close in relative earthmover distance to the true distribution, h. In the second part, we argue that for any two solutions v, w, their associated histograms are close. By the triangle inequality, these two parts prove the theorem.

We construct the feasible point v' in two stages, the first discretizes the true histogram h; the second makes small adjustments based on $\mathcal{F}$ so as to satisfy the second condition of the linear program – that the "total weight" including the probability mass in the empirical distribution derived from the fingerprint above probability $A+B$ is 1. We will then show that with high probability, the first and third conditions of the linear program are also satisfied by v'.

The discretization (in Step 1 below) proceeds by linear interpolation. That is, given probability mass at probability y, we find consecutive elements of X, that sandwich y, that is, x_i, x_{i+1} such that $x_i \leq y \leq x_{i+1}$, and distribute the histogram entry at y linearly between v_{x_i} and $v_{x_{i+1}}$. We note that such interpolation preserves both total probability mass, and the sum of the entries in the histogram. As mentioned above, if we were willing to sacrifice running time, a simple nearest-neighbor discretization to a much more finely spaced set X would suffice.

For a solution v to the linear program, let $v(i)$ denote v_{x_i}. We construct v' as follows:

1. (a) Initialize $v'(*) = 0$
   (b) For each $y < A + B/2$ such that $h(y) > 0$, let i be an index for which $y \in [x_i, x_{i+1}]$, or $i = 0$ if $y < x_1$
   (c) Modify v' by increasing $v'(i) \leftarrow v'(i) + h(y)\frac{x_{i+1}-y}{x_{i+1}-x_i}$, and $v'(i+1) \leftarrow v'(i+1) + h(y)\frac{y-x_i}{x_{i+1}-x_i}$
2. (a) Compute how much the second condition of the linear program is violated: let $w = \left(\sum_i x_i v'(i)\right) + \left(\sum_{i \geq k(A+B)} \frac{i}{k}\mathcal{F}(i)\right) - 1$
   (b) If $w < 0$, increase $v'_{A+B/2} \leftarrow v'_{A+B/2} + \frac{|w|}{A+B/2}$
   (c) Otherwise if $w > 0$, decrease v' arbitrarily (while still keeping it nonnegative) so as to satisfy the second condition of the linear program.

At least intuitively, we expect each fingerprint entry to be close to its expectation, and the high-probability portion of the empirical distribution to be similar to the high-probability portion of h, and thus v' will be in the feasible region with high probability. In addition, since v' was constructed from h, we expect its associated histogram $h^{v'}$ to be close in relative earthmover distance to h. This intuition is formalized in the following proposition (see Proposition 23 in the appendix for a proof):

Proposition. *For sufficiently large k, given a k-sample fingerprint $\mathcal{F}$ from a distribution of support size at most $k^{1.1}$, then with probability at least $1 - e^{-k^{.04}}$ the associated linear program given in Definition 12 has a feasible point v' whose associated histogram $h^{v'}$ is at most $17k^{-.4a}$ far in relative-earthmover distance from h, the actual histogram of the distribution.*

3.1 All Solutions are Good Solutions

We now argue that with high probability over the set of samples, for any pair of solutions v, w, to a linear program corresponding to a set of k samples from a distribution of support at most $n = \delta k \log k$, their associated histograms satisfy $R(h^v, h^w) \leq O(\sqrt{\delta} \cdot \max\{1, |\log \delta|\})$. Theorem 2 will then follow, via the triangle inequality, from the proposition above. To prove that the histograms yielded from a pair of solutions are close, we construct an earthmoving scheme that leverages the fact that the fingerprint expectations of h^w and h^v are close.

Before describing the intuition behind our construction, we start by defining a very natural class of earthmoving schemes.

Definition 14. *For a given k, a β-*bump earthmoving scheme *is defined by a sequence of positive real numbers $\{c_i\}$, the* bump centers*, and a sequence of functions $\{f_i\} : (0,1] \to \mathbb{R}$ such that $\sum_{i=0}^{\infty} f_i(x) = 1$ for each x, and each function f_i may be expressed as a linear combination of Poisson functions, $f_i(x) = \sum_{j=0}^{\infty} a_{ij} poi(kx, j)$, such that $\sum_{j=0}^{\infty} |a_{ij}| \leq \beta$.*

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

Definition 15. *For given n, k, a bump earthmoving scheme (c, f) is ϵ-*good *if for any generalized histogram h, the relative earthmover distance between h and $(c, f)(h)$ is at most ϵ.*

Perhaps the most natural bump earthmoving scheme—which we will end up using a refinement of—is where $f_i(x) = poi(kx, i)$ and $c_i = \frac{i}{k}$, where for $i = 0$, c_i is chosen, say, as $\frac{1}{2k}$ to avoid a logarithm of 0 when evaluating relative earthmover distance. This is a valid earthmoving scheme since $\sum_{i=0}^{\infty} f_i(x) = 1$ for any x.

The motivation for this construction is the fact that, for any i, the amount of probability mass that ends up at c_i in $(c, f)(h)$ is exactly c_i times the expectation of the ith fingerprint in a $Poi(k)$-sample from h. Thus if we apply this earthmover scheme to two histograms derived from solutions to the linear program, their fingerprint expectations will closely match, and we would be left with a pair of histograms h^1, h^2 such that $R(h^1, h^2)$ is small.

The problem with this "Poisson bump" earthmoving scheme is that it has bad relative earthmover distance, particularly towards the origin. This is due to the fact that most of the mass that starts at a probability below $\frac{1}{k}$ will end up in the zeroth bump, no matter if it has probability nearly $\frac{1}{k}$, or the rather lower $\frac{1}{n}$. The situation gets significantly better for higher Poisson functions: most of the mass of $Poi(i)$ lies within relative distance $O(\frac{1}{\sqrt{i}})$ of i. We will therefore construct a scheme that uses Poisson functions $poi(kx, i)$ for $i \geq \log k$, but takes great care to construct "narrower" bumps below this region.

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

Lemma 16. *For $n > k$, letting δ be such that $n = \delta k \log k$, there exists an $O(\sqrt{\delta} \cdot \max\{1, |\log \delta|\})$-good $k^{0.3}$-bump earthmoving scheme*

In fact, we will construct a single scheme for all δ.

Definition 17. *The* Chebyshev earthmoving scheme *is defined in terms of k as follows. Let $s = \frac{1}{5} \log k$. For $i \geq s$, let $f_i(x) = poi(kx, i)$ and $c_i = \frac{i}{k}$.*

Define $g(y) = \sum_{j=-s}^{s-1} \cos(jy)$. Define $g'(y) = g(y) + g(y - \frac{\pi}{s})$ and, for $i \in \{0, \ldots, s-1\}$ define $g_i''(y) = g'(y - \frac{i\pi}{s}) + g'(y + \frac{(i+1)\pi}{s})$. Let $t_i(x)$ be the linear combination of Chebyshev polynomials so that $t_i(\cos(y)) = g_i''(y)$. We thus define the final s bumps to be $f_i(x) = \frac{1}{4s} t_i(1 - \frac{xk}{2s}) \sum_{j=0}^{s-1} poi(xk, j)$, for $i \in \{0, \ldots, s-1\}$. That is, $f_i(x)$ is related to $g_i''(y)$ by the coordinate transformation $x = \frac{2s}{k}(1 - \cos(y))$, and scaling by $\frac{1}{4s} \sum_{j=0}^{s-1} poi(xk, j)$. For these bumps, define $c_i = \frac{2s}{k}(1 - \cos(\frac{(i+1)\pi}{s}))$.

4 A Final Remark

The vast majority of statistical estimators, including many of those proposed for the problem of estimating the entropy of a distribution, can be expressed as the computation of a dot product with the fingerprint of the samples, $\mathcal{F}$. In particular, these estimators are *linear*, in that they calculate a vector of coefficients $a_1, \ldots, a_k$, and then return the estimate $\phi := \sum_i a_i \mathcal{F}(i)$.

From this vantage point, our estimator makes the leap from harnessing the power of *linear algebra*, to harnessing the power of *linear programming.* In addition to the more obvious directions for future investigation, an intriguing question is whether this additional power is necessary; curiously, the nonconstructive proof of Paninski [27] shows the existence of a sublinear-sample *linear* estimator. Can a linear estimator achieve sample complexity $O(\frac{n}{\log n})$?

References

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

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

[3] Z. Bar-Yossef, T.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.*

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

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

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

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

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

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

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

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

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

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

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

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

[16] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika, 40(16):237264*, 1953.

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

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

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

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

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

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

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

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

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

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

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

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

[29] G. Valiant and P. Valiant. A CLT and tight lower bounds for estimating entropy. *Available at: http://www.cs.berkeley.edu/~gvaliant/papers/cltVV.pdf*, 2010.

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

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

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

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

Note: For clarity, in the following appendix we include the complete proofs in the context of a full write-up.

A Defining the Linear Program

Intuitively, given the fingerprint $\mathcal{F}$ of a set of samples from an unknown distribution with histogram h, we wish to reconstruct a distribution h' that is similar to h. For the frequently-occurring elements, say elements whose probabilities are at least k^{-1+a}, for some small constant $a \in (0,1)$, we can simply let h' agree with the empirical distribution, namely setting $h'(j/k) = \mathcal{F}(j)$. For the portion of h' below probability k^{-1+a}, we would like the fingerprint expectations for samples from h' to roughly agree with the observed fingerprints $\mathcal{F}(j)$ in this regime (roughly, for $j \leq k \cdot k^{-1+a} = k^a$).

To see why this makes sense, consider a high-probability element i (with $p(i) > k^{-1+a}$), and note that we expect to see it roughly $k^a \gg \log k$ times in our sample, and thus we can expect good concentration around this expectation. In contrast, for the portion of h below probability k^{-1+a}, we instead rely on the concentration of each fingerprint entry, $\mathcal{F}(j)$, about its expectation.

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

Given a k-sample fingerprint $\mathcal{F}$, choose $c \in [1,2]$ such that the total "mass" in $\mathcal{F}$ between frequencies ck^a and $ck^a + 4k^{.6a}$ is at most $4k^{-.4a}$. Namely,

$$\sum_{j=\lceil ck^a \rceil}^{\lceil ck^a + 4k^{.6a} \rceil} j\mathcal{F}(j) \leq 4k^{1-.4a}.$$

Note that such a choice of c can be found, for otherwise the total number of samples accounted for by fingerprint entries in the interval $[k^a, 2k^a]$ would exceed k.

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

Definition 18. *Given a k-sample fingerprint $\mathcal{F}$, bounds $A := ck^{-1+a}$, $B := 4k^{-1+.6a}$, and real number $\gamma := k^{-3/2}$, the linear program consists of variables $v_x \geq 0$ for all $x \leq A + B/2$ in the set $X := \{\gamma, 2^2\gamma, 3^2\gamma, 4^2\gamma, \ldots, A + B/2\}$, subject to the following three conditions:*

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

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

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

We remark that we carefully chose the set X of probabilities for which we solve. If we instead take the set X to be a very fine mesh—for example $\{\frac{1}{k^2}, \frac{2}{k^2}, \ldots, 1\}$—several of the proofs would simplify, but then the computation time to solve the resulting linear program would be $O(k^7)$. We instead opt to take a coarse quadratically-spaced mesh so as to minimize the number of variables for which we solve. Perhaps coincidentally, while our approach seems to require at least $k^{1/4}$ variables in the LP, we are able to show that fewer than $k^{1/3.5}$ variables suffice, and thus the LP can be solved in time linear in k, the number of samples[21].

Given a solution to the linear program v, the definition below extends v to yield the histogram h^v, which we refer to as the *histogram associated to the solution* v. Roughly, to obtain h^v, we start with v and first adjoin the empirical distribution for probabilities above $A+B$, then round each value down to the nearest integer. Finally, to compensate for the decrease in mass resulting from the rounding, we scale the support by a factor of $1+\epsilon$ (while keeping the values of the histogram fixed) thereby increasing the total mass in the histogram by a factor of $(1+\epsilon)$, where ϵ is chosen so as to make the total probability mass equal 1 after the rounding. We formalize this process below:

Definition 19. *Let $X = \{\gamma, 2^2\gamma, 3^2\gamma, 4^2\gamma, \ldots, A+B/2\}$ be the set of probabilities for which the linear program solves. Given a k-fingerprint $\mathcal{F}$ and a solution v to the associated linear program, the corresponding histogram h^v is derived from v according to the following process in which generalized histogram h' is constructed, then rounded to create h^v.*

1. *set $h'(*) = 0$ and $h^v(*) = 0$.*

2. *for all $x \in X$, let $h'(x) := v_x$.*

3. *for all integers $j \geq k(A+B)$, let $h'(j/k) := \mathcal{F}(j)$.*

4. *for all probabilities $x : h'(x) \neq 0$, set $h^v((1+\epsilon)x) := \lfloor h'(x) \rfloor$, where $\epsilon := \frac{\sum_{x\in X} x(v_x - \lfloor v_x \rfloor)}{1-\sum_{x\in X} x(v_x - \lfloor v_x \rfloor)}$.*

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

The following trivial proposition guarantees that the final rounding and scaling does not alter the histogram by much in the relative earthmover metric.

Proposition 20. *Given a histogram h^v associated to a solution v to the linear program, let h', as in Definition 19 denote the generalized histogram obtained prior to rounding, ie $h'(x) = v_x$ for $x \leq A+B$, and $h'(j/k) = \mathcal{F}(j)$ for integers $j \geq k(A+B)$. Then*

$$R(h^v, h') \leq k^{-1/2}.$$

Proof. Consider the earthmoving scheme for obtaining h^v from h', in which the probability mass $x\lfloor h^v(x) \rfloor$ at probability x is moved to probability $(1+\epsilon)x$, and then the tiny bit of remaining probability mass, $x(h^v(x) - \lfloor h^v(x) \rfloor)$, is moved anywhere, so as to obtain h'. We round at most $|X| \leq 2k^{1/4+a/4}$ of the entries of h', and the total probability mass that is changed in each rounding is at most $A+B/2 \leq 2k^{-1+a}$, so the total reduction in mass in the rounding step is at most $4k^{-3/4+2a}$, and thus $\epsilon \leq k^{-.6}$. The cost of the first stage of our earth moving scheme is thus at most $\log(1+\epsilon) \leq k^{-.6}$. In the second stage, the amount of remaining mass is precisely $\sum_{x\in X} x(v_x - \lfloor v_x \rfloor) \leq 4k^{-3/4+2a}$, and thus we can move this mass to any probability above $k^{-3/2}$ at cost at most $4k^{-3/4+2a}\log(k^{3/2}) \leq k^{-.6}$, for sufficiently large k. □

```
Algorithm:. The Estimator
Given a set of $k$ samples, with fingerprint $\mathcal{F}$:
    • Construct the linear program of Definition 18 corresponding to $\mathcal{F}$.
    • Find a solution $v$ to the the linear program.  If no solution exists, output
      FAIL.
    • Output histogram $h^v$ associated to solution $v$, as defined in Definition 19.
```

The correctness of our estimator is captured in the following theorem, from which Theorem 1 follows:

Theorem 3. *For a constant $\delta \in (0,1]$, consider a sample consisting of k independent samples from a distribution h of support size at most $\delta k \log k$. With probability at least $1 - e^{-k^{.04}}$, the linear program of Definition 18 has a solution and furthermore, for any solution to the linear program, v, the associated histogram h^v constructed from v in Definition 19 satisfies*

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

We note that that factor of $\max(1, |\log \delta|)$ in the statement of Theorem 3 is, likely, an artifact of our analysis, and perhaps can be removed with slightly tighter analysis.

The proof of Theorem 3 has two parts. In the first part, we show that, with the claimed probability, the linear program has a feasible point v, whose associated histogram h^v is close in relative earthmover distance to the true distribution, h. In the second part, we argue that for any two solutions v, w, their associated histograms are close. By the triangle inequality, these two parts prove the theorem.

We now prove an easy lemma which will enable us to prove Theorem 3 by arguing about $Poi(k)$-sample fingerprints, instead of simply k-sample fingerprint. As noted above, in a $Poi(k)$-sample, the number of occurrences of each support element will be independent (and Poisson distributed), allowing us to apply standard Chernoff bounds to various quantities. Intuitively, the distribution of a k-sample fingerprint should be very similar to that of a $Poi(k)$-sample fingerprint. The following easy proposition makes this intuition rigorous.

Proposition 21. *Given $k > 30$, and any set of fingerprints A, let $\overline{A}$ be the set of fingerprints that can be obtained by adding or removing at most $k^{.6}$ samples from some fingerprint in set A. Let $\mathcal{F}$ denote a random k-sample fingerprint, and let $\mathcal{F}'$ denote a fingerprint obtained from choosing $k' \leftarrow Poi(k)$, random samples. Then*

$$\Pr[\mathcal{F} \in A] \leq \Pr[\mathcal{F}' \in \overline{A}] + e^{-k^{.1}/2}.$$

Proof. By the bound on the Poisson tail probabilities given in Corollary 32, $\Pr[|k' - k| > k^{.6}] \leq e^{-k^{.1}/2}$. Thus for any set B, $\Pr[\mathcal{F}' \in B]$ and $\Pr\left[\mathcal{F}' \in B | k' \in [k - k^{.6}, k + k^{.6}]\right]$ will differ by at most $e^{-k^{.1}/2}$. Given that k lies in $[k - k^{.6}, k + k^{.6}]$, by a trivial coupling argument in which the first $\min(k, k')$ samples are common to both the k-sample and the k'-sample, we conclude that $\Pr[\mathcal{F} \in A] \leq \Pr[\mathcal{F}' \in \overline{A}] + e^{-k^{.1}/2}$, where $\overline{A}$ is the set of fingerprints obtained by adding or removing at most $k^{.6}$ samples from some fingerprint in set A. □

The following corollary will prove convenient, and illustrates the application of the above result:

Corollary 22. *Fix an integer $k > 30$. Given a distribution D with histogram h, and a k-sample fingerprint f yielded by taking k samples from D, for any integer i let $p_i := \sum_{x:h(x)\neq 0} h(x) poi(xk, i)$ denote the expected ith fingerprint entry. Then*

$$\Pr(|\mathcal{F}(i) - p_i| > 2k^{.6}) \leq 2e^{-k^{.1}/2}.$$

Proof. From Proposition 21, we consider taking $k' \leftarrow Poi(k)$ samples from D, and note that the number of occurrences of each element are independent random variables, since k' is taken from a Poisson distribution. Let $S_{i,j}$ denote the boolean random variable representing whether the jth element of the support is sampled exactly i times, and let $S_i := \sum_j S_{i,j}$. By elementary Chernoff bounds $\Pr(|S_i - E[S_i]| \geq k^{.6}) \leq e^{-k^{.2}/4}$.

Since adding or removing at most $k^{.6}$ samples can change S_i by at most $k^{.6}$, we conclude that $\Pr(|\mathcal{F}(i) - p_i| \geq 2k^{.6}) \leq e^{-k^{.2}/4} + e^{-k^{.1}/2} \leq 2e^{-k^{.1}/2}$, as desired. □

B A Feasible Point

In this section we show that with high probability, the linear program of Definition 18 has a feasible point, v', whose associated histogram $h^{v'}$ is close to h in relative-earthmover distance. We explicitly construct such a solution v' from the true histogram h. That v' is in the feasible region follows from our sufficiently dense choice of the set X of probabilities for which to solve, together with elementary concentration inequalities. We have the following proposition:

Proposition 23. *For sufficiently large k, given a k-sample fingerprint $\mathcal{F}$ from a distribution of support size at most $k^{1.1}$, then with probability at least $1 - e^{-k^{.04}}$ the associated linear program given in Definition 18 has a feasible point v' whose associated histogram $h^{v'}$ is at most $17k^{-.4a}$ far in relative-earthmover distance from h, the actual histogram of the distribution.*

To prove the proposition, we explicitly construct a potential solution, v', which is very similar to the portion of the true histogram h below probability $A + B/2$, but has support consisting of the probabilities for which the linear program solves. Additionally, we will make sure that v' has the appropriate mass (ie the second condition of the LP is satisfied). In Lemmas 24 and 25 we show that with high probability, v' satisfies the first and third conditions of the linear program, respectively, and thus v' is in the feasible region of the linear program. By Proposition 20 and the triangle inequality, the final rounding does not change the relative earthmover distance by much, and thus it suffices to analyze the generalized histogram obtained from v' prior to rounding.

We construct the feasible point v' in two stages, the first discretizes the true histogram h; the second makes small adjustments based on $\mathcal{F}$ so as to satisfy the second condition of the linear program – that the "total weight" including the probability mass in the empirical distribution derived from the fingerprint above probability $A + B$ is 1. We will then show that with high probability, the first and third conditions of the linear program are also satisfied by v'.

The discretization (in Step 1 below) proceeds by linear interpolation. That is, given probability mass at probability y, we find consecutive elements of X, that sandwich y, that is, x_i, x_{i+1} such that $x_i \leq y \leq x_{i+1}$, and distribute the histogram entry at y linearly between v_{x_i} and $v_{x_{i+1}}$. We note that such interpolation preserves both total probability mass, and the sum of the entries in the histogram. As mentioned above, if we were willing to sacrifice running time, a simple nearest-neighbor discretization to a much more finely spaced set X would suffice.

For a solution v to the linear program, let $v(i)$ denote v_{x_i}. We construct v' as follows:

1. (a) Initialize $v'(*) = 0$

   (b) For each $y < A + B/2$ such that $h(y) > 0$, let i be an index for which $y \in [x_i, x_{i+1}]$, or $i = 0$ if $y < x_1$

   (c) Modify v' by increasing $v'(i) \leftarrow v'(i) + h(y)\frac{x_{i+1}-y}{x_{i+1}-x_i}$, and $v'(i+1) \leftarrow v'(i+1) + h(y)\frac{y-x_i}{x_{i+1}-x_i}$

2. (a) Compute how much the second condition of the linear program is violated: let $w = \left(\sum_i x_i v'(i)\right) + \left(\sum_{i \geq k(A+B)} \frac{i}{k}\mathcal{F}(i)\right) - 1$

   (b) If $w < 0$, increase $v'_{A+B/2} \leftarrow v'_{A+B/2} + \frac{|w|}{A+B/2}$

   (c) Otherwise if $w > 0$, decrease v' arbitrarily (while still keeping it nonnegative) so as to satisfy the second condition of the linear program.

To show that with high probability the first condition is satisfied, we argue that since A was chosen so as to have relatively little mass in fingerprints $i \in [kA, k(A+B)]$, then with high probability, there could not have been much more than twice this mass in h in the probability interval $[A, A+B]$.

Additionally, with high probability the constructed v' after Step 1 has mass close to 1, and thus in Step 2, little mass is added to $v'_{A+B/2}$.

Lemma 24. *With probability at least $1 - 4e^{-k^{.1}/2}$, the constructed solution v' satisfies the first condition of the linear program; $\sum_{i \le A+B/2} x_i v'(i) \le 16k^{-.4a}$.*

Proof. First, we argue that with high probability the mass of h in the probability range $[A, A+B]$ is at most $10k^{-.4a}$. We will then argue that the total amount of mass added in Step 2 at probability $A+B/2$ is, with high probability, at most $6k^{-.4a}$.

Since the median of $Poi(\lambda) \in [\lambda, \lambda+1]$, and the tail bounds of Corollary 32 show that for $\lambda < k(A+B/2)$, $\Pr[X \leftarrow Poi(\lambda) > k(A+B)] \le e^{-k^{.1a}/2}$ and for $\lambda > k(A+B/2)$, $\Pr[X \leftarrow Poi(\lambda) < kA)] \le e^{-k^{.1a}/2}$, which we can crudely bound by $1/100$ for sufficiently large k, we conclude that given a $k' \leftarrow Poi(k)$-sample fingerprint, the expected mass in the fingerprint in the range $[kA, k(A+B)]$ is at least $\sum_{y \in [A,A+B]:h(y)\neq 0} yh(y)(1/2 - 2/100 - c)$, where c is the expected amount of mass in the fingerprint between λ and $\lambda+1$, which we can crudely bound by $1/100$, for sufficiently large k. Thus the expected mass in the fingerprint in the range $[kA, k(A+B)]$ is at least $(.47)\sum_{y \in [A,A+B]:h(y)\neq 0} yh(y)$.

Assume for the sake of contradiction that $\sum_{y \in [A,A+B]:h(y)\neq 0} yh(y) \ge 10k^{-.4a}$, and thus the expected number of samples in the fingerprint in the range $[kA, k(A+B)]$ is at least $(10 \cdot .47)k^{1-.4a}$. By elementary Chernoff bounds, with probability at most $e^{-k^{.2}/2}$ the number of samples in this range will be less than its expectation by at least $k^{.6}$. We will now apply Proposition 21, and note that the addition of $k^{.6}$ samples can alter the mass in the fingerprint above frequency H by at most $k^{.6}H/k$. For large k, since $(10 \cdot .47)k^{1-.4a} + 4k^{.6+a} \ge 4k^{1-.4a}$, by Proposition 21, given fingerprint $\mathcal{F}$,

$$\Pr[\sum_{i \in [kA, k(A+B)]} i\mathcal{F}(i) < 4k^{1-.4a}] \le e^{-k^{.2}/2} + e^{-k^{.1}/2} \le 2e^{-k^{.1}/2}.$$

Thus with probability at least $1 - 2e^{-k^{.1}/2}$, $\sum_{y \in [A,A+B]:h(y)\neq 0} yh(y) \le 10k^{-.4a}$.

We now bound the amount of extra mass that is added to v' at probability $A+B/2$ in Step 2. To do this, we first show that the expected amount of mass in $\mathcal{F}$ from frequencies below kA is, with high probability, not too much more than the amount of mass in the histogram h up to probability B. Then, we show that with high probability, the amount of mass in the fingerprint up to frequency kA is not much more than this expectation. Together with the bound that the fingerprint has relatively little mass in $[kA, k(A+B)]$, we conclude that with high probability, the fingerprint above frequency $k(A+B)$ has roughly the amount of mass it should, and thus Step 2 will only add at most a modest amount of mass.

Consider taking a $k' \leftarrow Poi(k)$-sample fingerprint. The probability of an element of probability $> B$ being sampled fewer than kA times, by Corollary 32 is at most $e^{-k^{.1a}/2}$, and thus the total mass in the expected fingerprint in frequencies below kA from elements of probability at least B is at most $ke^{-k^{.1a}/2} \le k^{.6}$, for sufficiently large k. By Chernoff bounds and Proposition 21,

$$\Pr[\sum_{i \le kA} i\mathcal{F}(i)/k \ge \sum_{y \le A+B:h(y)\neq 0} yh(y) + 4k^{-.4+a}] \le 2e^{-k^{.1}/2}.$$

Together with the bound on the amount of mass observed in $\mathcal{F}$ in the frequency range $[kA, k(A+B)]$, we conclude that with probability at least $1 - 2e^{-k^{.1}/2}$, the total amount of mass that must be added in Step 2 is at most $4k^{-.4+a} + 4k^{-.4a} \le 5k^{-.4a}$, from which the claim follows. □

We now argue that with high probability, the third condition of the linear program is satisfied by the constructed vector v'. The proof has three parts: first we show that the process of discretizing in Step 1, in which the true histogram h is modified so as to have support at the probabilities $\gamma, 2^2\gamma, 3^2\gamma, \ldots$ does not alter the fingerprint 'expectations' (where 'expectation' is in quotations to indicate that we mean the formal expression $\sum_i v'(i)poi(kx_i, j)$, which is well-defined even though v' has not been rounded into a true histogram). Next, we argue that Step 2, in which v' is modified so to have weight $1-w$ where $w = \sum_{i \geq k(A+B)} i\mathcal{F}(i)$, does not alter these 'expectations', because with high probability, prior to Step 2 v' has nearly the correct amount of mass. Thus we have established that the fingerprint expectations corresponding to v' are, with high probability, similar to the fingerprint expectations of the true histogram h. Finally, a union bound over Chernoff bounds shows that with high probability, the observed fingerprints $\mathcal{F}(i)$ will all be close to their expectations, and, thus, will be close to the 'expectations' of v' by the triangle inequality.

Lemma 25. *With probability at least* $1 - 10k^a e^{-k^{.1}/2}$, *every nonnegative integer* $j \leq k(A + B/4)$ *satisfies* $|\sum_i v'(i)poi(x_i k, j) - f(j)| \leq 4k^{.6+a}$.

Proof. Consider v' as it is at the end of Step 1: for each y such that $h(y) > 0$, we have "replaced" value $h(y)$ at probability y with the pair of values $h(y)\frac{x_{i+1}-y}{x_{i+1}-x_i}, h(y)\frac{y-x_i}{x_{i+1}-x_i}$ respectively at the corresponding discretized probabilities x_i and x_{i+1}, and we aim to bound

$$h(y)\left|\left(\frac{x_{i+1}-y}{x_{i+1}-x_i}poi(x_i k, j) + \frac{y-x_i}{x_{i+1}-x_i}poi(x_{i+1}k, j)\right) - poi(yk, j)\right| \qquad (1)$$

We note the basic calculus fact that for an arbitrary twice-differentiable function $g : \mathbb{R} \to \mathbb{R}$ and real numbers $a < y < b$, the linear interpolation $\frac{b-y}{b-a}g(a) + \frac{y-a}{b-a}g(b)$ approximates $g(y)$ to within $\frac{1}{8}(b-a)^2 \max_{z\in[a,b]} |g''(z)|$. Thus Equation 1 is bounded by $h(y)\frac{1}{8}(x_{i+1}-x_i)^2 \max_{z\in[x_i,x_{i+1}]} \left|\frac{d^2}{dz^2}poi(zk, j)\right|$. By Proposition 30 we see that $\left|\frac{d^2}{dz^2}poi(zk, j)\right| \leq 2k^2 \min\{1, \frac{1}{zk}\}$, yielding a bound on Equation 1 of

$$h(y)\frac{k^2}{4}(x_{i+1}-x_i)^2 \min\{1, \frac{1}{x_i k}\}.$$

Since $\gamma := k^{-3/2}$,

$$\begin{aligned} \max_{i:x_i\leq 1/k}\left(\frac{k^2}{4}(x_{i+1}-x_i)^2 \min\{1, \frac{1}{x_i k}\}\right) &= \max_{i:x_i\leq 1/k}\left(\frac{k^2}{4}(x_{i+1}-x_i)^2\right) \\ &\leq \frac{(2k^{1/4}+1)^2}{4k} \\ &\leq 4k^{-1/2}. \end{aligned}$$

Similarly,

$$\begin{aligned} \max_{i:x_i\geq 1/k}\left(\frac{k^2}{4}(x_{i+1}-x_i)^2 \min\{1, \frac{1}{x_i k}\}\right) &= \max_{i:x_i\geq 1/k}\left(\frac{k}{4x_i}(x_{i+1}-x_i)^2\right) \\ &\leq \frac{k^{-1/2}(2i+1)^2}{4i^2} \\ &\leq 4k^{-1/2}. \end{aligned}$$

Thus for any integer j,

$$|\sum_{y<A+B/2} h(y)poi(yk, j) - \sum_i v'(i)poi(x_i k, j)| \leq k(4k^{-1/2}) = 4k^{1/2}.$$

Next, we argue that the weight of v' before Step 2 is very close to what it should be, and thus the expectations are changed only slightly during Step 2. First note that if we take a $Poi(k)$-sample fingerprint $\mathcal{F}'$, the expected mass above frequency $k(A+B)$ is at most $\sum_{y>A+B/2:h(y)\neq 0} yh(y)+T$, where T is the expected mass contributed by the portion of h below $A+B/2$, which we can bound as $k\Pr[X \leftarrow poi(k(A+B/2)) > k(A+B)]$, which by the tail bound on Poisson distributions of Fact 31 and our choice of A, B, is at most $2ke^{-k^{.1a}/2}$, which we can bound by 1 for sufficiently large k. Thus by crudely applying the elementary multiplicative Chernoff bounds, the probability of seeing more than $k^{.5}$ elements of the support whose true probabilities are at most $A+B/2$ at least $k(A+B)$ times is at most $e^{-k^{.5}}$. By Proposition 21, $\Pr[\sum_{i>k(A+B)} i\cdot\mathcal{F}(i)/k > \sum_{y>A+B/2:h(y)\neq 0} yh(y)+k^{-.4+a}+k^{-.5+a}] \leq e^{-k^{.5}}+2e^{k^{.1}/2} < 3e^{k^{.1}/2}$.

Thus with at least this probability, after Step 1, v' will have mass at most $1-w+k^{-.4+a}+k^{-.5+a} \leq 1-w+2k^{-.4+a}$, where $w=\sum_{i\geq k(A+B)} i\mathcal{F}(i)$. If the mass is greater than $1-w$, then by arbitrarily decreasing the mass, the total change to the expected fingerprint will be at most $2kk^{-.4+a} = 2k^{.6+a}$, as desired. If the mass is less than $1-w$, then by adding the extra mass at probability $A+B/2$, this can effect the expectation of $\mathcal{F}(i)$, for $i \leq k(A+B/4$ by at most $k\cdot poi(k(A+B/2), \lfloor k(A+B/4)\rfloor) \leq 2ke^{-k^{.1}/2} \leq k^{.5}$, for sufficiently large k.

Combining the contributions to the error from discretizing h in Step 1 and from rearranging the mass in Step 2, we get that for each $j \leq k(A+B/4)$, with probability at least $1-3e^{-k^{.1}/2}$, $|\sum_{y:h(y)\neq 0} h(y)poi(ky,j) - \sum_i v'(i)poi(kx_i,j)| \leq 2k^{.6+a}$. To conclude, by Corollary 22, for each $j \leq k(A+B/4)$, $|\mathcal{F}(j)-E[\mathcal{F}(j)]| \leq 2k^{.6}$ with probability at least $1-2e^{-k^{.1}/2}$, and thus taking a union bound over the at most $k(A+B/4)$ fingerprint entries in this region, we get our claim. □

Proof of Proposition 23. By Lemmas 24 and 25, the constructed vector v' satisfies the first and third conditions of the linear program with probability at least $1-11k^a e^{-k^{.1}/2}$. By construction, it also satisfies the second condition, thus with at least this probability, it is in the feasible region of the linear program.

We now argue that the histogram $h^{v'}$ associated to the vector v' is close in relative-earthmover distance to the original histogram h. As above, let h' denote the histogram prior to the rounding step. Proposition 20 guarantees that $R(h', h^{v'}) \leq k^{-1/2}$, and thus $|R(h,h^{v'}) - R(h,h')| \leq k^{-1/2}$, by the triangle inequality. We show this distance is small by arguing that below probability $A+B/2$, the two histograms are very similar by construction, and that discretizing h in Step 1 of our construction of v' does not change the histogram much in relative-earthmover distance, and then in Step 2 of the construction, only a small amount of mass is adjusted, again resulting in a small change in earthmover distance. Above probability $A+B$, with high probability all the weight in h' at probability z/k can be sent to probability $(z \pm z^{.6})/k$ in h–essentially because each element of the fingerprint, with high probability, has true probability close to its observed probability.

For probabilities below $A+B/2$, our earthmoving scheme starts with h, discretizes (Step 1 of the construction of v'), then adjusts the mass if necessary (Step 2 of the construction). We now consider the cost of the discretization step. Since the support is at most $k^{1.1}$, at most $k^{-.2}$ mass in h can lie at probabilities below $k^{-1.3}$. For the mass at probability at least $k^{-1.3} = \gamma(k^{.1})^2$, the relative earthmover cost is at most $\log((k^{.1}+1)^2/(k^{.1})^2) \leq 2k^{-.1}$. For the mass between probabilities $k^{-3/2}$ and $k^{-1.3}$, the relative earthmover cost of discretization is at most $k^{-.4}\log(2^2/1) \leq 2k^{-.4}$. For the mass in h at probabilities below γ, note that there can be at most mass $m_z = k^{1.1}z$ at probability z. The cost of moving this mass to any higher probability is at most $m_z \log(1/z) \leq k^{1.1}z\log(1/z)$. Setting $z = k^{-c}$, and noting that for large k, $k^{1.1}\log(k) \leq k^{1.2}$, we have that this cost is at most $ck^{1.2-c} \leq k^{1.2-.9c}$, for $k > 40$, and since $c \geq 3/2$, Since the total mass in this region is at most 1, the cost of moving the mass in this region anywhere is at most $k^{1.2-.9(3/2)} \leq k^{-.1}$. Thus the total relative-earthmover cost of discretizing h to create v' is at most $4k^{-.1}$.

In Step 2 of the construction of v', as was shown in the proof of Lemma 25, with probability at least $1 - 11k^a e^{-k^{.1}/2}$, Step 2 decreases the mass in v' in probabilities less than $A + B/2$ by at most $2k^{-.4+a}$. The relative-earthmover cost of moving this mass to a probability less than $A + B$ is at most $2k^{-.4+a}|\log(\gamma/(A+B))| = 2k^{-.4+a}(3/2a)\log(2) \leq 2k^{-.4+2a}$, for sufficiently large k.

We now consider the earthmoving scheme for probabilities above $A + B$. Consider an element of the support of probability $k^{-1+\alpha}$, and let X be the random variable representing the number of occurrences of that element in a k-sample fingerprint; by Chernoff bounds $\Pr[|X - k^{\alpha}| > k^{.6\alpha}] \leq 2e^{-k^{.2\alpha}/4}$. Taking a union bound over all elements of the support that have probability at least $A + B/2$ shows that with probability at most $1 - k^{1.1}e^{-k^{.2a}/4} > 1 - e^{-k^{.1a}}$ (for large k) all the mass in the observed fingerprint with index above $k(A + B)$ can be accounted for by mass in h above $A + B/2$, and thus there is an earth-moving scheme for equating h and h' in which the entire mass of h' above probability $A + B$ is zeroed by an earthmover scheme where mass at probability k^{b-1} is moved within the range $[k^{b-1} - k^{.6b-1}, k^{b-1} + k^{.6b-1}]$, and thus the cost per unit mass is bounded by $\log\left(\frac{k^{b-1}}{k^{b-1}-k^{.6b-1}}\right) \leq 2k^{-.4b}$, and thus the disparity in h and h' above probability $A+B$ can be zeroed at relative earthmover cost at most $2 \cdot k^{-.4a}$.

Next, we account for the mass in h in the probability range $[A + B/2, A + B]$. In the proof of Lemma 24 we showed that with probability at least $1 - 2e^{-k^{.1}/2}$, the mass in the probability range $[A, A + B]$ is at most $10k^{-.4a}$, and this mass will, at worst, need to be moved from probability A to $A + B$, at constant cost per unit mass. Adding up these contributions we have $R(h, h') \leq 17k^{-\min(.1,.4a)}$, from which the claim follows. □

C All Solutions are Good Solutions

In this section, we now argue that with high probability over the set of samples, for *any* solution v to the linear program, the associated histogram h^v is close to the true histogram h. The previous section (Proposition 23) establishes that with high probability, the linear program will have at least one solution v whose corresponding histogram h^v is close to the true histogram h. Specifically, with probability at least $1 - e^{-k^{.04}}$, such a v exists with the property that $R(h, h^v) \leq 17k^{-.4a}$. To prove Theorem 3, we must now argue that for any pair of solutions v, w, their associated histograms satisfy $R(h^v, h^w) \leq O(\sqrt{\delta} \cdot \max\{1, |\log\delta|\})$, from which the theorem will follows by the triangle inequality. To prove this, in the following section, we construct an earthmoving scheme that leverages the fact that the fingerprint expectations of h^w and h^v are close.

C.1 The Bumps

Before describing the intuition behind our construction, we start by defining a very natural class of earthmoving schemes.

Definition 26. *For a given k, a β-*bump earthmoving scheme *is defined by a sequence of positive real numbers $\{c_i\}$, the* bump centers, *and a sequence of functions $\{f_i\} : (0,1] \to \mathbb{R}$ such that $\sum_{i=0}^{\infty} f_i(x) = 1$ for each x, and each function f_i may be expressed as a linear combination of Poisson functions, $f_i(x) = \sum_{j=0}^{\infty} a_{ij} poi(kx, j)$, such that $\sum_{j=0}^{\infty} |a_{ij}| \leq \beta$.*

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

Definition 27. *For given n, k, a bump earthmoving scheme (c, f) is* ϵ-good *if for any generalized histogram h, the relative earthmover distance between h and $(c, f)(h)$ is at most ϵ.*

Perhaps the most natural bump earthmoving scheme—which we will end up using a refinement of—is where $f_i(x) = poi(kx, i)$ and $c_i = \frac{i}{k}$, where for $i = 0$, c_i is chosen, say, as $\frac{1}{2k}$ to avoid a logarithm of 0 when evaluating relative earthmover distance. This is a valid earthmoving scheme since $\sum_{i=0}^{\infty} f_i(x) = 1$ for any x.

The motivation for this construction is the fact that, for any i, the amount of probability mass that ends up at c_i in $(c, f)(h)$ is exactly c_i times the expectation of the ith fingerprint in a $Poi(k)$-sample from h. Thus if we apply this earthmover scheme to two histograms derived from solutions to the linear program, their fingerprint expectations will closely match, and we would be left with a pair of histograms h^1, h^2 such that $R(h^1, h^2)$ is small.

The problem with this "Poisson bump" earthmoving scheme is that it has bad relative earthmover distance, particularly towards the origin. This is due to the fact that most of the mass that starts at a probability below $\frac{1}{k}$ will end up in the zeroth bump, no matter if it has probability nearly $\frac{1}{k}$, or the rather lower $\frac{1}{n}$. The situation gets significantly better for higher Poisson functions: most of the mass of $Poi(i)$ lies within relative distance $O(\frac{1}{\sqrt{i}})$ of i. We will therefore construct a scheme that uses Poisson functions $poi(kx, i)$ for $i \geq \log k$, but takes great care to construct "narrower" bumps below this region.

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

Lemma 28. *For $n > k$, letting δ be such that $n = \delta k \log k$, there exists an $O(\sqrt{\delta} \cdot \max\{1, |\log \delta|\})$-good $k^{0.3}$-bump earthmoving scheme*

In fact, we will construct a single scheme for all δ.

Definition 29. *The* Chebyshev earthmoving scheme *is defined in terms of k as follows. Let $s = \frac{1}{5} \log k$. For $i \geq s$, let $f_i(x) = poi(kx, i)$ and $c_i = \frac{i}{k}$.*

Define $g(y) = \sum_{j=-s}^{s-1} \cos(jy)$. Define $g'(y) = g(y) + g(y - \frac{\pi}{s})$ and, for $i \in \{0, \ldots, s-1\}$ define $g_i''(y) = g'(y - \frac{i\pi}{s}) + g'(y + \frac{(i+1)\pi}{s})$. Let $t_i(x)$ be the linear combination of Chebyshev polynomials so that $t_i(\cos(y)) = g_i''(y)$. We thus define the final s bumps to be $f_i(x) = \frac{1}{4s} t_i(1 - \frac{xk}{2s}) \sum_{j=0}^{s-1} poi(xk, j)$, for $i \in \{0, \ldots, s-1\}$. That is, $f_i(x)$ is related to $g_i''(y)$ by the coordinate transformation $x = \frac{2s}{k}(1 - \cos(y))$, and scaling by $\frac{1}{4s} \sum_{j=0}^{s-1} poi(xk, j)$. For these bumps, define $c_i = \frac{2s}{k}(1 - \cos(\frac{(i+1)\pi}{s}))$.

Proof of Lemma 28. We first show that the scheme of Definition 29 is in fact a $k^{0.3}$-bump earthmoving scheme.

Since $\sum_{i=0}^{s-1} g_i''(y) = \sum_{i=-s}^{s-1} g'(y - \frac{i\pi}{s})$, and since $g'(y)$ is a linear combination of cosines at integer frequencies j, shifted by all possible multiples of $\frac{j\pi}{s}$, we note that all but the $j = 0$ term will be canceled out; when $j = 0$, $cos(jy) = 1$, which term occurs twice in $g'(y)$, and hence $4s$ times in the sum, to yield a total of $\sum_{i=0}^{s-1} g_i''(y) = 4s$. Thus, correspondingly, $\sum_{i=0}^{s-1} t_i(x) = 4s$, and hence $\sum_{i=0}^{s-1} f_i(x) = \sum_{j=0}^{s-1} poi(xk, j)$. Since these s Poisson functions are exactly those missing from the Poisson bumps for $i \geq s$, we have $\sum_{i=0}^{\infty} f_i(x) = \sum_{i=0}^{\infty} poi(xk, i) = 1$, as desired.

We next check that each f_i may be expressed as $\sum_{j=0}^{\infty} a_{ij} poi(kx, j)$ for a_{ij} satisfying $\sum_{j=0}^{\infty} |a_{ij}| \leq k^{0.3}$. For $i \geq s$, we may trivially let $a_{ii} = 1$ and for all $j \neq i$, let $a_{ij} = 0$. For $i < s$, we first consider decomposing $g_i''(y)$ into a linear combination of $cos(jy)$, for $j \in \{0, \ldots, s\}$. Since $cos(-jy) = \cos(jy)$, g consists of 1 copy of $\cos(sy)$, two copies of $\cos(jy)$ for each j between 0 and s, and one copy of $\cos(0y)$; g' shifts some of these to introduce sine components, but these are canceled out in the formation of g_i'', which is a symmetric function for each i. Thus each g_i'' may be regarded as a linear combination $\sum_{j=0}^{s} \cos(yj) b_{ij}$ where the sth term has coefficient at most 4, and all the remaining terms have coefficients at most 8. Next, we note that 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 leave out the $\frac{1}{4s}$ term until later). 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. We note that since $\ell \leq i \leq s$, we have $s+\ell \leq 2s$, from which we see $\frac{1}{(2s)^\ell}\frac{(s+\ell)!}{s!} \leq 1$. We thus bound $\beta_{ij} \sum_{\ell=0}^{i} \binom{i}{\ell} \frac{1}{(2s)^\ell} \frac{1}{l+1} \frac{(s+\ell)!}{s!} \leq \beta_{ij} \sum_{\ell=0}^{i} \binom{i}{\ell} = \beta_{ij} 2^i$. Thus, in sum, we desire, for any $j \leq s$, to bound $\sum_{i=0}^{j} \beta_{ij} 2^i$, where β_{ij} are the coefficients of the jth Chebyshev polynomial. We note that since Chebyshev polynomials have coefficients whose signs repeat in the pattern $(+,0,-,0)$, we may evaluate this sum exactly as $|T_j(2\mathbf{i})|$, for $\mathbf{i} = \sqrt{-1}$. Explicitly, $|T_j(2\mathbf{i})| = \frac{1}{2}\left[(2-\sqrt{5})^j + (2+\sqrt{5})^j\right] \leq (2+\sqrt{5})^j$. To yield our final bound on the Poisson coefficients, recall that, before multiplying by $\frac{1}{4s}$, we have a factor at most 4 on the sth term, and factors at most 8 on each term for $j < s$, yielding that, for $s > 1$, Definition 29 is a $(2+\sqrt{5})^s$-bump earthmoving scheme.

We now turn to the main thrust of the argument, showing that the scheme is $O(\sqrt{\delta})$-good, where $n = \delta k \log k$, and $\delta \geq \frac{1}{\log k}$.

We first consider the cost of the portion of the scheme associated with bumps f_i for $i \geq s$, specifically, the relative earthmover cost of moving $poi(xk,i)$ mass from x to $\frac{i}{k}$, summed over $i \geq s$.

By definition of relative earthmover distance, the cost of moving mass from x to $\frac{i}{k}$ is $|\log \frac{xk}{i}|$, which, since $\log y \leq y-1$, we bound by $\frac{xk}{i} - 1$ when $i < xk$ and $\frac{i}{xk} - 1$ otherwise. We thus split the sum into two parts.

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

For $i \leq \lceil xk \rceil - 1$ we have, since $i \geq s$, that $poi(xk,i)(\frac{xk}{i} - 1) \leq poi(xk,i)((1+\frac{1}{s})\frac{xk}{i+1} - 1) = (1+\frac{1}{s}) poi(xk,i+1) - poi(xk,i)$. The $\frac{1}{s}$ term sums to at most $\frac{1}{s}$, and the rest telescopes to $poi(xk,\lceil xk \rceil) - poi(xk,s) = O(\frac{1}{\sqrt{s}})$.

Thus in total, f_i for $i \geq s$ contributes $O(\frac{1}{\sqrt{s}})$ to the relative earthmover cost, per unit of weight moved.

We now turn to the bumps $f_i(x)$ for $i < s$. The simplest case is when x is outside the region that corresponds to the cosine of a real number – that is, when $xk \geq 4s$. It is straightforward to show that $f_i(x)$ is very small in this region. We note the general expression for Chebyshev polynomials: $T_j(x) = \frac{1}{2}\left[(x-\sqrt{x^2-1})^j + (x+\sqrt{x^2-1})^j\right]$, whose magnitude we bound by $|2x|^j$. Further, since $2x \leq \frac{2}{e} e^x$, we bound this by $(\frac{2}{e})^j e^{|x|j}$, which we apply when $|x| > 1$. Recall the definition $f_i(x) = \frac{1}{4s} t_i(1-\frac{xk}{2s}) \sum_{j=0}^{s-1} poi(xk,j)$, where t_i is the polynomial defined so that $t_i(\cos(y)) = g_i''(y)$, that is, t_i is a linear combination of Chebyshev polynomials of degree at most s and with coefficients summing in magnitude to at most $8s$. Since $xk > s$, we may bound $\sum_{j=0}^{s-1} poi(xk,j) \leq s \cdot poi(xk,s)$. Further, since $z \leq e^{z-1}$ for all z, letting $z = \frac{x}{4s}$ yields $x \leq 4s \cdot e^{\frac{x}{4s}-1}$, from which we may bound $poi(xk,s) = \frac{(xk)^s e^{-xk}}{s!} \leq \frac{e^{-xk}}{s!}(4s \cdot e^{\frac{xk}{4s}-1})^s = \frac{4^s s^s}{e^s \cdot e^{3xk/4} s!} \leq 4^s e^{-3xk/4}$. We combine this with the above bound on the magnitude of Chebyshev polynomials, $T_j(z) \leq (\frac{2}{e})^j e^{|z|j} \leq (\frac{2}{e})^s e^{|z|s}$, where $z = (1-\frac{xk}{2s})$ yields $T_j(z) \leq (\frac{2}{e^2})^s e^{\frac{xk}{2}}$. Thus $f_i(x) \leq \frac{(8s)s}{4s} 4^s e^{-3xk/4} (\frac{2}{e^2})^s e^{\frac{xk}{2}} = 2s(\frac{8}{e^2})^s e^{-\frac{xk}{4}}$. Since $\frac{xk}{4} \geq s$ by definition of this case, f_i is exponentially small in both x and s; the total cost of this earthmoving scheme, per unit of weight above $\frac{4s}{k}$ is obtained by multiplying this by the logarithmic relative distance the weight has to move, and summing over the s values of $i < s$, yielding something that remains exponentially small, and thus trivially our goal of $O(\frac{1}{\sqrt{s}})$.

To bound the cost in the remaining case, when $xk \leq 4s$ and $i < s$, we work with the trigonometric functions g''_ℓ, instead of t_ℓ directly. For $y \in (0, \pi]$, consider the relative earthmover cost of, for each ℓ, moving $g''_\ell(y)$ mass from $\frac{2s}{k}(1 - \cos(y))$ to $\frac{2s}{k}(1 - \cos(\frac{\ell\pi}{s}))$, that is, $\sum_{\ell=1}^{s} |g''_\ell(y)(\log(1 - \cos(y)) - \log(1 - \cos(\frac{\ell\pi}{s}))|$. To simplify the analysis, we compare $\log(1 - \cos(y))$ with $2\log y$ when $y \in (0, \pi]$, noting that their derivatives respectively are $\frac{\sin(y)}{1-\cos(y)}$ and $\frac{2}{y}$, and we claim that the second expression is always greater. To compare the two expressions, cross-multiply and take the difference, to yield $y \sin y - 2 + 2\cos y$, which we show is always at most 0 by noting that it is 0 when $y = 0$ and has derivative $y \cos y - \sin y$, which is negative since $\cot y \leq \frac{1}{y}$. Thus we have that $|\log(1 - \cos(y)) - \log(1 - \cos(\frac{\ell\pi}{s}))| \leq 2|\log y - \log \frac{\ell\pi}{s}|$; we use this bound in all but the last step of the analysis.

We now turn to bounding the relative earthmover cost. We note that since $cos(y) = \Re(e^{\mathbf{i}y})$, for $\mathbf{i} = \sqrt{-1}$, we may express $g(y)$ as the real part of the sum of $2s$ terms of a geometric series to compute $g(y) = \Re\left(\frac{e^{\mathbf{i}y^{2s}}-1}{(e^{\mathbf{i}y}-1)e^{\mathbf{i}y^{s}}}\right) = \Re\left(\frac{e^{\mathbf{i}ys}-e^{-\mathbf{i}ys}}{e^{\mathbf{i}y}-1}\right)$. We have that $e^{\mathbf{i}ys} - e^{-\mathbf{i}ys} = 2\mathbf{i}\sin(ys)$, and $\Im(\frac{1}{e^{\mathbf{i}y}-1}) = -\frac{1}{2}\cot(\frac{y}{2})$, yielding that $g(y) = \sin(ys)\cot(\frac{y}{2})$; while this expression is undefined for $y = 0$, we note that $g(0) = 2s$.

Since $\sin(ys) = -\sin((y - \frac{\pi}{s})s)$, we have that $g'(y) = \sin(ys)\left[\cot(\frac{y}{2}) - \cot(\frac{y}{2} - \frac{\pi}{2s})\right]$. Since the cotangent is concave between 0 and $\frac{\pi}{2}$ and thus has decreasing derivative, we may bound $\cot(\frac{y}{2}) - \cot(\frac{y}{2} - \frac{\pi}{2s})$ in terms of the derivative of cotangent at the left endpoint, $\frac{y}{2} - \frac{\pi}{2s}$, which we compute as $\cot'(\frac{y}{2} - \frac{\pi}{2s}) = -\frac{1}{\cos^2(\frac{y}{2}-\frac{\pi}{2s})} = \frac{1}{\cos(y-\frac{\pi}{s})-1}$; thus $\cot(\frac{y}{2}) - \cot(\frac{y}{2} - \frac{\pi}{2s}) \leq \frac{\pi}{2s}\frac{1}{\cos(y-\frac{\pi}{s})-1}$. We may crudely bound this by noting that, for $y \in [0, \pi]$, $\cos(y) \leq 1 - 2(\frac{y}{\pi})^2$, yielding that, for $y \in (\frac{\pi}{s}, \pi]$, $|g'(y)| \leq \frac{\pi^3}{4s}\frac{1}{(y-\frac{\pi}{s})^2}$. By symmetry, for $y \in [-\pi + \frac{\pi}{s}, 0)$, we have $|g'(y)| = O(\frac{1}{sy^2})$. In general, since g' is the sum of $4s$ (shifted) cosine functions, $g'(y) \leq 4s$, which we will use instead of the previous bounds when $y \in [-\frac{\pi}{s}, 2\frac{\pi}{s}]$, combining the previous bounds to $O(\frac{1}{sy^2})$ otherwise.

We next bound $g''_i(y) = g'(y - \frac{i\pi}{s}) + g'(y + \frac{(i+1)\pi}{s})$, applying our bound on $g'(y)$ to yield $g''_i(y) = O(\frac{1}{s(y-\frac{i\pi}{s})^2})$ for $y \in [0, \frac{(i-1)\pi}{s}] \cup [\frac{(i+2)\pi}{s}, \pi]$, and $g''_i(y) \leq 8s$ for $y \in (\frac{(i-1)\pi}{s}, \frac{(i+2)\pi}{s})$.

We may now bound the relative earthmover distance. We ignore the term $\sum_{j=0}^{s-1} poi(xk, j)$ as it is always at most 1.

Case 1: $\frac{ys}{\pi} \geq 1$.

To bound $\frac{1}{4s}\sum_{i=0}^{s-1} |g''_i(y)(\log y - \log \frac{(i+1)\pi}{s})|$, we use each of the three bounds for g''_i just derived. In the middle region, when $y \in (\frac{(i-1)\pi}{s}, \frac{(i+2)\pi}{s})$, we note that $|(\log y - \log \frac{i\pi}{s})| = O(\frac{1}{sy})$, which we combine with the term $\frac{1}{4s}$ and the bound of $g''_i(y) \leq 8s$ in this region to yield $O(\frac{1}{sy})$.

For the high region, when $i \geq \frac{sy}{\pi} + 2$, we have bounded $g''_i(y) = O(\frac{1}{s(y-\frac{i\pi}{s})^2})$, which yields $\frac{1}{4s}\sum_{i \geq \frac{sy}{\pi}+2}^{s-1} |g''_i(y)(\log y - \log \frac{(i+1)\pi}{s})| = \frac{1}{4s}O(\sum_{i \geq \frac{sy}{\pi}+2}^{\infty} \frac{s}{|\frac{sy}{\pi}-i|^2}|\log \frac{sy}{\pi} - \log i|)$, which is easily seen to be $O(\frac{\log(1+\frac{sy}{\pi})}{sy})$, since $\int_{z+1}^{\infty} \frac{\log i - \log z}{(i-z)^2}\, di = \frac{(z+1)\log(z+1) - z\log z}{z} = O(\frac{\log(z+1)}{z})$. The same bounds hold for the region $i \leq \frac{sy}{\pi} - 1$.

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

To bound $\frac{1}{4s}\sum_{i=0}^{s-1} |g''_i(y)(\log \frac{ys}{\pi} - \log(i+1))|$, we note that $\log \frac{ys}{\pi} < 0$ and $\log(i+1) \geq 0$, and hence split the sum into two terms. To bound $\frac{1}{4s}\sum_{i=0}^{s-1} |g''_i(y)\log(i+1)|$, we note that for $i = 0$ the logarithm is 0, for $i = 1$, we bound $g''_i(y) \leq 8s$ to yield a constant bound on this term (when multiplied by $\frac{1}{4s}$), and when $i \geq 2$ we bound $g_i''(y) = O(\frac{1}{si^2})$, to yield $\frac{1}{4s}\sum_{i=0}^{s-1} |g_i''(y)\log(i+1)| = O(1) + \sum_{i=2}^{s-1} \frac{\log(i+1)}{i^2} = O(1)$.

Since $\sum_i |g_i''(y)| \leq 8s$, we bound the remaining term as $\frac{1}{4s}\sum_{i=0}^{s-1} |g_i''(y)\log\frac{ys}{\pi}| = O(\log\frac{ys}{\pi}$, yielding a total bound on the relative earthmover distance in this case of $O(1 + |\log\frac{ys}{\pi}|)$. Since for $z \in (0,1)$, $|\log z| < \frac{1}{z}$, we may bound this by $O(\frac{1}{ys})$.

Having concluded the case analysis, recall that we have been using the change of variables $x = \frac{2s}{k}(1-\cos(y))$. Since $1 - \cos(y) = O(y^2)$, we have $xk = O(sy^2)$. Thus the fact that the preceding case analysis yielded a bound of $\frac{\max\{1,\log sy\}}{sy}$ implies that we may express this as $O(\frac{\max\{1,\log sxk\}}{\sqrt{sxk}})$, our final bound on the per-unit cost of moving weight from location $x \leq \frac{4s}{k}$, under bumps f_i for $i < s$.

For a distribution with (generalized) histogram h, the cost of moving earth on this region, for bumps f_i where $i < s$ is thus $O(\sum_{x:h(x)\neq 0} h(x) \cdot x \cdot \frac{\max\{1,\log sxk\}}{\sqrt{sxk}})$. Because $\frac{\max\{1,\log z\}}{z}$ is a decreasing function, for $x \geq \frac{1}{n} = \frac{1}{5\delta sk}$, we have that $\frac{\max\{2,\log sxk\}}{\sqrt{sxk}}) \leq \max\{2, -\log 5\delta\}\sqrt{5\delta}$. Since $\sum_{x:h(x)\neq 0} h(x) \cdot x = 1$, we thus have a bound of $O(\max\{1, -\log\delta\}\sqrt{\delta})$ for this region. Otherwise, when $x \leq \frac{1}{n}$, since $x \cdot \frac{\max\{1,\log sxk\}}{\sqrt{sxk}}$ is an increasing function, it is maximized when $x = \frac{1}{n}$. Since the remaining term, $h(x)$, sums to at most n, by assumption, we thus have, as above, a bound for this region of $\frac{\max\{1,\log skx\}}{\sqrt{skx}} = O(\max\{1, -\log\delta\}\sqrt{\delta})$, which is the desired bound. As we have already bounded the relative earthmover cost for bumps $f_i, i \geq s$ at least this tightly, this concludes the proof. □

C.2 Proof of Theorem 3

We are now equipped to prove Theorem 3.

Proof of Theorem 3. By Proposition 23, with probability at least $1 - e^{-k^{.04}}$ the linear program has a solution v whose associated histogram h^v satisfies $R(h, h^v) \leq 17k^{-.4a}$. We now argue that for any pair of solutions v, w, their associated histograms satisfy $R(h^w, h^{w'}) \leq O(\sqrt{\delta} \cdot \max\{1, |\log\delta|\})$, from which the theorem will follows by the triangle inequality.

Consider a pair of histograms h^v, h^w derived from the solution to the linear program, and let $h^{v'}, h^{w'}$ represent the generalized histograms that yielded, respectively, h^v and h^w prior to the rounding step in Definition 19. By Proposition 20, $R(h^v, h^{v'}) \leq k^{-1/2}$, $R(h^w, h^{w'}) \leq k^{-1/2}$, and thus by the triangle inequality, it suffices to consider the pair $h^{v'}, h^{w'}$. We now exhibit an earth-moving scheme.

Consider applying the earth-moving scheme of Definition 29 to each of the generalized histograms $h^{v'}$ and $h^{w'}$, yielding the pair of generalized histograms h^1, h^2. By Lemma 28, $R(h^1, h^{v'}) \leq O(\max(1, |\log\delta|)\sqrt{\delta}$, and similarly for $R(h^2, h^{w'})$. Additionally, all of the probability mass in h^1, h^2 lies at the "bump" centers c_i. For each $c_i \leq A$, we now consider how much probability mass lies at c_i in each of the two histograms.

From Definition 29 we have

$$h^1(c_i) := \sum_{x:h^{v'}(x)\neq 0} x h^{v'}(x) f_i(x) = \sum_{x:h^{v'}(x)\neq 0} x h^{v'}(x) \sum_j a_{ij} poi(kx, j).$$

Since $\sum_j |a_{ij}| \leq \beta \leq k^{.3}$ by Lemma 28 we have:

$$\begin{aligned}
|h^1(c_i) - h^2(c_i)| &= |\left(\sum_j a_{ij} \sum_{x:h^{v'}(x)\neq 0} xh^{v'}(x)poi(kx,j)\right) - \left(\sum_j a_{ij} \sum_{x:h^{w'}(x)\neq 0} xh^{w'}(x)poi(kx,j)\right)| \\
&\leq \sum_j |a_{ij}| \left(|\sum_{x:h^{v'}(x)\neq 0} xh^{v'}(x)poi(kx,j) - \sum_{x:h^{w'}(x)\neq 0} xh^{w'}(x)poi(kx,j)|\right) \\
&\leq k^{.3} \cdot 9k^{.6+a} \leq k^{.9+a},
\end{aligned}$$

where the last line comes from the third condition of our linear program, which guarantees that $\sum_{x<A+B/2:h^{v'}(x)\neq 0} xh^1(x)poi(kx,j) \in [\mathcal{F}(j) - 4k^{.6+a}, \mathcal{F}(j) + 4k^{.6+a}]$, and the fact that the contribution from probability $x \geq A + B/2$ to the expectation of the jth fingerprint entry for $j < kA$ can be trivially crudely bounded by $k^{.5}$ (using the tail bounds in Fact 31, for example).

Thus the total residual mass for $c_i \leq A + B/4$ that can not be canceled is at most $\sum_{c_i \leq A+B/4} c_i \, |h^1(c_i) - h^2(c_i)| \leq 2k^a \cdot (k^{.9+a} 2k^{-1+a}) \leq 4k^{-.1+3a}$, where we used the fact that $c_{\log k} = \frac{\log k}{k}$, and for $i \geq \log k$, $c_i = i/k$, thus $|\{i : c_i \leq A + B/4\}| \leq k(A+B)$. Since $c_0 = \Theta(\frac{1}{k \log k})$, we can move this residual mass anywhere, at relative-earthmover cost at most $4k^{-.09+3a}$, for sufficiently large k.

We now consider the relative earthmover cost above probability $A + B/4$. This is easy–there is little probability mass in the probability range $[A+B/4, A+3B/4]$, and above probability $A+3B/4$ h^1 and h^2 are identical, except for the probability mass brought to region from probabilities below $A+B/2$ by our earth-moving scheme. Furthermore, throughout this range the "bumps" are Poisson bumps, namely 1-bumps. Rigorously, we bound the total probability mass that our earthmoving scheme brings into the range $[A + B/4, A + 3B/4]$ from probabilities below A and above B by, say, k^{-1} for sufficiently large k, by our tail bound for Poisson distributions given in Corollary 32. Thus the total mass in the range $[A+B/4, A+3/4B]$ in h^1 and h^2 is at most $k^{-.4a} + k^{-1}$, and thus we can move this mass anywhere within this range at cost at most $2k^{-.4a}$. To conclude, the entire disparity in h^1, h^2 above probability $A + 3B/4$ comes from the probability mass in $h^{v'}, h^{w'}$ below probability $A + B/2$, and thus, as above, we can bound this mass by k^{-1}, and thus this contributes a negligible amount to the total earthmover distance between h^1 and h^2 (namely at most $k^{-1}\log k < k^{-1/2}$ for large k).

Putting the pieces together, by the triangle inequality:

$$\begin{aligned}
R(h^v, h^w) &\leq R(h^{v'}, h^{w'}) + k^{-1/2} \\
&\leq R(h^1, h^2) + O(\sqrt{\delta} \cdot \max\{1, |\log \delta|\}) + k^{-1/2} \\
&\leq 5k^{-.09+3a} + O(\sqrt{\delta} \cdot \max\{1, |\log \delta|\}) \\
&= O(\sqrt{\delta} \cdot \max\{1, |\log \delta|\}).
\end{aligned}$$

□

D Properties of Poissons

In this section we collect the useful facts about the Poisson distribution, and the "Poisson functions," that are used throughout the paper.

D.1 Second Derivative of Poisson Functions

Proposition 30. *Letting $poi_{xx}(x,j)$ denote the second derivative of the jth Poisson function, for all $x > 0$, $j \geq 0$ we have $|poi_{xx}(x,j)| \leq \min\{2, \frac{2}{x}\}$.*

Proof. Since $poi(x,j) \triangleq \frac{x^j e^{-x}}{j!}$, we have $poi_{xx}(x,j) = (x^j - 2jx^{j-1} + j(j-1)x^{j-2})\frac{e^{-x}}{j!}$.
Case 1: $j = 0$ or 1. We have from the above expression that $poi_{xx}(x,0) = e^{-x}$, which is easily seen to be less than $\min\{2, \frac{2}{x}\}$. Similarly, for $j = 1$ we have $poi_{xx}(x,1) = (x-2)e^{-x}$, where, for $x \in (0,1)$ we have that $|(x-2)e^{-x}| \leq 2e^{-x} \leq 2$. For $x \geq 1$, we must show that $|(x-2)e^{-x}| \leq \frac{2}{x}$, or equivalently, $|\frac{1}{2}x^2 - x| \leq e^x$. Since $|\frac{1}{2}x^2 - x| \leq \frac{1}{2}x^2 + x$, and this last expression is just two terms from the power series of e^x, all of whose terms are positive, it is thus bounded by e^x as desired.
Case 2: $x < 1$ and $j \geq 2$.

In this case we must show $|poi_{xx}(x,j)| \leq 2$. For $j \geq 2$, we note that we may simplify the above expression for $poi_{xx}(x,j)$ to $((x-j)^2 - j)\frac{x^{j-2}e^{-x}}{j!}$. Noting that for $x \in (0,1)$ we have $x^{j-2} \leq 1$ and $e^{-x} < 1$, we may bound the absolute value of this last expression by $\frac{|(x-j)^2 - j|}{j!}$. Since $(x-j)^2 \geq 0$ and $-j \leq 0$, we may bound this expression as $\max\left\{\frac{(x-j)^2}{j!}, \frac{j}{j!}\right\}$; since we have $j \geq 2$ and $x \in (0,1)$, we note that $\frac{(x-j)^2}{j!} \leq \frac{j^2}{j!} \leq 2$, and $\frac{j}{j!} \leq 1$, as desired.
Case 3: $x \geq 1$ and $j \geq 2$.

We reexpress $|poi_{xx}(x,j)|$ as $\left|(1-\frac{j}{x})^2 - \frac{j}{x^2}\right| \cdot poi(x,j)$, which we may bound by $\max\{(1-\frac{j}{x})^2, \frac{j}{x^2}\} \cdot poi(x,j)$.

We consider the second term first. For $j > x+1$, consider the ratio of the expression $\frac{j}{x^2}poi(x,j)$ for consecutive values of j:

$$\frac{j}{j-1}\frac{x^j(j-1)!}{x^{j-1}j!} = \frac{x}{j-1}$$

and note that this is always at most 1. Thus $\frac{j}{x^2}poi(x,j)$ attains its maximum (over j) for $j \leq x+1$. We may thus bound $\frac{j}{x^2}poi(x,j)$ by taking $j \leq x+1$ and noting that, since $poi(x,j) \leq 1$ we have $\frac{j}{x^2}poi(x,j) \leq \frac{x+1}{x^2} \leq \frac{2}{x}$ as desired.

We now consider the first term, $(1-\frac{j}{x})^2 poi(x,j)$ and show that it attains its maximum for j in the interval $[x - 2\sqrt{x}, x + 2\sqrt{x} + 1]$. Consider the ratio of $(1-\frac{j}{x})^2 poi(x,j)$ to $(1 - \frac{j-1}{x})^2 poi(x, j-1)$:

$$\frac{(1-\frac{j}{x})^2}{(1-\frac{j-1}{x})^2}\frac{e^{-x}x^j(j-1)!}{e^{-x}x^{j-1}j!} = \left(\frac{x-j}{x-j+1}\right)^2 \frac{x}{j} \qquad (2)$$

We now show that this ratio is at most 1 for $j \geq x + 2\sqrt{x} + 1$, and at least 1 for $j \leq x - 2\sqrt{x} + 1$, thereby showing that $(1-\frac{j}{x})^2 poi(x,j)$ attains its maximum in the interval $j \in [x - 2\sqrt{x}, x + 2\sqrt{x} + 1]$. We note that both $\frac{x-j}{x-j+1}$ and $\frac{x}{j}$ are decreasing functions of j, outside the interval $[x, x+1]$, so it suffices to check the claim for $j = x + 2\sqrt{x} + 1$ and $j = x - 2\sqrt{x} + 1$. We have

$$\left(\frac{x - (x + 2\sqrt{x} + 1)}{x - (x + 2\sqrt{x} + 1) + 1}\right)^2 \frac{x}{x + 2\sqrt{x} + 1} = \frac{(2\sqrt{x}+1)^2}{(2\sqrt{x}+2)^2} \leq 1$$

and

$$\left(\frac{x - (x - 2\sqrt{x} + 1)}{x - (x - 2\sqrt{x} + 1) + 1}\right)^2 \frac{x}{x - 2\sqrt{x} + 1} = \frac{(2\sqrt{x}-1)^2}{(2\sqrt{x}-2)^2} \geq 1$$

Thus $(1-\frac{j}{x})^2 poi(x,j)$ attains its maximum for j in the interval $[x - 2\sqrt{x}, x + 2\sqrt{x} + 1]$. We note that on the sub-interval $[x - 2\sqrt{x}, x + 2\sqrt{x}]$, we have $(1-\frac{j}{x})^2 \leq (\frac{2\sqrt{x}}{x})^2 \leq \frac{4}{x}$, and that, for $x \geq 1$, $poi(x,j) \leq \frac{1}{e}$, implying that $(1-\frac{j}{x})^2 poi(x,j) \leq \frac{2}{x}$ as desired. Finally, for the remainder of the interval, we have, since $x \geq 1$ that $(1-\frac{j}{x})^2 \leq \frac{(2\sqrt{x}+1)^2}{x^2} \leq \frac{9}{x}$. On this sub-interval $j > x + 2\sqrt{x}$, and thus we have, since $x \geq 1$ and j is an integer, that $j \geq 4$. Since $poi(x,j)$ is maximized with respect to x when $x = j$, this maximum has value $\frac{j^j e^{-j}}{j!}$, which, by Stirling's approximation, is at most $\frac{1}{\sqrt{2\pi j}} < \frac{2}{9}$ (for $j \geq 4$). Combining these two bounds yields the desired bound of $\frac{2}{x}$. □

D.2 Tail Bounds for Poisson Distributions

Fact 31. *(From [15]) 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 $n \geq \lambda$,

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

Corollary 32. *For $\lambda > 30$, let $X \leftarrow Poi(\lambda)$,*

$$\Pr[|X - \lambda| > \lambda^{.6}] \leq e^{-\frac{\lambda^{.1}}{2}}.$$

Proof. We first show that the claim holds for $\Pr[X > \lambda + \lambda^{.6}]$. By Stirling's approximation, we have $n! \geq \sqrt{n}\frac{n^n}{e^n}$, and thus Fact 31 yields $\Pr[X \geq a] \leq \frac{\lambda^a e^{-\lambda} e^a}{\sqrt{a} a^a (1 - \lambda/a)}$. Letting $a := \lambda + \lambda^{.6}$ and simplifying slightly we get:

$$\begin{aligned}\Pr[X \geq \lambda + \lambda^{.6}] &\leq \frac{\sqrt{(\lambda + \lambda^{.6})} e^{\lambda^{.6} - (\lambda + \lambda^{.6}) \log(1 + \lambda^{-.4})}}{\lambda^{.6}} \\ &\leq e^{\lambda^{.6} - (\lambda + \lambda^{.6}) \log(1 + \lambda^{-.4})}.\end{aligned}$$

Considering the Taylor expansion of $\log(1 + x)$, and noting that for $x > 0$, $\log(1 + x) \geq x - x^2/2$, and thus $\log(1 + \lambda^{-.4}) \geq \lambda^{-.4} - \lambda^{-.8}/2$, we have:

$$\begin{aligned}\Pr[X \geq \lambda + \lambda^{.6}] &\leq e^{\lambda^{.6} - (\lambda + \lambda^{.6})(\lambda^{-.4} - \lambda^{-.8}/2)} \\ &\leq e^{(1/2)\lambda^{.2} - \lambda^{.2} + (1/2)\lambda^{-.2}} \\ &\leq e^{-\frac{\lambda^{.2} - \lambda^{-.2}}{2}}.\end{aligned}$$

For $\lambda > 30$, $\lambda^{.2} - \lambda^{-.2} > \lambda^{.1}$, yielding the claimed bound. We now apply an analogous argument to $\Pr[X < \lambda - \lambda^{.6}]$, again using Fact 31.

$$\begin{aligned}Pr[X < \lambda - \lambda^{.6}] &\leq \frac{\lambda^{.4} e^{-\lambda^{.6} - (\lambda - \lambda^{.6}) \log(1 - \lambda^{-.4})}}{\sqrt{(\lambda + \lambda^{.6})}} \\ &\leq e^{-\lambda^{.6} - (\lambda - \lambda^{.6}) \log(1 - \lambda^{-.4})}.\end{aligned}$$

Using the fact that for $x \in (0, .2]$, $\log(1 - x) > -x - \frac{5}{8}x^2$, we get

$$\begin{aligned}Pr[X < \lambda - \lambda^{.6}] &\leq e^{-\lambda^{.6} - (\lambda - \lambda^{.6})(-\lambda^{-.4} - \frac{5}{8}\lambda^{-.8})} \\ &\leq e^{-\frac{3\lambda^{.2} + 5\lambda^{-.2}}{8}}.\end{aligned}$$

Since $\frac{3\lambda^{.2} + 5\lambda^{-.2}}{8} > \lambda^{.1}/2$, the claimed bound holds.

□

For completeness, we state the elementary Chernoff bounds that we use throughout:

Fact 33. *Let $X_1, \ldots, X_n$ be independent $0, 1$ random variables, with $\Pr[X_i = 1] = p_i$. Let $X := \sum X_i$, and $\mu := \sum p_i$ Then:*

- $\Pr[X < (1 - \delta)\mu] \leq e^{-\mu\delta^2/2}$.
- *For $\delta \leq 2e - 1$, $\Pr[X > (1 + \delta)\mu] \leq e^{-\mu\delta^2/4}$, and for $\delta > 2e - 1$, $\Pr[X > (1 + \delta)\mu] \leq 2^{-\mu\delta}$.*