Data Amplification: Instance-Optimal Property Estimation

Yi Hao [1] **Alon Orlitsky** [1]

Abstract

The best-known and most commonly used technique for distribution-property estimation uses a plug-in estimator, with empirical frequency replacing the underlying distribution. We present novel linear-time-computable estimators that significantly "amplify" the effective amount of data available. For a large variety of distribution properties including four of the most popular ones and for every underlying distribution, they achieve the accuracy that the empirical-frequency plug-in estimators would attain using a logarithmic-factor more samples. Specifically, for Shannon entropy and a broad class of Lipschitz properties including the L_1 distance to a fixed distribution, the new estimators use n samples to achieve the accuracy attained by the empirical estimators with $n \log n$ samples. For support-size and coverage, the new estimators use n samples to achieve the performance of empirical frequency with sample size n times the logarithm of the property value. Significantly strengthening the traditional min-max formulation, these results hold not only for the worst distributions, but for each and every underlying distribution. Furthermore, the logarithmic amplification factors are optimal. Experiments on a wide variety of distributions show that the new estimators outperform the previous state-of-the-art estimators designed for each specific property.

1. Introduction

Recent years have seen significant interest in estimating properties of distributions over large domains (Valiant & Valiant, 2013; Jiao et al., 2015; 2018; Wu & Yang, 2016; Orlitsky et al., 2016; Acharya et al., 2017a; Hao et al., 2018; Wu & Yang, 2019; Hao & Orlitsky, 2019a;c; Charikar et al., 2019; Hao & Li, 2020). Chief among these properties are support size and coverage, Shannon entropy, and L_1 distance to a known distribution. The main achievement of these papers is essentially estimating properties of distributions with alphabet size k using just $k/\log k$ samples.

In practice however, the underlying distributions are often simple, and their properties can be accurately estimated with significantly fewer than $k/\log k$ samples. For example, if the distribution is concentrated on a small part of the domain, or is exponential, very few samples may suffice to estimate the property. To address this discrepancy, Hao et al. (2018) took the following competitive approach.

The best-known distribution property estimator is the *empirical estimator* that replaces the unknown underlying distribution by the observed empirical distribution. For example, with n samples, it estimates entropy by the formula $-\sum_i (N_i/n)\log(N_i/n)$ where N_i is the number of times symbol i appeared. Besides its simple and intuitive form, the empirical estimator is also consistent, stable, and universal. It is therefore the most commonly used property estimator for data-science applications.

The estimator in Hao et al. (2018) uses n samples and for any underlying distribution achieves the same performance that the empirical estimator would achieve with $n\sqrt{\log n}$ samples. It therefore provides an effective way to *amplify* the amount of data available by a factor of $\sqrt{\log n}$, regardless of the domain or structure of the underlying distribution.

In this paper we present novel estimators that increase the amplification factor for all sufficiently smooth properties including those mentioned above from $\sqrt{\log n}$ to the information-theoretic bound of $\log n$. Namely, for *every* distribution their expected estimation error with n samples is that of the empirical estimator with $n \log n$ samples and no further uniform amplification is possible.

It can further be shown (Valiant & Valiant, 2013; Jiao et al., 2015; Acharya et al., 2017a; Wu & Yang, 2019) that the empirical estimator estimates all of the aforementioned four properties with linearly many samples, hence the sample size required by the new estimators is always at most the $k/\log k$ guaranteed by the state-of-the-art estimators.

The current formulation has several additional advantages over previous approaches, which we illustrate as follows.

[1]Department of Electrical and Computer Engineering, University of California, San Diego, USA. Correspondence to: Yi Hao <yih179@eng.ucsd.edu>, Alon Orlitsky <alon@ucsd.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Fewer assumptions It eliminates the need for some commonly used assumptions. For example, support size cannot be estimated with any number of samples, as arbitrarily-many low-probabilities may be missed. Hence previous research (Acharya et al., 2017a; Wu & Yang, 2019) unrealistically assumed prior knowledge of the alphabet size k, and additionally that all positive probabilities exceed $1/k$. By contrast, the current formulation does not need these assumptions. Intuitively, if a symbol's probability is so small that it won't be detected even with $n \log n$ samples, we do not need to worry about it.

Refined bounds For some properties, our results are more refined than previously shown. For example, existing results estimate the support size to within $\pm\epsilon k$, rendering the estimates rather inaccurate when the true support size S is much smaller than k. By contrast, the new estimation errors are bounded by $\pm\epsilon S$, and are therefore accurate regardless of the support size. A similar improvement holds for the support coverage that we introduce below.

Graceful degradation For the previous results to work, one needs at least $k/\log k$ samples. With fewer samples, the estimators have no guarantees. By contrast, the guarantees of the new estimators work for any sample size n. If $n < k/\log k$, the performance may degrade, but will still track that of the empirical estimators with a factor $\log n$ more samples. See Theorem 1 for an example.

Instance optimality With the recent exception of Hao et al. (2018), all modern property-estimation research took a min-max-related approach, evaluating the estimation improvement based on the worst possible distribution for the property. In reality, practical distributions are rarely the worst possible and often quite simple, rendering min-max approach overly pessimistic, and its estimators, typically suboptimal in practice. In fact, for this very reason, practical distribution estimators do not use min-max based approaches (Gale & Sampson, 1995). By contrast, our *competitive*, or *instance-optimal*, approach provably ensures amplification for every underlying distribution, regardless of its complexity or support size.

In addition, the proposed estimators run in time near-linear in the sample size, and the constants involved are very small, attributes shared by some, though not all existing estimators.

Below, we formalize the foregoing discussion in definitions.

Let Δ_k denote the collection of discrete distributions over $[k] := \{1, \ldots, k\}$. A distribution *property* is a mapping $F : \Delta_k \to \mathbb{R}$. It is *additive* if it can be written as

$$F(p) := \sum_{i\in[k]} f_i(p_i),$$

where $f_i : [0, 1] \to \mathbb{R}$ are real functions. Many important distribution properties are additive:

Shannon entropy $H(p) := \sum_{i\in[k]} -p_i \log p_i$, is the principal measure of information (Cover & Thomas, 2012), and arises in a variety of machine-learning (Chow & Liu, 1968; Quinn et al., 2013; Bresler, 2015), neuroscience (Mainen & Sejnowski, 1995; Steveninck et al., 1997; Gerstner & Kistler, 2002), and other applications.

L_1 distance $D_q(p) := \sum_{i\in[k]} |p_i - q_i|$, where q is a given distribution, is one of the most basic and well-studied properties in the field of distribution property testing (Batu et al., 2000; Ron, 2010; Valiant & Valiant, 2016; Canonne, 2017).

Support size $S(p) := \sum_{i\in[k]} \mathbb{1}_{p_i>0}$, is a fundamental quantity for discrete distributions, and plays an important role in vocabulary size (McNeil, 1973; Efron & Thisted, 1976; Thisted & Efron, 1987) and population estimation (Good, 1953; Mao & Lindsay, 2007).

Support coverage $C(p) := \sum_{i\in[k]}(1 - (1 - p_i)^m)$, for a given m, represents the number of distinct elements we would expect to see in m independent samples, arises in many ecological (Chao, 1984; Chao & Lee, 1992; Colwell et al., 2012; Chao & Chiu, 2014), biological (Chao, 1984; Kroes et al., 1999), genomic (Ionita-Laza et al., 2009) as well as database (Haas et al., 1995) studies.

2. Prior and New Results

Given an additive property F and sample access to an unknown distribution p, we would like to estimate the value of $F(p)$ as accurately as possible. Let $[k]^n$ denote the collection of all length-n sequences, an estimator is a function $\hat{F} : [k]^n \to \mathbb{R}$ that maps a sample sequence $X^n \sim p$ to a property estimate $\hat{F}(X^n)$. We evaluate the performance of $\hat{F}$ in estimating $F(p)$ via its *mean absolute error* (MAE) [1],

$$L_{\hat{F}}(p, n) := \mathop{\mathbb{E}}_{X^n\sim p} \left|\hat{F}(X^n) - F(p)\right|.$$

Since we do not know p, the common approach is to consider the worst-case MAE of $\hat{F}$ over Δ_k,

$$L_{\hat{F}}(n) := \max_{p\in\Delta_k} L_{\hat{F}}(p, n).$$

The best-known and most commonly-used property estimator is the *empirical plug-in estimator*. Upon observing X^n, let N_i denote the number of times symbol $i \in [k]$ appears in X^n. The empirical estimator estimates $F(p)$ by

$$\hat{F}^{\mathrm{E}}(X^n) := \sum_{i\in[k]} f_i\left(\frac{N_i}{n}\right).$$

[1] As we aim to estimate only a single property value, the estimators in this paper all have negligible variances, e.g., $\mathcal{O}(1/n^{0.9})$. Hence the MAE is the same as MSE for our purpose, and we choose the former because it induces cleaner expressions.

Starting with Shannon entropy, it has been shown (Wu & Yang, 2016) that for $n \geq k$, the worst-case (max) MAE of the empirical estimator $\hat{H}^{\mathrm{E}}$ is

$$L_{\hat{H}^{\mathrm{E}}}(n) = \Theta\left(\frac{k}{n} + \frac{\log k}{\sqrt{n}}\right). \tag{1}$$

On the other hand, Jiao et al. (2015); Wu & Yang (2016); Acharya et al. (2017a); Hao & Orlitsky (2019a;c) showed that for $n \geq k/\log k$, more sophisticated estimators achieve the best min-max performance of

$$L(n) := \min_{\hat{F}} L_{\hat{F}}(n) = \Theta\left(\frac{k}{n\log n} + \frac{\log k}{\sqrt{n}}\right). \tag{2}$$

Hence up to constant factors, for the "worst" distributions, the MAE of these estimators with n samples equals that of the empirical estimator with $n\log n$ samples. A similar relation holds for the other three properties we consider.

However, the min-max formulation is pessimistic as it evaluates the estimator's performance for the worst distributions. In many practical applications, the underlying distribution is fairly simple and does not attain this worst-case loss, rather, a much smaller MAE can be achieved. Several recent works have therefore gone beyond worst-case analysis and designed algorithms that perform well for all distributions, not just those with the worst performance (Orlitsky & Suresh, 2015; Valiant & Valiant, 2016; Hao & Orlitsky, 2019b).

For property estimation, Hao et al. (2018) designed an estimator $\hat{F}^{\mathrm{A}}$ that for any underlying distribution uses n samples to achieve the performance of the $n\sqrt{\log n}$-sample empirical estimator, hence effectively multiplying the data size by a $\sqrt{\log n}$ *amplification factor*.

Lemma 1 (Hao et al. (2018))**.** *For every property F in a large class including the aforementioned four properties, there is an absolute constant c_F such that for all distributions p, all $\varepsilon \leq 1$, and all $n \geq 1$,*

$$L_{\hat{F}^{\mathrm{A}}}(p,n) - L_{\hat{F}^{\mathrm{E}}}\left(p, \varepsilon n\sqrt{\log n}\right) \leq c_F \cdot \varepsilon.$$

In this work, we fully strengthen the above result and establish the limits of data amplification for all sufficiently smooth additive properties including four of the most important ones, and all that are appropriately Lipschitz.

Using Shannon entropy as an example, we achieve a $\log n$ amplification factor. Equations (1) and (2) imply that the improvement over the empirical estimator cannot always exceed $\mathcal{O}(\log n)$, hence up to an absolute constant, this amplification factor is information-theoretically optimal. Similar optimality arguments hold for our results on the other three properties (e.g., see Table 1 in Acharya et al. (2017a)).

Specifically, we derive efficient estimators $\hat{H}$, $\hat{D}$, $\hat{S}$, $\hat{C}$, and $\hat{F}$ for the Shannon entropy, L_1 distance, support size, support coverage, and a broad class of additive properties which we refer to as *Lipschitz properties*. These estimators run in *near-linear time*, take a single parameter ε, and given samples $X^n \sim p$, amplify the data as described below.

For brevity, henceforth we shall write $a \wedge b$ and $a \lesssim b$ instead of $\min\{a,b\}$ and $a = \mathcal{O}(b)$, respectively, and abbreviate support size $S(p)$ by S_p and coverage $C(p)$ by C_p.

The following five theorems hold for all $\varepsilon \leq 1$, all distributions p, and all $n \geq 1$.

Theorem 1 (Shannon entropy)**.**

$$L_{\hat{H}}(p,n) - L_{\hat{H}^{\mathrm{E}}}(p, \varepsilon n\log n) \lesssim \varepsilon \wedge \left(\frac{S_p}{n} + \frac{1}{n^{0.49}}\right).$$

Note that the estimator requires no knowledge of S_p or k. When $\varepsilon = 1$, the estimator amplifies the data by a factor of $\log n$. As ε decreases, the amplification factor decreases, and so does the extra additive inaccuracy. One can also set ε to be a vanishing function of n, e.g., $\varepsilon = 1/\log\log n$.

This result may be interpreted as follows. For distributions with large support sizes such that the min-max estimators provide no or only very weak guarantees, our estimator with n samples always tracks the performance of the $n\log n$-sample empirical estimator. On the other hand, for distributions with relatively small support sizes, our estimator achieves a near-optimal $\mathcal{O}(S_p/n)$-error rate.

Similarly, for L_1 distance to a fixed distribution q,

Theorem 2 (L_1 distance)**.** *For any q, we can construct an estimator $\hat{D}_q$ for D_q such that*

$$L_{\hat{D}_q}(p,n) - L_{\hat{D}_q^{\mathrm{E}}}\left(p, \varepsilon^2 n\log n\right) \lesssim \varepsilon \wedge \left(\sqrt{\frac{S_p}{n}} + \frac{1}{n^{0.49}}\right).$$

Besides having an interpretation similar to that of Theorem 1, the above result shows that for each q and each p, we can use just n samples to achieve the performance of the $n\log n$-sample empirical estimator. More generally, for any additive property $F(p) := \sum_{i\in[k]} f_i(p_i)$ that satisfies the simple condition: f_i is $\mathcal{O}(1)$-Lipschitz, for all i,

Theorem 3 (General additive properties)**.** *Given F, we can construct an estimator $\hat{F}$ such that*

$$L_{\hat{F}}(p,n) - L_{\hat{F}^{\mathrm{E}}}\left(p, \varepsilon^2 n\log n\right) \lesssim \varepsilon \wedge \left(\sqrt{\frac{S_p}{n}} + \frac{1}{n^{0.49}}\right).$$

The results in Kamath et al. (2015) show that no plug-in estimators provide those theoretical guarantees presented in Theorem 2 and 3. Henceforth, we refer to the above collection of distribution properties as the class of *Lipschitz properties*. The L_1 distance D_q, for any q, is in this class.

Lipschitz properties are essentially bounded by absolute constants and Shannon entropy grows at most logarithmically in the support size, and we were able to approximate all with just an additive error. Support size and support coverage

can grow linearly in k and m, and can be approximated only multiplicatively. We therefore evaluate the estimator's normalized performance, regarding the property value.

Note that for both properties, the amplification factor is logarithmic in the property value, which can be arbitrarily larger than the sample size n.

The following two theorems hold for $\varepsilon \le e^{-2}$,

Theorem 4 (Support size)**.**

$$\frac{1}{S_p}\left(L_{\hat{S}}(p,n) - L_{\hat{S}^{\mathrm{E}}}\left(p, n\cdot\frac{\log S_p}{\log^2\varepsilon}\right)\right) \lesssim \varepsilon + S_p^{\frac{1}{|\log\varepsilon|}-\frac{1}{2}}.$$

To make the slack term vanish, one can simply set ε to be a vanishing function of n (or S_p), e.g., $\varepsilon = 1/\log n$. Note that in this case, the slack term modifies the multiplicative error in estimating S_p by only $o(1)$, which is negligible in most applications. Similarly, for support coverage,

Theorem 5 (Support coverage)**.**

$$\frac{1}{C_p}\left(L_{\hat{C}}(p,n) - L_{\hat{C}^{\mathrm{E}}}\left(p, n\cdot\frac{\log C_p}{\log^2\varepsilon}\right)\right) \lesssim \varepsilon + C_p^{\frac{1}{|\log\varepsilon|}-\frac{1}{2}}.$$

The next section presents implications of these results.

3. Implications

Data amplification Numerous modern scientific applications, such as those emerging in social networks and genomics, deal with properties of distributions whose support size S_p is equal to or even larger than the sample size n.

In this data-sparse regime, the estimation error of the empirical estimator often decays at a slow rate, e.g., $1/\log^c n$ for some $c \in (0,1)$, hence the proposed estimators yield a much more accurate estimate, paralleling that of the empirical with $n\log n$ samples. For applications where $n \ge 25,000$ and regardless of the distribution structure, our approach significantly amplifies the number of samples by at least a factor of 10, known by practitioners as an "order of magnitude".

As for the data-rich regime where $n \gg S_p$, our method essentially recovers the the standard $\sqrt{S_p/n}$ rate of maximum likelihood methods in general, without knowing S_p.

Instance optimality With just n samples, our method emulates the performance of the $n\log n$-sample empirical estimator for *every distribution instance*. The method hence possesses the vital ability of strengthening all MAE guarantees of the empirical estimator by a logarithmic factor, which is optimal in many settings.

The significance of such "instance optimality" arises from 1) empirical estimators are often simple and easy to analyze; 2) there is a rich literature on their estimation attributes, e.g., Bustamante (2017) and the references therein; 3) empirical estimators are the best-known and most-used.

Consequently, we can work on a simple problem, analyzing the performance of the empirical estimator, and immediately strengthen the result we obtain by a logarithmic-factor using the theorems in this paper. In many cases, the strengthened results are challenging to establish via other statistical methods. We present two examples below.

Entropy Consider entropy estimation over Δ_k. As Equation 2 shows, the min-max MAE is known for $n \ge k/\log k$, and essentially becomes a constant when n gets close to the $k/\log k$ lower bound. Nevertheless, over an alphabet of size k, the value of entropy can go up to $\log k$. Hence, it is still possible to get meaningful estimation results in the $n = o(k/\log k)$ large-alphabet regime.

We follow the above strategy to solidify the statement. First, for empirical estimator $\hat{H}^{\mathrm{E}}$, Paninski (2003) [see Proposition 1] provides a short argument showing that its worst-case MAE, for all n and k, satisfies

$$L_{\hat{H}^{\mathrm{E}}}(n) \le \log\left(1+\frac{k-1}{n}\right) + \frac{\log n}{\sqrt{n}}.$$

Consolidating this inequality with Theorem 1 then implies

Corollary 1. *In the $n = o(k/\log k)$ large-alphabet regime, the min-max MAE of estimating Shannon entropy satisfies*

$$L(n) \le (1+o(1))\log\left(1+\frac{k-1}{n\log n}\right).$$

Lipschitz Property The same type of arguments apply to any Lipschitz property F. Again, we begin with characterizing the performance of the empirical estimator $\hat{F}^{\mathrm{E}}$. By Lemma 3 and the Cauchy-Schwarz inequality, the bias of $\hat{F}^{\mathrm{E}}$ is at most $\mathcal{O}(\sqrt{k/n})$. By the Efron-Stein inequality, the standard deviation is no more than $\mathcal{O}(1/\sqrt{n})$.

It then follows by Theorem 3 that: $\hat{F}$ estimates F over Δ_k to an MAE of ε with $\mathcal{O}(k/(\varepsilon^3\log k))$ samples. Note that 1) this yields the first estimator for Lipschitz properties with optimal sample dependence on k; 2) after a draft of this paper became available online, Hao & Orlitsky (2019c) improved the sample dependence on ε to the optimal ε^2.

4. Estimator Construction and Analysis

For clarity, we focus on the proof of Theorem 1 about entropy estimation, and explain only necessary modifications for similar arguments to go through for other properties. We begin by relating the empirical entropy estimator to the "Bernstein polynomial" of function $-x\log x$.

Notation For a sampling parameter n and accuracy $\varepsilon \le 1$, define the *amplification factor* as $a := \varepsilon\log n$. Without loss of generality, assume that $\varepsilon \ge 1/\log n$ and hence $a \ge 1$. For simplicity, write $h(x) := -x\log x$, $m := na$, $\tau_n := \Theta(\log n/n)$ and $d_n := \Theta(\log n)$, where the asymptotic notations hide only properly chosen absolute constants.

4.1. Bernstein Polynomial

Drawing i.i.d. samples Y^m from any distribution p, the expected value of the empirical estimator for entropy is

$$\mathbb{E}[\hat{H}^{\mathrm{E}}(Y^m)] = \sum_{i\in[k]} \mathop{\mathbb{E}}_{M_i\sim\mathrm{bin}(m,p_i)} \left[h\left(\frac{M_i}{m}\right)\right].$$

Note that for any function f, $m\in\mathbb{N}$, and $x\in[0,1]$, the degree-m *Bernstein polynomial* of f is

$$B_m(f,x) := \sum_{j=0}^{m} f\left(\frac{j}{m}\right)\binom{m}{j}x^j(1-x)^{m-j}.$$

Therefore, we can express the expectation of the empirical entropy estimator as

$$\mathop{\mathbb{E}}_{Y^m\sim p}[\hat{H}^{\mathrm{E}}(Y^m)] = \sum_{i\in[k]} B_m(h,p_i).$$

As modifying a sample changes the value of $\hat{H}^{\mathrm{E}}(Y^m)$ by at most $2\log m/m$, the Efron-Stein inequality bounds its variance by $2\log^2 m/m$, which is negligible in magnitude. Hence, for our purpose, we focus on finding a good approximation of each $B_m(h,p_i)$.

4.2. Estimator Construction and Computation

In the subsequent sections, given i.i.d. samples $X^n\sim p$, we construct our estimator as follows.

Substitute n by $2n$ for simplicity. According to Section 4.4, we first split the samples into two halves, X_1^n and X_{n+1}^{2n}, and respectively denote by N_i and N_i' the empirical counts of each symbol $i\in[k]$ in them.

Then, we follow Dobrushin (1958) to classify the symbols into two categories and decompose the sum

$$\mathop{\mathbb{E}}_{Y^m\sim p}[\hat{H}^{\mathrm{E}}(Y^m)] = \sum_{i\in[k]} B_m(h,p_i)$$

into two parts by thresholding the empirical counts N_i' at level $1/\varepsilon$. The first part, $\mathcal{V}_L := \sum_{i\in[k]} B_m(h,p_i)\mathbb{1}_{N_i'>1/\varepsilon}$, corresponds to the contribution of symbols with potentially large probabilities. Illustrated in Section 4.3, this quantity is well approximated by the *large-probability estimator*

$$\hat{\mathcal{V}}_L := \sum_{i\in[k]} h\left(\frac{N_i}{n}\right)\cdot\mathbb{1}_{N_i'>\frac{1}{\varepsilon}},$$

to an MAE of $2(\varepsilon\wedge S_p/n)$. As for the small-probability part,

$$\mathcal{V}_S := \sum_{i\in[k]} B_m(h,p_i)\cdot\mathbb{1}_{N_i'\le\frac{1}{\varepsilon}},$$

we follow the arguments in Section 4.4 and 4.5 to learn each summand adaptively (to the magnitude of the probability) and compute the summation.

Concretely, recall $\tau_n = \Theta(\log n/n)$ and $d_n = \Theta(\log n)$. For a given function and domain, the polynomial achieving the minimal maximum deviation from the function over the domain is the *min-max polynomial*. Then, denote by

$$\tilde{h}_m(x) := \sum_{t=0}^{d_n} b_t x^t$$

the degree-d_n min-max polynomial of $B_m'(h,p_i)$ over interval $I_n := [0,\tau_n]$. The *small-probability estimator* is

$$\hat{\mathcal{V}}_S := \sum_{i\in[k]}\left(\sum_{t=1}^{d+1}\frac{b_{t-1}}{t}\cdot\frac{N_i^{\underline{t}}}{n^t}\right)\cdot\mathbb{1}_{N_i\lesssim\log n}\cdot\mathbb{1}_{N_i'\le\frac{1}{\varepsilon}},$$

where for each symbol i, the term in the parentheses is an unbiased estimator for $\tilde{H}_m(p_i) := \int_0^{p_i}\tilde{h}_m(s)ds$. Next, we illustrate the technique and intuition behind the construction.

Differential smoothing The construction of $\hat{\mathcal{V}}_S$ presents a generic method for designing a polynomial $\tilde{G}$ that closely approximates a given differentiable function G *with pointwise error bounds*.

More precisely, for a fixed interval $I := [0,\tau]$ and degree bound $d\in\mathbb{N}$, we want to find a polynomial $\tilde{G}$ of degree at most d, satisfying

$$\max_{x\in I}|\tilde{G}(x)-G(x)|\le c\cdot x,$$

for a number $c\ge 0$ that is *as small as possible*.

We propose a novel method, *differential smoothing*, that addresses this approximation problem and operates as follows.

1. Compute $G'(x)$ and write $g := G'$.
2. Approximate g by its min-max polynomial $\tilde{g}$ over I.
3. Let c be the min-max approximation error in Step 2.
4. Compute $\tilde{G}(x) := \int_0^x \tilde{g}(t)dt$.

By the triangle inequality for integrals, the resulting c and $\tilde{G}$ satisfy the desired inequality. Besides, Step 2 and 3 can be jointly performed using the well-known Remez algorithm (Pachón & Trefethen, 2009; Trefethen, 2013).

Turning back to our estimator $\hat{V}_S$, by the reasoning in Section 4.6 and 4.7, the min-max polynomial $\tilde{h}_m(x)$ approximates $B_m'(h,x)$ to within $\mathcal{O}(\varepsilon)$ over I_n. Hence, applying the method of differential smoothing yields

$$|B_m(h,x)-\tilde{H}_m(x)|\lesssim\varepsilon\cdot x.$$

Further relating this inequality to the expectation of the empirical entropy estimator implies

$$\left|\mathop{\mathbb{E}}_{Y^m\sim p}[\hat{H}^{\mathrm{E}}(Y^m)] - \sum_{i\in[k]}\tilde{H}_m(p_i)\right|\lesssim\sum_{i\in[k]}\varepsilon\cdot p_i = \varepsilon.$$

In Section 6.1 of the supplementary, we prove that the absolute bias is also at most $\mathcal{O}(S_p/n)$, which requires some additional work. Finally, Section 7.1 bounds the mean absolute deviation of the estimator by $\mathcal{O}(1/n^{0.49})$.

Consequently, we approximate $H(p)$ by

$$\hat{H} := \hat{\mathcal{V}}_L + \hat{\mathcal{V}}_S.$$

Computational complexity The dominant computation step is finding the min-max polynomial of $B'_m(h,x)$, for which we utilize the well-known Remez algorithm (Pachón & Trefethen, 2009; Trefethen, 2013). In Section 9 of the supplementary, we shall argue that the algorithm takes only $\tilde{O}(n)$ time to well approximate $B'_m(h,x)$.

4.3. Large-Probability Estimator

Following the previous arguments, we say that $i \in [k]$ is a *large-probability symbol* if $N'_i > 1/\varepsilon$. To the expectation of the m-sample empirical estimator, these symbols contribute

$$\mathcal{V}_L = \sum_{i\in[k]} B_m(h,p_i)\cdot \mathbb{1}_{N'_i > \frac{1}{\varepsilon}}.$$

We estimate $\mathcal{V}_L$ by respectively reweighing the empirical estimator associated with the first-half samples:

$$\hat{\mathcal{V}}_L = \sum_{i\in[k]} h\left(\frac{N_i}{n}\right)\cdot \mathbb{1}_{N'_i > \frac{1}{\varepsilon}}.$$

To bound the estimation bias, we leverage the next lemma, stating that the Bernstein polynomial of h closely approximates the function over $[0,1]$.

Lemma 2. *For any $t \in \mathbb{Z}^+$ and $x \in [0,1]$,*

$$-\frac{1-x}{t} \le B_t(h,x) - h(x) \le 0.$$

The number of symbols satisfying $N'_i > 1/\varepsilon$ is at most $n\varepsilon$. Together with the lemma and triangle inequality, this yields

$$|\mathbb{E}[\mathcal{V}_L] - \mathbb{E}[\hat{\mathcal{V}}_L]| \le \sum_{i\in[k]} \left(\frac{1+a}{m}\right)(1-p_i)\mathbb{E}\left[\mathbb{1}_{N'_i > \frac{1}{\varepsilon}}\right] \le 2\varepsilon.$$

Furthermore, the number of such symbols is also at most S_p, implying an alternative upper bound of $2S_p/n$.

For Shannon entropy, we note that adding $1/(2n)$ to the empirical estimate $h(N_i/n)$ may reduce its bias. This particular method, known as the "Miller-Mallow estimator", appears in Miller (1955) and eliminates the first-order term of $B_n(h,x) - h(x)$. Applying the method will introduce extra complications in the analysis, and hence for entropy and general non-differentiable properties, we employ the original empirical estimator. On the other hand, substituting the Miller-Mallow estimate into our algorithm in Theorem 1 retains its theoretical guarantee.

For Lipschitz properties, the rich literature on Bernstein operators presents us with the following bound.

Lemma 3 ((Bustamante, 2017) Proposition 4.9)**.** *For any $t \in \mathbb{Z}^+$, $x \in [0,1]$, and c-Lipschitz function f,*

$$|B_t(f,x) - f(x)| \le c\cdot\sqrt{\frac{x(1-x)}{t}}.$$

Combined with the Cauchy-Schwarz inequality, the lemma shows that the estimation bias of the respective $\hat{\mathcal{V}}_L$ admits

$$|\mathbb{E}[\mathcal{V}_L] - \mathbb{E}[\hat{\mathcal{V}}_L]| \le 2\left(\varepsilon \wedge \sqrt{\frac{S_p}{n}}\right).$$

This completes the bias analysis of the large-probability estimator, while Section 6.2 in the supplementary provides additional technical details. For the variance analysis, see Section 7.2. The following three sections proceed to construct the small-probability estimator and introduce fundamental results from polynomial approximation theory.

4.4. Choice of Parameters and Sample Splitting

Section 4.1 calls for estimating $B_m(h,x)$. Applying the method of *differential smoothing* in Section 4.2, we first choose some domain $I = [0,\tau]$ and degree d, and estimate $B'_m(h,x)$ by its min-max polynomial $\tilde{h}_m(x) = \sum_{t=0}^d b_t x^t$ over I. Then, we approximate $B_m(h,x)$ by

$$\tilde{H}_m(x) = \int_0^x \tilde{h}_m(t)dt = \sum_{t=0}^d \frac{b_t}{t+1}x^{t+1}.$$

To estimate $\tilde{H}_m(x)$, note that given a binomial variable $X \sim \mathrm{bin}(n,x)$, an unbiased estimator for x^t is $X^{\underline{t}}/n^{\underline{t}}$, where $t \in \mathbb{N}$ and $A^{\underline{B}}$ denotes *the B-th order falling factorial* of A. Hence, we employ

$$\hat{H}_m(X) := \sum_{t=1}^{d+1} \frac{b_{t-1}}{t}\cdot\frac{X^{\underline{t}}}{n^{\underline{t}}},$$

an *unbiased* estimator for $\tilde{H}_m(x)$ that corresponds to the parenthetical component in estimator $\hat{\mathcal{V}}_S$'s expression. Next, we illustrate the intuitions behind our choice of τ and d.

For any $X \sim \mathrm{bin}(n,x)$, the variance of $\hat{H}_m(X)$ generally gets larger as the degree d increases. On the other hand, a higher-degree polynomial is able to achieve a lower approximation error. To balance this bias-variance trade-off, we want to reduce both the interval length, τ, and the polynomial degree, d, while maintaining the approximation power.

As in Section 4.2, we set parameter $\tau = \tau_n = \Theta(\log n/n)$ since below this threshold, sample statistics are insufficient for inferring the relative magnitudes of the underlying probabilities with high confidence. Regarding the degree parameter $\tau = \tau_n = \Theta(\log n)$, below the $\log n$ threshold, the approximation $\tilde{H}_m$ loses the $\varepsilon\cdot x$ guarantee; in contrast, above the threshold, the final estimator may no longer possess a vanishing variance. For more details, see derivations in Section 7.1 and Appendix A of the supplementary.

One thing that follows the construction of $\tilde{H}_m$ and $\hat{H}_m$ is how to apply these approximations to only probabilities of order τ_n. This issue arises from the fact that we observe symbol counts, not ranges of the actual probability values. It is straightforward to deal with such uncertainty by inferring the magnitudes of unknowns leveraging the counting statistics concentration.

For concentration, binomial random variables are sums of independent indicator variables and possess Gaussian-type tail bounds. To avoid introducing additional statistical dependency, we 1) split the sample sequence into two halves of equal length; 2) denote respectively the empirical counts of each symbol i in the first and second halves by N_i and N_i' (where we slightly abused the notation); 3) classify each $i \in [k]$ as a large- or small- probability symbol by thresholding the count N_i' at $1/\varepsilon$. The supplementary material presents relevant details in Section 5 and 6.2.

In the literature, the above procedure is often referred to as *sample splitting*. This idea of classifying the symbols in the alphabet into two categories dates back to Dobrushin (1958), and has been applied to estimate a variety of specific distribution properties in the past decade (Acharya et al., 2014; Jiao et al., 2015; Wu & Yang, 2016; Hao et al., 2018). Recently, Hao & Orlitsky (2019c) generalize this idea to estimate general properties by partitioning the unit interval into $\tilde{\Theta}(\sqrt{n})$ pieces; Hao & Orlitsky (2019b) apply the method to derive state-of-the-art distribution estimators.

Sample splitting and additiveness of the property enable us to estimate the contributions from the large and small probabilities separately. The rest sections assume this separation and address the small-probability approximation error.

4.5. Min-Max Polynomial

Polynomials have extensive applications to statistical inference, ranging from approximating the norms of Gaussian parameters (Cai & Low, 2011) to learning structured distributions (Chan et al., 2014; Acharya et al., 2017b; Hao & Orlitsky, 2019b) to estimating properties of distributions (Jiao et al., 2015; Orlitsky et al., 2016; Wu & Yang, 2016; Hao et al., 2018; Hao & Orlitsky, 2019c).

As illustrated in Section 4.2 and 4.4, we aim to find a polynomial $\tilde{h}_m(x)$ of degree $d_n = \Theta(\log n)$ that satisfies the pointwise bound $|B_m'(h,x) - \tilde{h}_m(x)| \lesssim \varepsilon$ over $I_n = [0, \tau_n]$.

The task naturally calls for a polynomial achieving the minimal maximum deviation from $B_m'(h,x)$, commonly known as the respective *min-max polynomial*. Moreover, direct computation shows that $B_m'(h,x)$ is the order-$(m-1)$ Bernstein polynomial of another function:

$$B_m'(h,x) = B_{m-1}(h_m, x),$$

where function h_m is defined as

$$h_m(y) := -\log\frac{m-1}{m} + (m-1)\left(h\left(y + \frac{1}{m-1}\right) - h(y)\right).$$

Hence, our objective reduces to bounding the error of min-max polynomial approximations of $B_{m-1}(h_m, x)$ over I_n. As one could expect, the analysis gets more involved since 1) $B_{m-1}(h_m, x)$ is a high-degree polynomial with transcendental coefficients; 2) in general, there are no closed-form formulas for the min-max polynomials of a real function.

Though sophisticated in its form, function $B_{m-1}(h_m, x)$ is continuous and relatively smooth, as hinted by Lemma 2. This simple observation serves as the starting point for our subsequent analysis. In the next section, we dive into approximation theory and present fundamental connections between the smoothness of a function (characterized by specific quantities) and its min-max polynomial approximation error over a given interval. The desired result then follows by a sequence of inequalities and simplifications that enable us to gauge the smoothness of $B_{m-1}(h_m, x)$.

For the proof of the derivative identity on h_m and a more straightforward argument leading to a weaker result, see Section 4 and 5 of the supplementary.

4.6. Moduli of Smoothness

In this section, we introduce some notable results in approximation theory (Ditzian & Totik, 2012) that are crucial for simplifying the problem. Denote $\varphi(x) := \sqrt{x(1-x)}$. For any function $f : [0,1] \to \mathbb{R}$, *the first- and second- order Ditzian-Totik moduli of smoothness* quantities of f are

$$w_\varphi^1(f,t) := \sup\left\{|f(u) - f(v)| : 0 \le u, v \le 1, \; |u-v| \le t \cdot \varphi\left(\frac{u+v}{2}\right)\right\},$$

and

$$w_\varphi^2(f,t) := \sup\left\{\left|f(u) + f(v) - 2f\left(\frac{u+v}{2}\right)\right| : 0 \le u, v \le 1, |u-v| \le 2t \cdot \varphi\left(\frac{u+v}{2}\right)\right\},$$

respectively. Let $\boldsymbol{P}_d$ denote the collection of polynomials with real coefficients and degree at most d. For any $d \in \mathbb{Z}^+$, interval $I \subset \mathbb{R}$, and function $f : I \to \mathbb{R}$, denote by

$$E_d[f, I] := \min_{\tilde{f} \in \boldsymbol{P}_d} \max_{x \in I} |f(x) - \tilde{f}(x)|$$

the *best approximation error* of the degree-d min-max polynomial of f over I. For a bounded domain I, we can always shift and rescale f to make it a real function over $[0,1]$. Hence, without loss of generality, it suffices to consider and analyze $E_d[f] := E_d[f, [0,1]]$.

The connection between the best polynomial-approximation error $E_d[f]$ of a continuous function f and the second-order Ditzian-Totik moduli of smoothness $w_\varphi^2(f,t)$ is established in the following lemma (Ditzian & Totik, 2012).

Lemma 4. *There are absolute constants C_1 and C_2 such that for any continuous function f over $[0,1]$ and $d > 2$,*

$$E_d[f] \le C_1 w_\varphi^2(f, d^{-1}),$$

and

$$\frac{1}{d^2}\sum_{t=0}^{d}(t+1)E_t[f] \ge C_2 w_\varphi^2(f, d^{-1}).$$

The above lemma shows that the second-order smoothness quantity $w_\varphi^2(f,\cdot)$ essentially characterizes $E.[f]$, and thus transforms the problem of showing

$$|\tilde{h}_m(x) - B_{m-1}(h_m, x)| \lesssim \varepsilon,\ \forall x \in I_n,$$

to that of establishing

$$w_\varphi^2(B_{m-1}(h_m, \tau_n \cdot x), d_n^{-1}) \lesssim \varepsilon,$$

where $\tau_n = \Theta(\log n/n)$ and $d_n = \Theta(\log n)$ by definition.

4.7. Simplification via Poissonization

The last block in our analysis is Poissonization, which helps decompose and simplify the function to approximate. For any $y \in [0,\infty]$, define two functions:

$$f_1(y) := \mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(y)}[h(X)] = -e^{-y}\sum_{j=1}^{\infty}\frac{y^j}{j!}(j\log j)$$

and

$$f_2(y) := \mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(y)}[h(X+1)].$$

Let $z(x) := (m-1)x$ for simplicity. The following lemma, appearing in Appendix A.1 of the supplementary relates $B_{m-1}(h_m, x)$ to these functions and base function $h(x)$.

Lemma 5. *For any $m \in \mathbb{Z}^+$ and $x \in [0, \log^4 m/m]$,*

$$h_m(x) - B_{m-1}(h_m, x) = [h(z(x)+1) - f_2(z(x))] - [h(z(x)) - f_1(z(x))] + \tilde{\mathcal{O}}\left(\frac{1}{m}\right).$$

In particular, the above equation holds for any sufficiently large n and $x \in I_n = [0, \tau_n]$. Since $1/m = 1/(na-1) \le \min\{1/\log n, S_p/n\}$, the last term on the right-hand side is negligible. These results, together with the function-wise triangle inequality on w_φ^2, further reduce the last inequality in Section 4.6 to bounds in the form of

$$w_\varphi^2(g(x), d_n^{-1}) \lesssim \varepsilon,$$

for function $g(x)$ being $h_m(\tau_n \cdot x)$, $h(z(x))$, $h(z(x)+1)$, $f_1(z(x))$, and $f_2(z(x))$, respectively.

We prove these bounds in Appendix A.2 and A.3 of the supplementary. In Appendix B, a similar yet more involved argument extends the result to all Lipschitz properties. One reason for the extra complication is the absence of concrete expression, as we impose only the Lipschitz condition.

While these proofs are technical, a critical insight is that the optimization problems induced by computing w_φ^2 for the above choices of g are all convex. Consequently, it suffices to consider only the boundary cases of parameters.

5. Experiments

We demonstrate the efficacy of the proposed estimators by comparing their performance to two state-of-the-art estimators (Wu & Yang, 2016; 2019), and empirical estimators with logarithmic larger sample sizes. Due to method similarity, we present only the results for entropy and support size. Additional estimators for both properties were compared in Orlitsky et al. (2016); Wu & Yang (2016; 2019); Hao et al. (2018); Hao & Orlitsky (2019a) and found to perform similarly to or worse than the estimators we tested, hence we exclude them here. For each property, we considered nine natural-synthetic distributions, shown in Figure 1 and 2.

Settings We experimented with nine distributions having *support size* $S = 10{,}000$: uniform distribution; a two-steps distribution with probability values $0.5S^{-1}$ and $1.5S^{-1}$; Zipf distribution with power $1/2$; Zipf distribution with power 1; binomial distribution with success probability 0.3; geometric distribution with success probability 0.9; Poisson distribution with mean $0.3S$; a distribution drawn from Dirichlet prior with parameter 1; a distribution drawn from Dirichlet prior with parameter 1/2.

The geometric, Poisson, and Zipf distributions were truncated at S and re-normalized. The horizontal axis shows the number of samples, n, ranging from $S^{0.2}$ to S. Each experiment was repeated 100 times and the reported results, shown on the vertical axis, reflect their mean values and standard deviations. Specifically, the real property value is drawn as a dashed black line, and the other estimators are color/shape coded, with the solid line displaying their mean estimate, and the shaded area corresponding to one standard deviation.

We compared the estimators' performance with n samples to that of two other recent estimators as well as the empirical estimator with n, $n\sqrt{\log A}$, and $n\log A$ samples, where for Shannon entropy, $A = n$ and for support size, $A = S_p$, the actual distribution support size (which is S). We chose the parameter $\varepsilon = 1$. The graphs denote our proposed estimator by Proposed, $\hat{F}^{\mathrm{E}}$ with n samples by Empirical, $\hat{F}^{\mathrm{E}}$ with $n\sqrt{\log A}$ samples by Empirical+, $\hat{F}^{\mathrm{E}}$ with $n\log A$ samples by Empirical++, the entropy and support-size estimators in Wu & Yang (2016) and Wu & Yang (2019) by WY.

Results As Theorem 1 and 4 would imply and the experiments confirmed, for both properties, the proposed estimators with n samples achieved the accuracy as the empirical estimators with at least $n\log n$ samples for entropy and $n\log S_p$ samples for support size. In particular, for entropy, the proposed estimator with n samples performed significantly better than the $n\log n$-sample empirical estimator, for all tested distributions and all values of sample size n. For both properties, the proposed estimators outperformed the state-of-the-art estimators in terms of accuracy and stability regarding distribution structures.

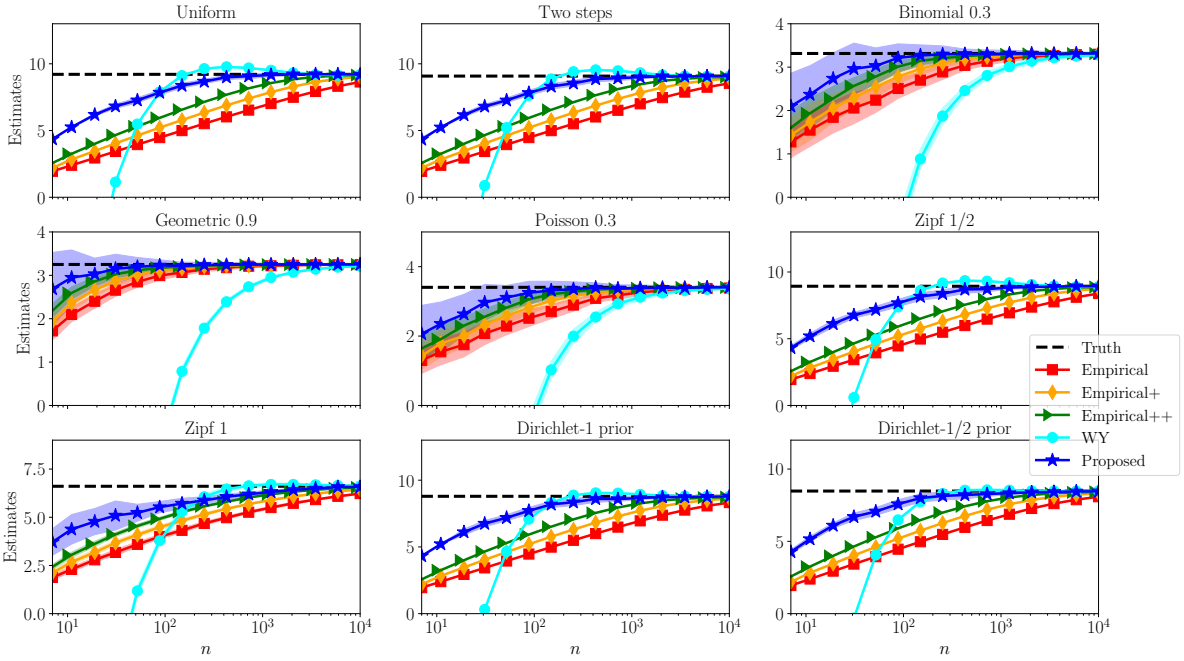

Figure 1. Shannon entropy estimation. For clarity, the horizontal axis is in logarithmic scale. The WY curve is flipped vertically around Truth for all the curves to have similar trends. Besides the samples, the WY estimator takes as input an upper bound of the support size, which is set to be the actual support size in the experiments. The vertical axis shows only nonnegative values.

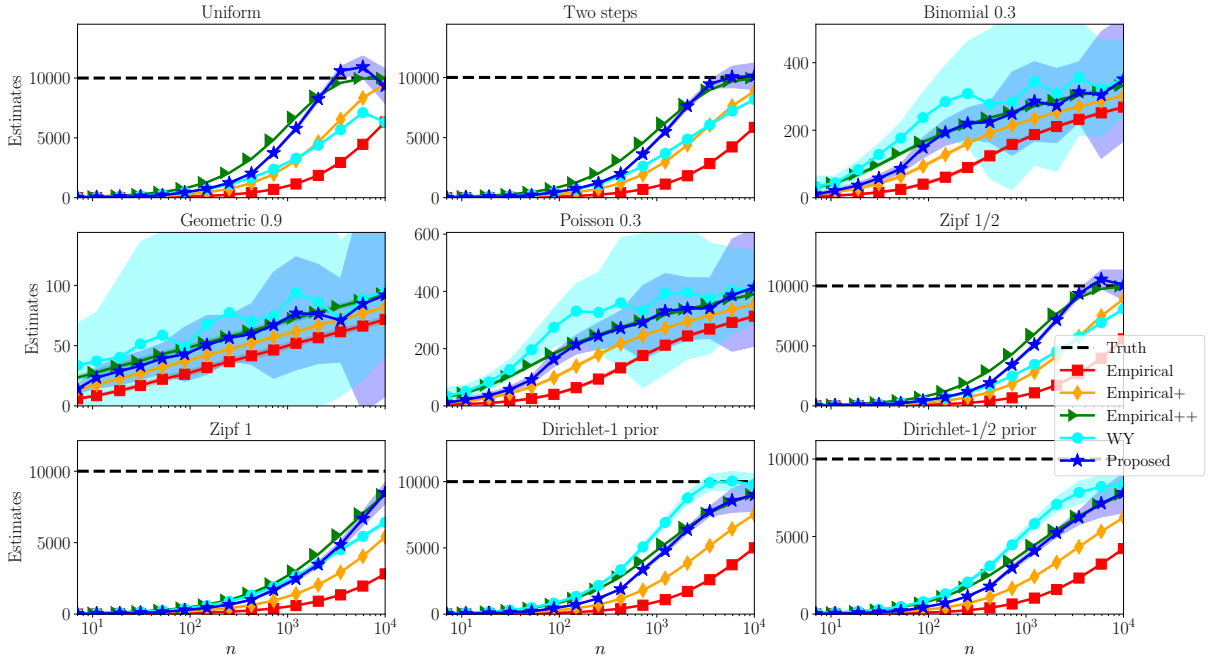

Figure 2. Support size estimation. For clarity, the horizontal axis is in logarithmic scale. Besides the samples, the WY estimator takes as input a lower bound of the smallest positive probability $p^+_{\min}$, which is set to be $\max\{1/(10S), 4p^+_{\min}\}$ in the experiments. Here, $1/(10S)$ is used to avoid division by zero in numerical computation, and factor 4 represents a reasonable uncertainty about $p^+_{\min}$. For several distributions, such as uniform and geometric, knowing $p^+_{\min}$ yields the full knowledge of the entire probability multiset. Finally, while estimator WY's bias is slightly lower on a few distributions, the corresponding standard deviation is too high to be acceptable.

Acknowledgements

We thank the reviewers for helpful comments, and are grateful to the National Science Foundation (NSF) for supporting this work through grants CIF-1564355 and CIF-1619448.

References

Acharya, J., Orlitsky, A., Suresh, A. T., and Tyagi, H. The complexity of estimating Rényi entropy. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1855–1869. SIAM, 2014.

Acharya, J., Das, H., Orlitsky, A., and Suresh, A. T. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *International Conference on Machine Learning*, pp. 11–21, 2017a.

Acharya, J., Diakonikolas, I., Li, J., and Schmidt, L. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1278–1289. SIAM, 2017b.

Batu, T., Fortnow, L., Rubinfeld, R., Smith, W. D., and White, P. Testing that distributions are close. In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, pp. 259–269. IEEE, 2000.

Bresler, G. Efficiently learning Ising models on arbitrary graphs. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pp. 771–782, 2015.

Bustamante, J. *Bernstein operators and their properties*. Springer, 2017.

Cai, T. T. and Low, M. G. Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics*, 39(2): 1012–1041, 2011.

Canonne, C. L. A survey on distribution testing. *Your Data is Big. But is it Blue.*, 2017.

Chan, S.-O., Diakonikolas, I., Servedio, R. A., and Sun, X. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pp. 604–613, 2014.

Chao, A. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, pp. 265–270, 1984.

Chao, A. and Chiu, C.-H. Species richness: Estimation and comparison. *Wiley StatsRef: Statistics Reference Online*, pp. 1–26, 2014.

Chao, A. and Lee, S.-M. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, 87(417):210–217, 1992.

Charikar, M., Shiragur, K., and Sidford, A. Efficient profile maximum likelihood for universal symmetric property estimation. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pp. 780–791, 2019.

Chow, C. and Liu, C. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

Colwell, R. K., Chao, A., Gotelli, N. J., Lin, S.-Y., Mao, C. X., Chazdon, R. L., and Longino, J. T. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, 5(1):3–21, 2012.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.

Ditzian, Z. and Totik, V. *Moduli of smoothness*, volume 9. Springer Science & Business Media, 2012.

Dobrushin, R. L. A simplified method of experimentally evaluating the entropy of a stationary sequence. *Theory of Probability & Its Applications*, 3(4):428–430, 1958.

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

Gale, W. A. and Sampson, G. Good-Turing frequency estimation without tears. *Journal of Quantitative Linguistics*, 2(3):217–237, 1995.

Gerstner, W. and Kistler, W. M. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge University Press, 2002.

Good, I. J. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4): 237–264, 1953.

Haas, P. J., Naughton, J. F., Seshadri, S., and Stokes, L. Sampling-based estimation of the number of distinct values of an attribute. In *Proceedings of the 21st International Conference on Very Large Data Bases*, pp. 311–322. Morgan Kaufmann Publishers Inc., 1995.

Hao, Y. and Li, P. Bessel smoothing and multi-distribution property estimation. In *Conference on Learning Theory*, pp. 1817–1876, 2020.

Hao, Y. and Orlitsky, A. The broad optimality of profile maximum likelihood. In *Advances in Neural Information Processing Systems*, pp. 10991–11003, 2019a.

Hao, Y. and Orlitsky, A. Doubly-competitive distribution estimation. In *International Conference on Machine Learning*, pp. 2614–2623, 2019b.

Hao, Y. and Orlitsky, A. Unified sample-optimal property estimation in near-linear time. In *Advances in Neural Information Processing Systems*, pp. 11104–11114, 2019c.

Hao, Y., Orlitsky, A., Suresh, A. T., and Wu, Y. Data amplification: A unified and competitive approach to property estimation. In *Advances in Neural Information Processing Systems*, pp. 8834–8843, 2018.

Ionita-Laza, I., Lange, C., and Laird, N. M. Estimating the number of unseen variants in the human genome. *Proceedings of the National Academy of Sciences*, 106 (13):5008–5013, 2009.

Jiao, J., Venkat, K., Han, Y., and Weissman, T. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

Jiao, J., Han, Y., and Weissman, T. Minimax estimation of the L_1 distance. *IEEE Transactions on Information Theory*, 64(10):6672–6706, 2018.

Kamath, S., Orlitsky, A., Pichapati, D., and Suresh, A. T. On learning distributions from their samples. In *Conference on Learning Theory*, pp. 1066–1100, 2015.

Kroes, I., Lepp, P. W., and Relman, D. A. Bacterial diversity within the human subgingival crevice. *Proceedings of the National Academy of Sciences*, 96(25):14547–14552, 1999.

Mainen, Z. F. and Sejnowski, T. J. Reliability of spike timing in neocortical neurons. *Science*, 268(5216):1503–1506, 1995.

Mao, C. X. and Lindsay, B. G. Estimating the number of classes. *The Annals of Statistics*, pp. 917–930, 2007.

McNeil, D. R. Estimating an author's vocabulary. *Journal of the American Statistical Association*, 68(341):92–96, 1973.

Miller, G. Note on the bias of information estimates. *Information Theory in Psychology: Problems and Methods*, 1955.

Orlitsky, A. and Suresh, A. T. Competitive distribution estimation: Why is Good-Turing good. In *Advances in Neural Information Processing Systems*, pp. 2143–2151, 2015.

Orlitsky, A., Suresh, A. T., and Wu, Y. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016.

Pachón, R. and Trefethen, L. N. Barycentric-Remez algorithms for best polynomial approximation in the Chebfun system. *BIT Numerical Mathematics*, 49(4):721, 2009.

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

Quinn, C. J., Kiyavash, N., and Coleman, T. P. Efficient methods to compute optimal tree approximations of directed information graphs. *IEEE Transactions on Signal Processing*, 61(12):3173–3182, 2013.

Ron, D. Algorithmic and analysis techniques in property testing. *Foundations and Trends® in Theoretical Computer Science*, 5(2):73–205, 2010.

Steveninck, R., Lewen, G. D., Strong, S. P., Koberle, R., and Bialek, W. Reproducibility and variability in neural spike trains. *Science*, 275(5307):1805–1808, 1997.

Thisted, R. and Efron, B. Did Shakespeare write a newly-discovered poem? *Biometrika*, 74(3):445–455, 1987.

Trefethen, L. N. *Approximation theory and approximation practice*, volume 128. SIAM, 2013.

Valiant, G. and Valiant, P. Estimating the unseen: Improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems*, pp. 2157–2165, 2013.

Valiant, G. and Valiant, P. Instance optimal learning of discrete distributions. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pp. 142–155, 2016.

Wu, Y. and Yang, P. Minimax rates of entropy estimation on large alphabets via best polynomial approximation. *IEEE Transactions on Information Theory*, 62(6):3702–3720, 2016.

Wu, Y. and Yang, P. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *The Annals of Statistics*, 47(2):857–883, 2019.