Data Amplification: Instance-Optimal Property Estimation

Yi Hao
Dept. of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093
yih179@ucsd.edu

Alon Orlitsky
Dept. of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093
alon@ucsd.edu

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.

Contents

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

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.

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, our 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 shouldn't 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 structural 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.

1 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 samples $X^n \sim p$ to a property estimate $\hat{F}(X^n)$. We evaluate the performance of $\hat{F}$ in estimating F by *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

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

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 sample size $n \geq k/\log k$, 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. *For every F in a large property 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 the subsequent sections, 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 (see Table 1 in Acharya et al. (2017a)).

[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.

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 $\varepsilon \le 1$, all distributions p, and all $n \ge 1$.

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)**.**

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

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)**.**

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

The next section presents implications of these results.

2 Implications and Outline

2.1 Theoretical 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 $n \geq 25,000$ 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, for example, 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 $n \geq k/\log k$, 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 all k, satisfies the elegant bound

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

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, which can be as large as $\log k$, satisfies*

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

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 18 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.

2.2 Supplementary Outline

For notational convenience, let $h(p) := -p\log p$ for entropy, $\ell_q(p) := |p-q|-q$ for L_1 distance, $s(p) := \mathbb{1}_{p>0}$ for support size, and $c(p) := 1-(1-p)^m$ for support coverage. Below, we provide an outline of the remaining contents and a high-level overview of our techniques.

In the main body, we focus on Shannon entropy and prove a weaker version of Theorem 1.

Theorem 6. *For all $\varepsilon \le 1$ and all distributions p, the estimator $\hat{H}$ described in Section 5 satisfies*

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

The proof of Theorem 6 in the rest of the paper is organized as follows. In Section 3, we present a few useful concentration inequalities for Poisson and binomial random variables. In Section 4, we relate the n-sample empirical estimator's bias to the degree-n Bernstein polynomial $B_n(h,x)$ via $B_n(h,p_i) = \mathbb{E}[h(N_i/n)]$. In Section 4.1, we show that the absolute difference between the *derivative* of $B_n(h,x)$ and a simple function $h_n(x)$ is at most 1, uniformly for all $x \le 1-(n-1)^{-1}$.

Let $a := \varepsilon\log n$ be an amplification parameter. In Section 4.2, we approximate $h_{na}(x)$ by a degree-$\Theta(\log n)$ polynomial $\tilde{h}_{na}(x)$ and bound the approximation error uniformly by $c\cdot\varepsilon$. Let $\tilde{H}_{na}(x) := \int_0^x \tilde{h}_{na}(t)dt$. By construction, $|B'_{na}(h,x)-\tilde{h}_{na}(x)| \le |B'_{na}(h,x)-h_{na}(x)| + |h_{na}(x)-\tilde{h}_{na}(x)| \le 1+c\cdot\varepsilon$, implying $|\tilde{H}_{na}(x)-B_{na}(h,x)| \le x(1+c\cdot\varepsilon)$.

In Section 5, we construct our estimator $\hat{H}$ as follows.

First, we divide the symbols into small- and large- probability symbols according to their counts in an independent n-element sample sequence. The concentration inequalities in Section 3 imply that this step can be performed with relatively high confidence. Then, we estimate the partial entropy of each small-probability symbol i with a near-unbiased estimator of $\tilde{H}_{na}(p_i)$, and the combined partial entropy of the large-probability symbols with a simple variant of the empirical estimator. The final estimator is the sum of these small- and large- probability estimators.

In Section 6, we bound the bias of $\hat{H}$. In Sections 6.1 and 6.2, we use properties of $\tilde{H}_{na}$ and the Bernstein polynomials to bound the partial biases of the small- and large-probability estimators in terms of n, respectively. The critical observation is $|\sum_i(\tilde{H}_{na}(p_i)-B_{na}(h,p_i))| \le \sum_i p_i(1+c\cdot\varepsilon) = 1+c\cdot\varepsilon$, implying that the small-probability estimator has a low bias. To bound the bias of the large-probability estimator, we principally rely on the elegant inequality $|B_n(h,x)-h(x)| \le 1/n$.

By the triangle inequality, it remains to bound the mean absolute deviation of $\hat{H}$. We bound this quantity by bounding the partial variances of the small- and large- probability estimators in Section 7.1 and Section 7.2, respectively. Intuitively speaking, the small-probability estimator has a small variance because it is constructed based on a low-degree polynomial; the large-probability estimator has a small variance because $h(x)$ is smoother for larger values of x.

To demonstrate the efficacy of our methods, in Section 8, we compare the experimental performance of our estimators with that of the state-of-the-art property estimators for Shannon entropy and support size over nine distributions. Our competitive estimators outperformed these existing algorithms on nearly all the experimented instances.

Replacing the simple function $h_n(x)$ by a much finer approximation of $B_n(h,x)$ based on *differential smoothing*, we establish the full version of Theorem 1 in Appendix A. Applying similar techniques, we prove the other four results in Appendices B (Theorem 2 and 3), D (Theorem 4), and E (Theorem 5).

Computational complexity Section 9 presents the Remez algorithm (Remez, 1934; Pachón & Trefethen, 2009; Trefethen, 2013) for computing the best polynomial approximation of a function, and shows that it takes only $\tilde{O}(n)$ time to compute our approximation-based estimators.

3 Concentration Inequalities

The following lemma gives tight tail probability bounds for Poisson and binomial random variables.

Lemma 2 (Chung & Lu (2006)). *Let X be a Poisson or binomial random variable with mean μ, then for any $\delta > 0$,*

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu} \leq e^{-(\delta^2 \wedge \delta)\mu/3},$$

and for any $\delta \in (0,1)$,

$$\mathbb{P}(X \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\right)^{\mu} \leq e^{-\delta^2\mu/2}.$$

4 Approximating Bernstein Polynomials

With n samples, the bias of the empirical estimator in estimating $H(p)$ is

$$\mathrm{Bias}_n(\hat{H}^E, p) := \mathbb{E}[\hat{H}^E(X^n)] - H(p).$$

By the linearity of expectation, the right-hand side equals

$$\mathbb{E}[\hat{H}^E(X^n)] - H(p) = \sum_{i\in[k]} \left(\mathbb{E}\left[h\left(\frac{N_i}{n}\right)\right] - h(p_i)\right).$$

Noting that the degree-n Bernstein polynomial of h is

$$B_n(h,x) := \mathop{\mathbb{E}}_{N_x\sim\mathrm{bin}(n,x)}\left[h\left(\frac{N_x}{n}\right)\right] = \sum_{j=0}^{n} h\left(\frac{j}{n}\right)\binom{n}{j}x^j(1-x)^{n-j},$$

we can express the bias of the empirical estimator as

$$\mathrm{Bias}_n(\hat{H}^E, p) = \sum_{i\in[k]} \left(B_n(h,p_i) - h(p_i)\right).$$

Given a sampling number n and a parameter $\varepsilon \leq 1$, define the amplification factor $a := \varepsilon \log n$. Let c_l and c_s be sufficiently large and small absolute constants, respectively. In the following sections, we find a polynomial $\tilde{h}_{na}(x)$ of degree $d - 1 := d_n - 1 := c_s \log n - 1$, whose error in approximating $B'_{na}(h,x)$ over $I_n := [0,\tau_n] := [0, c_l(\log n)/n]$ satisfies

$$|B'_{na}(h,x) - \tilde{h}_{na}(x)| \leq 1 + \mathcal{O}(\varepsilon).$$

By the triangle inequality of integrals, the degree-d polynomial

$$\tilde{H}_{na}(x) := \int_0^x \tilde{h}_{na}(t)dt,$$

approximates $B_{na}(h,x)$ with the following pointwise error guarantee.

Lemma 3. *For any $x \in I_n$,*

$$|B_{na}(h,x) - \tilde{H}_{na}(x)| \leq x\left(1 + \mathcal{O}\left(\varepsilon\right)\right).$$

In Section 4.1, we relate $B'_n(h,x)$ to a simple function $h_n(x)$ that can be expressed in terms of $h(x)$. In Section 4.2, we approximate $h_n(x)$ by a linear combination of degree-d min-max polynomials of $h(x)$ over different intervals. The resulting polynomial is $\tilde{h}_{na}(x)$.

4.1 Derivative of Bernstein Polynomials

According to Bustamante (2017), the first-order derivative of the Bernstein polynomial $B_n(h,x)$ is

$$B'_n(h,x) := \sum_{j=0}^{n-1} n\left(h\left(\frac{j+1}{n}\right) - h\left(\frac{j}{n}\right)\right)\binom{n-1}{j}x^j(1-x)^{(n-1)-j}.$$

Hence, letting

$$h_n(x) := n\left(h\left(\left(\frac{n-1}{n}\right)x + \frac{1}{n}\right) - h\left(\left(\frac{n-1}{n}\right)x\right)\right),$$

we can write derivative B_n' as

$$B_n'(h,x) = \sum_{j=0}^{n-1} h_n\left(\frac{j}{n-1}\right)\binom{n-1}{j}x^j(1-x)^{(n-1)-j} = B_{n-1}(h_n,x).$$

Recall that $h(x) = -x\log x$. After some algebra, we get

$$h_n(x) = -\log\left(\frac{n-1}{n}\right) + (n-1)\left(h\left(x + \frac{1}{n-1}\right) - h(x)\right).$$

Furthermore, utilizing analytical attributes of $h(x)$ (Berens et al., 1972), we can bound the absolute difference between $h(x)$ and its Bernstein polynomial as follows.

Lemma 4. *For any $m > 0$ and $x \in [0,1]$,*

$$-\frac{1-x}{m} \le B_m(h,x) - h(x) \le 0.$$

As an immediate corollary,

Corollary 2. *For any $x \in [0, 1-(n-1)^{-1}]$,*

$$|B_n'(h,x) - h_n(x)| = |B_{n-1}(h_n,x) - h_n(x)| \le 1.$$

Proof. Given the equality $B_n'(h,x) = B_{n-1}(h_n,x)$ for $x \in [0, 1-(n-1)^{-1}]$,

$$\begin{aligned}
|B_{n-1}(h_n,x) - h_n(x)| &\le (n-1)|(B_{n-1}(h, x+(n-1)^{-1}) - h(x+(n-1)^{-1})) \\
&\quad - (B_{n-1}(h,x) - h(x))| \\
&\le (n-1)\left|\max\left\{\frac{1-x-(n-1)^{-1}}{n-1}, \frac{1-x}{n-1}\right\}\right| \\
&\le 1,
\end{aligned}$$

where the second inequality follows by Lemma 4. □

4.2 Approximating the Derivative Function

Denote the degree-d min-max polynomial of h over $[0,1]$ by

$$\tilde{h}(x) := \sum_{j=0}^{d} b_j x^j.$$

As shown in Wu & Yang (2016), the coefficients of $\tilde{h}(x)$ satisfy

$$|b_j| \lesssim 2^{3d},$$

and the error of $\tilde{h}(x)$ in approximating $h(x)$ admits

$$\max_{x\in[0,1]} |h(x) - \tilde{h}(x)| \lesssim \frac{1}{\log^2 n}.$$

By a change of variables, the degree-d min-max polynomial of h over $I_n = [0, c_l \log n/n]$ is

$$\tilde{h}_1(x) := \sum_{j=0}^{d} b_j \left(\frac{n}{c_l \log n}\right)^{j-1} x^j + \left(\log\frac{n}{c_l\log n}\right)x.$$

Correspondingly, for any $x \in I_n$, we have

$$\max_{x\in I_n} |h(x) - \tilde{h}_1(x)| \lesssim \frac{1}{n\log n}.$$

To approximate $h_{na}(x)$, we approximate $h(x)$ by $\tilde{h}_1(x)$, and $h(x+(na-1)^{-1})$ by $\tilde{h}_1(x+(na-1)^{-1})$. Then, the resulting polynomial is

$$\begin{aligned}\tilde{h}_{na}(x) &:= -\log\frac{na-1}{na} + (na-1)\left(\tilde{h}_1(x+(na-1)^{-1}) - \tilde{h}_1(x)\right)\\ &= -\log\frac{na-1}{c_l a\log n} + (na-1)\left(\sum_{j=0}^{d} b_j\left(\frac{n}{c_l\log n}\right)^{j-1}\left(\left(x+\frac{1}{na-1}\right)^j - x^j\right)\right).\end{aligned}$$

By the above reasoning, the error of $\tilde{h}_{na}$ in approximating h_{na} over I_n satisfies

$$\max_{x\in I_n}|h_{na}(x)-\tilde{h}_{na}(x)| \lesssim \frac{na}{n\log n} \lesssim \varepsilon.$$

Moreover, by an application of Corollary 2,

$$\max_{x\in[0,1/2]}|B'_{na}(h,x)-h_{na}(x)| = \max_{x\in[0,1/2]}|B_{na-1}(h_{na},x)-h_{na}(x)| \le 1.$$

The triangle inequality combines the above two inequalities and yields

$$\max_{x\in I_n}|B'_{na}(h,x)-\tilde{h}_{na}(x)| \le 1+\mathcal{O}(\varepsilon).$$

Therefore, denoting

$$\tilde{H}_{na}(x) := \int_0^x \tilde{h}_{na}(t)dt,$$

and noting that $B_{na}(h,0)=0$, we have

Lemma 5. *For any $x\in I_n$,*

$$|B_{na}(h,x)-\tilde{H}_{na}(x)| \le \int_0^x |B'_{na}(h,t)-\tilde{h}_{na}(t)|dt \le x\left(1+\mathcal{O}(\varepsilon)\right).$$

5 A Competitive Entropy Estimator

In this section, we design an explicit entropy estimator $\hat{H}$ based on $\tilde{H}_{na}$ and the empirical estimator. Note that $\tilde{H}_{na}(x)$ is a polynomial with a zero constant term. For $t\ge 1$, denote

$$g_t := \sum_{j=t}^{d}\frac{b_j}{j+1}\left(\frac{n}{c_l\log n}\right)^{j-1}\left(\frac{1}{na-1}\right)^{j-t}\binom{j+1}{j-t+1}.$$

Setting $b'_t = g_t$ for $t\ge 2$ and $b'_1 = g_1 - \log\frac{na-1}{c_l a\log n}$, we have the following lemma.

Lemma 6. *The function $\tilde{H}_{na}(x)$ can be written as*

$$\tilde{H}_{na}(x) = \sum_{t=1}^{d} b'_t x^t.$$

In addition, its coefficients satisfy

$$|b'_t| \lesssim 2^{4d}\left(\frac{n}{c_l\log n}\right)^{t-1}.$$

The proof of the above lemma is postponed to the end of this section.

To simplify our analysis and remove the dependency between symbol counts, we use the conventional *Poisson sampling* technique (Wu & Yang, 2016; Acharya et al., 2017a). Specifically, instead of drawing exactly n samples, we make the sample size an independent Poisson random variable N with mean n. This does not change the statistical nature of the problem as $N\sim\mathrm{Poi}(n)$ highly concentrates around its mean (see Lemma 2). We still define N_i as the count of symbol i in X^N. Due to Poisson sampling, these counts are now mutually independent and satisfy $N_i\sim\mathrm{Poi}(np_i),\ \forall i\in[k]$.

For each $i\in[k]$, let $N_i^{\underline{t}} := \prod_{m=0}^{t-1}(N_i-m)$ be the t-th order *falling factorial* of N_i. The following identity is well-known:

$$\mathbb{E}[N_i^{\underline{t}}] = (np_i)^t,\ \forall t\le n.$$

Note that for sufficiently small c_s or sufficiently large n, the degree parameter $d = c_s \log n \le n, \forall n$. By the linearity of expectation, the unbiased estimator of $\tilde{H}_{na}(p_i)$ is

$$\hat{H}_{na}(N_i) := \sum_{t=1}^{d} b'_t \frac{N_i^{\underline{t}}}{n^t}.$$

Let N' be an independent Poisson variable with mean n, and $X^{N'}$ be an independent length-N' sample sequence drawn from p. Analogously, we denote by N'_i the number of times that symbol $i \in [k]$ appears. Depending on whether $N'_i > \varepsilon^{-1}$ or not, we classify $p_i, i \in [k]$, into two categories: small- and large- probabilities. For small probabilities, we apply a simple variant of $\hat{H}_{na}(N_i)$; for large probabilities, we estimate $h(p_i)$ by an empirical-estimator variant.

Specifically, for each $i \in [k]$, we estimate $h(p_i)$ by

$$\hat{h}(N_i, N'_i) := \hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i \le c_l \log n} \cdot \mathbb{1}_{N'_i \le \varepsilon^{-1}} + h\left(\frac{N_i}{n}\right) \cdot \mathbb{1}_{N'_i > \varepsilon^{-1}}.$$

Consequently, we approximate $H(p)$ by

$$\hat{H}(X^N, X^{N'}) := \sum_{i \in [k]} \hat{h}(N_i, N'_i).$$

For the simplicity of illustration, we will refer to

$$\hat{H}_S(X^N, X^{N'}) := \sum_{i \in [k]} \hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i \le c_l \log n} \cdot \mathbb{1}_{N'_i \le \varepsilon^{-1}}$$

as the *small-probability estimator*, and

$$\hat{H}_L(X^N, X^{N'}) := \sum_{i \in [k]} h\left(\frac{N_i}{n}\right) \cdot \mathbb{1}_{N'_i > \varepsilon^{-1}}$$

as the *large-probability estimator*. Then, $\hat{H}$ is the sum of these two estimators.

In the next two sections, we analyze the bias and mean absolute deviation of $\hat{H}$. In Section 6, we show that for any p, the absolute bias of $\hat{H}$ satisfies

$$\left|\mathbb{E}[\hat{H}(X^N, X^{N'})] - H(p)\right| \le \left|\mathrm{Bias}(\hat{H}^E, na)\right| + (1 + \mathcal{O}(\varepsilon)) \left(1 \wedge (\varepsilon^{-1} + 1)\frac{S_p}{n}\right).$$

In Section 7, we further show that the mean absolute deviation of $\hat{H}$ satisfies

$$\mathbb{E}\left|\hat{H}(X^N, X^{N'}) - \mathbb{E}[\hat{H}(X^N, X^{N'})]\right| \lesssim \frac{1}{n^{1-\Theta(c_s)}}.$$

For sufficiently small c_s, the triangle inequality combines the above inequalities and yields

$$\mathbb{E}\left|\hat{H}(X^N, X^{N'}) - H(p)\right| \le \mathrm{Bias}(\hat{H}^E, na) + (1 + c \cdot \varepsilon) \wedge \left(\frac{S_p}{\varepsilon n} + \frac{1}{n^{0.49}}\right).$$

This basically completes the proof of Theorem 6.

Proof of Lemma 6

We begin by proving the first claim:

$$\tilde{H}_{na}(x) = -\sum_{t=1}^{d} b'_t x^t.$$

By definition, $\tilde{H}_{na}(x)$ satisfies

$$\begin{aligned}
&\tilde{H}_{na}(x) + \left(\log \frac{na-1}{c_l a \log n}\right) x \\
&= (na-1)\left(\sum_{j=1}^{d} \frac{b_j}{j+1}\left(\frac{n}{c_l \log n}\right)^{j-1}\left(\left(x + \frac{1}{na-1}\right)^{j+1} - \left(\frac{1}{na-1}\right)^{j+1} - x^{j+1}\right)\right) \\
&= \sum_{j=1}^{d} \frac{b_j}{j+1}\left(\frac{n}{c_l \log n}\right)^{j-1}\left(\sum_{m=0}^{j-1}\left(\frac{1}{na-1}\right)^{m} x^{j-m}\binom{j+1}{m+1}\right) \\
&= \sum_{t=1}^{d} x^t \left(\sum_{j=t}^{d} \frac{b_j}{j+1}\left(\frac{n}{c_l \log n}\right)^{j-1}\left(\frac{1}{na-1}\right)^{j-t}\binom{j+1}{j-t+1}\right),
\end{aligned}$$

where the last step follows by reorganizing the indices.

Next we establish the second claim. Recall that $d = c_s \log n$, thus,

$$\log \frac{na-1}{c_l a \log n} \lesssim 2^{4d}.$$

Since $b'_t = g_t$ for $t \geq 2$ and $b'_1 = g_1 - \log \frac{na-1}{c_l a \log n}$, it suffices to bound the magnitude of g_t:

$$\begin{aligned}
|g_t| &\leq \sum_{j=t}^{d} \frac{|b_j|}{j+1} \left(\frac{n}{c_l \log n}\right)^{j-1} \left(\frac{1}{na-1}\right)^{j-t} \binom{j+1}{j-t+1} \\
&\leq \sum_{j=t}^{d} |b_j| \left(\frac{1}{c_l \log n}\right)^{j-1} n^{t-1} \binom{j}{t} \\
&\leq \left(\frac{n}{c_l \log n}\right)^{t-1} \sum_{j=t}^{d} |b_j| \binom{j}{t} \\
&\leq \left(\frac{n}{c_l \log n}\right)^{t-1} \sum_{j=t}^{d} |b_j| \binom{d}{j-t} \\
&\lesssim 2^{4d} \left(\frac{n}{c_l \log n}\right)^{t-1}.
\end{aligned}$$

6 Bounding the Bias of $\hat{H}$

By the triangle inequality, the absolute bias of $\hat{H}$ in estimating $H(p)$ satisfies

$$\begin{aligned}
\left| \sum_{i \in [k]} \left(\mathbb{E}[\hat{h}(N_i, N'_i)] - h(p_i)\right) \right| \leq & \left| \sum_{i \in [k]} \left(B_{na}(h, p_i) - h(p_i)\right) \right| \\
& + \left| \sum_{i \in [k]} \left(\mathbb{E}[\hat{h}(N_i, N'_i)] - B_{na}(h, p_i)\right) \right|.
\end{aligned}$$

Note that the first term on the right-hand side is the absolute bias of the empirical estimator with sample size $na = \varepsilon n \log n$, that is,

$$\mathrm{Bias}_{na}(\hat{H}^E, p) = \left| \sum_{i \in [k]} \left(B_{na}(h, p_i) - h(p_i)\right) \right|.$$

Hence, we need to consider only the second term on the right-hand side, which admits

$$\left| \sum_{i \in [k]} \left(\mathbb{E}[\hat{h}(N_i, N'_i)] - B_{na}(h, p_i)\right) \right| \leq \mathrm{Bias}_S + \mathrm{Bias}_L,$$

where

$$\mathrm{Bias}_S := \left| \sum_{i \in [k]} \mathbb{E}\left[\left(\hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i \leq c_l \log n} - B_{na}(h, p_i)\right) \cdot \mathbb{1}_{N'_i \leq \varepsilon^{-1}} \right] \right|$$

is the absolute bias of the small-probability estimator, and

$$\mathrm{Bias}_L := \left| \sum_{i \in [k]} \mathbb{E}\left[\left(h\left(\frac{N_i}{n}\right) - B_{na}(h, p_i) \right) \cdot \mathbb{1}_{N'_i > \varepsilon^{-1}} \right] \right|$$

is the absolute bias of the large-probability estimator.

Assume that c_l is sufficiently large. In Section 6.1, we bound the small-probability bias by

$$\mathrm{Bias}_S \leq (1 + \mathcal{O}(\varepsilon)) \left(1 \wedge (\varepsilon^{-1} + 1) \frac{S_p}{n}\right).$$

In Section 6.2, we bound the large-probability bias by

$$\mathrm{Bias}_L \leq 2 \left(\varepsilon \wedge \frac{S_p}{n}\right).$$

6.1 Bias of the Small-Probability Estimator

We first consider and analyze Bias_S. By the triangle inequality,

$$\begin{aligned}\mathrm{Bias}_S \le &\sum_{i:p_i\notin I_n} \left|\mathbb{E}[\hat{H}_{na}(N_i)\cdot \mathbb{1}_{N_i\le c_l\log n}] - B_{na}(h,p_i)\right|\cdot \mathbb{E}[\mathbb{1}_{N_i'\le\varepsilon^{-1}}]\\ &+\sum_{i:p_i\in I_n} \left|\mathbb{E}\left[\hat{H}_{na}(N_i)\right] - B_{na}(h,p_i)\right|\cdot \mathbb{E}[\mathbb{1}_{N_i'\le\varepsilon^{-1}}]\\ &+\sum_{i:p_i\in I_n} \left|\mathbb{E}\left[\hat{H}_{na}(N_i)\cdot \mathbb{1}_{N_i> c_l\log n}\right]\cdot \mathbb{E}\left[\mathbb{1}_{N_i'\le\varepsilon^{-1}}\right]\right|.\end{aligned}$$

Let us assume $\varepsilon\log n\ge 1$ and consider the first sum on the right-hand side. By the general reasoning in the proof of Lemma 7, we can show that

$$\hat{H}_{na}(N_i)\cdot \mathbb{1}_{N_i\le c_l\log n}\lesssim 2^{5d}\cdot\frac{\log^2 n}{n}.$$

Further assume that c_s and c_l are sufficiently small and large, respectively. For large enough n, the above inequality bounds the first sum by

$$\sum_{i:p_i\notin I_n}\left|\hat{H}_{na}(N_i)\cdot \mathbb{1}_{N_i\le c_l\log n} - B_{na}(h,p_i)\right|\cdot\mathbb{E}[\mathbb{1}_{N_i'\le\varepsilon^{-1}}]\le \sum_{i:p_i\notin I_n}\mathbb{E}[\mathbb{1}_{N_i'\le\varepsilon^{-1}}]\le\frac{1}{n^5}\cdot\frac{n}{c_l\log n}\le\frac{1}{n^4}.$$

For the second sum on the right-hand side, by Lemma 5,

$$\begin{aligned}\sum_{i:p_i\in I_n}\left|\mathbb{E}\left[\hat{H}_{na}(N_i)\right] - B_{na}(h,p_i)\right|\cdot\mathbb{E}[\mathbb{1}_{N_i'\le\varepsilon^{-1}}] &\le \sum_{i:p_i\in I_n}\left|\mathbb{E}\left[\hat{H}_{na}(N_i)\right] - B_{na}(h,p_i)\right|\cdot\mathbb{E}[\mathbb{1}_{N_i'\le\varepsilon^{-1}}]\\ &= \sum_{i:p_i\in I_n}\left|\tilde{H}_{na}(p_i) - B_{na}(h,p_i)\right|\cdot\mathbb{E}[\mathbb{1}_{N_i'\le\varepsilon^{-1}}]\\ &\le \sum_{i:p_i\in I_n}(1+\mathcal{O}(\varepsilon))\,p_i\cdot\mathbb{E}[\mathbb{1}_{N_i'\le\varepsilon^{-1}}]\\ &\le (1+\mathcal{O}(\varepsilon))\left(1\wedge(\varepsilon^{-1}+1)\frac{S_p}{n}\right).\end{aligned}$$

The following lemma bounds the last sum and completes our argument.

Lemma 7. *For sufficiently large c_l,*

$$\sum_{i\in[k]}\left|\mathbb{E}\left[\hat{H}_{na}(N_i)\cdot\mathbb{1}_{N_i>c_l\log n}\right]\cdot\mathbb{E}\left[\mathbb{1}_{N_i'\le\varepsilon^{-1}}\right]\right|\le\frac{1}{n^5}.$$

Proof. For simplicity, we assume that $c_l\ge 4$ and $\varepsilon\log n\ge 1$. By the triangle inequality,

$$\begin{aligned}&\left|\mathbb{E}\left[\hat{H}_{na}(N_i)\cdot\mathbb{1}_{N_i>c_l\log n}\right]\cdot\mathbb{E}\left[\mathbb{1}_{N_i'\le\varepsilon^{-1}}\right]\right|\\ &\le\sum_{j=1}^{\infty}\left|\mathbb{E}\left[\hat{H}_{na}(N_i)\cdot\mathbb{1}_{c_l(j+1)\log n\ge N_i>c_l j\log n}\right]\cdot\mathbb{E}\left[\mathbb{1}_{N_i'\le\varepsilon^{-1}}\right]\right|.\end{aligned}$$

To bound the last term, we rely on the following result: For any $j\ge 1$,

$$\left|\mathbb{E}\left[\mathbb{1}_{c_l(j+1)\log n\ge N_i>c_l j\log n}\right]\cdot\mathbb{E}\left[\mathbb{1}_{N_i'\le\varepsilon^{-1}}\right]\right|\le\left(1+\varepsilon^{-1}\right)np_i\cdot e^{-\Theta(c_l j\log n)}.$$

To prove this inequality, we apply Lemma 2 and consider two cases:
Case 1: If $np_i<(3c_l/8)j\log n$, then

$$\mathbb{E}\left[\mathbb{1}_{c_l(j+1)\log n\ge N_i>c_l j\log n}\right]\le np_i\cdot e^{-\Theta(c_l j\log n)}.$$

Case 2: If $np_i\ge(3c_l/8)j\log n$, then

$$\mathbb{E}\left[\mathbb{1}_{N_i'\le\varepsilon^{-1}}\right]\le np_i\varepsilon^{-1}\cdot e^{-\Theta(c_l j\log n)}.$$

This essentially completes the proof. Next, we bound $\hat{H}_{na}(N_i)$ for $N_i \in [c_l j \log n, c_l(j+1)\log n]$:

$$\begin{aligned}
|\hat{H}_{na}(N_i)| &= \left|\left(\log\frac{na-1}{c_l a \log n}\right)\frac{N_i}{n} + \sum_{t=1}^{d} b'_t \frac{N_i^{\underline{t}}}{n^t}\right| \\
&\lesssim 2^{4d} \cdot \sum_{t=1}^{c_s \log n} \left(\frac{n}{c_l \log n}\right)^{t-1} \frac{(c_l(j+1)\log n)^t}{n^t} \\
&\lesssim 2^{5d} \cdot \frac{c_l j \log n}{n} \sum_{t=1}^{c_s \log n} j^{t-1} \\
&\lesssim 2^{5d} \cdot \frac{c_l j \log n}{n}(j^{c_s \log n} + c_s \log n).
\end{aligned}$$

Hence, for sufficiently large c_l,

$$\begin{aligned}
&\left|\mathbb{E}\left[\hat{H}_{na}(N_i)\cdot \mathbb{1}_{N_i > c_l \log n}\right]\cdot \mathbb{E}\left[\mathbb{1}_{N'_i \le \varepsilon^{-1}}\right]\right| \\
&\le \sum_{j=1}^{\infty} \left|\mathbb{E}\left[\hat{H}_{na}(N_i)\cdot \mathbb{1}_{c_l(j+1)\log n \ge N_i > c_l j \log n}\right]\cdot \mathbb{E}\left[\mathbb{1}_{N'_i \le \varepsilon^{-1}}\right]\right| \\
&\le \sum_{j=1}^{\infty} \mathcal{O}(2^{5d}) \cdot c_l j \log n (j^{c_s \log n} + c_s \log n)\cdot \mathbb{E}\left[\mathbb{1}_{c_l(j+1)\log n \ge N_i > c_l j \log n}\right]\cdot \mathbb{E}\left[\mathbb{1}_{N'_i \le \varepsilon^{-1}}\right] \\
&\lesssim 2^{5d} \cdot \sum_{j=1}^{\infty} \left(1+\varepsilon^{-1}\right) p_i \cdot e^{-\Theta(c_l j \log n)} \cdot c_l j \log n (j^{c_s \log n} + c_s \log n) \\
&\le p_i \sum_{j=1}^{\infty} \frac{1}{2n^{5j}} \\
&\le \frac{p_i}{n^5}.
\end{aligned}$$

Summing the right-hand side over $i \in [k]$ yields the desired result. □

6.2 Bias of the Large-Probability Estimator

This section proves the bias bound $\mathrm{Bias}_L \le 2(\varepsilon \wedge (S_p/n))$. By the triangle inequality,

$$\begin{aligned}
\mathrm{Bias}_L &\le \sum_{i\in[k]} \left|\mathbb{E}\left[h\left(\frac{N_i}{n}\right) - B_{na}(h,p_i)\right]\right| \cdot \mathbb{E}\left[\mathbb{1}_{N'_i > \varepsilon^{-1}}\right] \\
&\le \sum_{i\in[k]} |h(p_i) - B_{na}(h,p_i)| \cdot \mathbb{E}\left[\mathbb{1}_{N'_i > \varepsilon^{-1}}\right] + \sum_{i\in[k]} \left|\mathbb{E}\left[h\left(\frac{N_i}{n}\right) - h(p_i)\right]\right| \cdot \mathbb{E}\left[\mathbb{1}_{N'_i > \varepsilon^{-1}}\right].
\end{aligned}$$

We need the following inequality to bound the right-hand side.

$$0 \le x\log x - (x-1) \le (x-1)^2, \ \forall x \in [0,1].$$

For simplicity, denote $\hat{p}_i := N_i/n$. Then,

$$\begin{aligned}
\left|\mathbb{E}\left[h\left(\frac{N_i}{n}\right) - h(p_i)\right]\right| &= |\mathbb{E}[p_i \log p_i - \hat{p}_i \log \hat{p}_i]| \\
&\le |\mathbb{E}[p_i \log p_i - \hat{p}_i \log p_i]| + |\mathbb{E}[\hat{p}_i \log p_i - \hat{p}_i \log \hat{p}_i]| \\
&= p_i \cdot \left|\mathbb{E}\left[\frac{\hat{p}_i}{p_i}\log\frac{\hat{p}_i}{p_i}\right]\right| \\
&\le p_i \cdot \left|\mathbb{E}\left[\left(\frac{\hat{p}_i}{p_i} - 1\right) + \left(\frac{\hat{p}_i}{p_i} - 1\right)^2\right]\right| \\
&= \frac{1}{n}.
\end{aligned}$$

Replacing n by na in the above argument yields

$$|h(p_i) - B_{na}(h,p_i)| \le \frac{1}{na}.$$

Consider the first term on the right-hand side. By the last bound and Markov's inequality,

$$\begin{aligned}\sum_{i\in[k]} |h(p_i) - B_{na}(h,p_i)| \cdot \mathbb{E}\left[\mathbb{1}_{N_i'>\varepsilon^{-1}}\right] &\le \frac{1}{na} \sum_{i\in[k]} \mathbb{E}\left[\mathbb{1}_{N_i'>\varepsilon^{-1}}\right] \\ &\le \frac{1}{na} \sum_{i\in[k]} (\mathbb{1}_{p_i>0} \wedge \varepsilon n p_i) \\ &\le \varepsilon \wedge \frac{S_p}{n}.\end{aligned}$$

For the second term, an analogous argument yields

$$\sum_{i\in[k]} \left|\mathbb{E}\left[h\left(\frac{N_i}{n}\right) - h(p_i)\right]\right| \cdot \mathbb{E}\left[\mathbb{1}_{N_i'>\varepsilon}\right] \le \varepsilon \wedge \frac{S_p}{n}.$$

7 Bounding the Mean Absolute Deviation of $\hat{H}$

By Jensen's inequality,

$$\mathbb{E}|\hat{H}(X^N, X^{N'}) - \mathbb{E}[\hat{H}(X^N, X^{N'})]| \le \sqrt{\mathrm{Var}(\hat{H}(X^N, X^{N'}))}.$$

Hence, to bound the mean absolute deviation of $\hat{H}$, it suffices to bound its variance. Note that the symbol counts are mutually independent. The inequality $\mathrm{Var}(X+Y) \le 2(\mathrm{Var}(X)+\mathrm{Var}(Y))$ implies

$$\mathrm{Var}(\hat{H}(X^N, X^{N'})) = \sum_{i\in[k]} \mathrm{Var}(\hat{h}(N_i, N_i')) \le 2\mathrm{Var}_S + 2\mathrm{Var}_L,$$

where

$$\mathrm{Var}_S := \sum_{i\in[k]} \mathrm{Var}\left(\hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i\le c_l \log n} \cdot \mathbb{1}_{N_i'\le\varepsilon^{-1}}\right)$$

is the variance of the small-probability estimator, and

$$\mathrm{Var}_L := \sum_{i\in[k]} \mathrm{Var}\left(h\left(\frac{N_i}{n}\right) \cdot \mathbb{1}_{N_i'>\varepsilon^{-1}}\right)$$

is the variance of the large-probability estimator. Assume that c_l and c_s are sufficiently large and small absolute constants. In Section 7.1 and 7.2, we will respectively establish

$$\mathrm{Var}_S \lesssim \frac{1}{n^{1-\Theta(c_s)}} \text{ and } \mathrm{Var}_L \lesssim \frac{(\log n)^3}{n}.$$

7.1 Variance of the Small-Probability Estimator

First we bound the small-probability variance Var_S and prove $\mathrm{Var}_S \le \mathcal{O}\left(1/n^{1-\Theta(c_s)}\right)$. Following the sequence of derivations in Section 6.1,

$$\begin{aligned}\mathrm{Var}_S &\le 2 \sum_{i\in[k]} \mathrm{Var}\left(\hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i>c_l\log n} \cdot \mathbb{1}_{N_i'\le\varepsilon^{-1}}\right) \\ &\quad + 2 \sum_{i\in[k]} \mathrm{Var}\left(\hat{H}_{na}(N_i) \cdot \mathbb{1}_{N_i'\le\varepsilon^{-1}}\right) \\ &\le 2 \sum_{i\in[k]} \mathbb{E}[(\hat{H}_{na}(N_i))^2 \cdot \mathbb{1}_{N_i>c_l\log n}] \cdot \mathbb{E}[\mathbb{1}_{N_i'\le\varepsilon^{-1}}] \\ &\quad + 2 \sum_{i\in[k]} \mathrm{Var}\left(\hat{H}_{na}(N_i)\right) \cdot \mathbb{E}[\mathbb{1}_{N_i'\le\varepsilon^{-1}}] + 2 \sum_{i\in[k]} (\mathbb{E}[\hat{H}_{na}(N_i)])^2 \cdot \mathrm{Var}(\mathbb{1}_{N_i'\le\varepsilon^{-1}}) \\ &\le 2 \sum_{i\in[k]} \mathbb{E}[(\hat{H}_{na}(N_i))^2 \cdot \mathbb{1}_{N_i>c_l\log n}] \cdot \mathbb{E}[\mathbb{1}_{N_i'\le\varepsilon^{-1}}] \\ &\quad + 2 \sum_{i\in[k]} \mathrm{Var}\left(\hat{H}_{na}(N_i)\right) \cdot \mathbb{E}[\mathbb{1}_{N_i'\le\varepsilon^{-1}}] + 2 \sum_{i\in[k]} (\tilde{H}_{na}(p_i))^2 \cdot \mathrm{Var}(\mathbb{1}_{N_i'\le\varepsilon^{-1}}),\end{aligned}$$

where the first step follows by $\mathrm{Var}(X-Y) \le 2(\mathrm{Var}(X)+\mathrm{Var}(Y))$, the second step follows from $\mathrm{Var}(A\cdot B) = \mathbb{E}[A^2]\mathrm{Var}(B) + \mathrm{Var}(A)(\mathbb{E}[B])^2$ for any independent random variables A and B, and the last step follows from our construction, which satisfies $\mathbb{E}[\hat{H}_{na}(N_i)] = \tilde{H}_{na}(p_i)$.

Similar to the proof of Lemma 7, for the first term on the right-hand side and sufficiently large c_l,

$$\sum_{i\in[k]} \left|\mathbb{E}\left[(\hat{H}_{na}(N_i))^2 \cdot \mathbb{1}_{N_i > c_l \log n}\right] \cdot \mathbb{E}\left[\mathbb{1}_{N_i' \le \varepsilon^{-1}}\right]\right| \le \sum_{i\in[k]} \frac{p_i}{n^3} = \frac{1}{n^3}.$$

As for the second term on the right-hand side,

$$\begin{aligned}
\sum_{i\in[k]} \mathrm{Var}\left(\hat{H}_{na}\right) \cdot \mathbb{E}[\mathbb{1}_{N_i' \le \varepsilon^{-1}}] &\lesssim 2^{8d} \cdot \sum_{i\in[k]} d \sum_{t=1}^{d} \left(\frac{n}{c_l \log n}\right)^{2(t-1)} \frac{\mathrm{Var}(N_i^t)}{n^{2t}} \cdot \mathbb{E}[\mathbb{1}_{N_i' \le \varepsilon^{-1}}] \\
&\le 2^{8d} \cdot \frac{d}{n^2} \sum_{i\in[k]} \sum_{t=1}^{d} \left(\frac{1}{c_l \log n}\right)^{2(t-1)} \mathrm{Var}(N_i^t) \cdot \mathbb{E}[\mathbb{1}_{N_i' \le \varepsilon^{-1}}] \\
&\le 2^{8d} \cdot \frac{d}{n^2} \sum_{i\in[k]} \sum_{t=1}^{d} \left(\frac{1}{c_l \log n}\right)^{2(t-1)} (np_i)^t \sum_{j=0}^{t-1} \binom{t}{j} (np_i)^j \frac{t!}{j!} \cdot \mathbb{E}[\mathbb{1}_{N_i' \le \varepsilon^{-1}}] \\
&\le 2^{8d} \cdot \frac{d}{n^2} \sum_{i\in[k]} \sum_{t=1}^{d} \left(\frac{1}{c_l \log n}\right)^{2(t-1)} (np_i)^t (t+np_i)^t \cdot \mathbb{E}[\mathbb{1}_{N_i' \le \varepsilon^{-1}}] \\
&\le 2^{8d} \cdot \frac{d}{n^2} \sum_{i\in[k]} \sum_{t=1}^{d} \left(\frac{1}{c_l \log n}\right)^{2(t-1)} 2^t((np_i)^{2t} + (np_i)^t t^t) \cdot \Pr(N_i' \le \varepsilon^{-1}) \\
&\le 2^{8d} \cdot \frac{d}{n} \sum_{i\in[k]} p_i \sum_{t=1}^{d} \left(\frac{1}{c_l \log n}\right)^{2(t-1)} 2^t \Big((\varepsilon^{-1}+2t)^{2t-1} \cdot \Pr(N_i' \le \varepsilon^{-1}+2t) \\
&\quad + (\varepsilon^{-1}+t)^{t-1} t^t \cdot \Pr(N_i' \le \varepsilon^{-1}+t)\Big) \\
&\lesssim 2^{9d} \cdot \frac{d}{n}.
\end{aligned}$$

It remains to bound the third term. Leveraging $|\tilde{H}_{na}(p_i)| \lesssim p_i 2^{5d}$ shows that

$$\begin{aligned}
&\sum_{i\in[k]} (\tilde{H}_{na}(p_i))^2 \cdot \mathrm{Var}(\mathbb{1}_{N_i' \le \varepsilon^{-1}}) \\
&\lesssim 2^{8d} \cdot \sum_{i\in[k]} \sum_{t=1}^{d} \left(\frac{n}{c_l \log n}\right)^{2(t-1)} p_i^{2t} \cdot \mathrm{Var}(\mathbb{1}_{N_i' \le \varepsilon^{-1}}) \\
&\le 2^{8d} \cdot \sum_{i\in[k]} \sum_{t=1}^{d} \left(\frac{n}{c_l \log n}\right)^{2(t-1)} p_i^{2t} \cdot \Pr(N_i' \le \varepsilon^{-1}) \\
&= 2^{8d} \cdot \sum_{i\in[k]} p_i \sum_{t=1}^{d} \left(\frac{n}{c_l \log n}\right)^{2(t-1)} p_i^{2t-1} \cdot \sum_{m=0}^{\varepsilon^{-1}} e^{-np_i} \frac{(np_i)^m}{m!} \\
&\le 2^{8d} \cdot \sum_{i\in[k]} p_i \sum_{t=1}^{d} \left(\frac{n}{c_l \log n}\right)^{2(t-1)} \left(\frac{2t-1+\varepsilon^{-1}}{n}\right)^{2t-1} \Pr(N_i \le 2t-1+\varepsilon^{-1}) \\
&\le 2^{8d} \cdot \sum_{i\in[k]} p_i \cdot c_s \log n \cdot \frac{c_l \log n}{n} \\
&\lesssim \frac{2^{9d}}{n}.
\end{aligned}$$

Consolidating all the three bounds above yields

$$\mathrm{Var}_S \le \frac{2}{n^3} + \mathcal{O}(2^{9d}) \cdot \frac{d}{n} + \mathcal{O}\left(\frac{2^{9d}}{n}\right) \le \frac{1}{n^{1-\Theta(c_s)}},$$

where the last step follows by $d = c_s \log n$.

7.2 Variance of the Large-Probability Estimator

In this section we bound the quantity Var_L and establish $\mathrm{Var}_L \lesssim (\log n)^3/n$. Due to independence,

$$\mathrm{Var}_L = \sum_{i\in[k]} \mathrm{Var}\left(h\left(\frac{N_i}{n}\right)\cdot \mathbb{1}_{N_i'>\varepsilon^{-1}}\right).$$

The following lemma bounds the right-hand-side summation.

Lemma 8. *For any integer $s \geq 1$,*

$$\sum_{i\in[k]} \mathrm{Var}\left(h\left(\frac{N_i}{n}\right)\cdot \mathbb{1}_{N_i'>s}\right) \leq (\log n)^2 \frac{4s}{n}.$$

Proof. First, we effectively decompose the variances:

$$\begin{aligned}\sum_{i\in[k]} \mathrm{Var}\left(h\left(\frac{N_i}{n}\right)\mathbb{1}_{N_i'>s}\right) &= \mathrm{Var}(\mathbb{1}_{N_i'>s})\mathbb{E}\left[h^2\left(\frac{N_i}{n}\right)\right] + \sum_{i\in[k]} \left(\mathbb{E}[\mathbb{1}_{N_i'>s}]\right)^2 \mathrm{Var}\left(h\left(\frac{N_i}{n}\right)\right)\\ &\leq \mathrm{Var}(\mathbb{1}_{N_i'>s})\mathbb{E}\left[h^2\left(\frac{N_i}{n}\right)\right] + \sum_{i\in[k]} \mathrm{Var}\left(h\left(\frac{N_i}{n}\right)\right).\end{aligned}$$

To bound the first term on the right-hand side, note that

$$\begin{aligned}\mathrm{Var}(\mathbb{1}_{N_i'>s})\mathbb{E}\left[h^2\left(\frac{N_i}{n}\right)\right] &\leq \mathrm{Var}(\mathbb{1}_{N_i'>s})\mathbb{E}\left[(\log n)^2\left(\frac{N_i}{n}\right)^2\right]\\ &\leq (\log n)^2 \frac{p_i}{n}\left(1 + np_i\mathrm{Var}(\mathbb{1}_{N_i'>s})\right),\end{aligned}$$

where the term in the parentheses further admits

$$\begin{aligned}p_i\,\mathrm{Var}(\mathbb{1}_{N_i'>s}) &\leq p_i\cdot\mathbb{P}[N_i' \leq s]\\ &= e^{-np_i}\sum_{j=0}^{s}\frac{(np_i)^{j+1}}{(j+1)!}\frac{j+1}{n}\\ &\leq \frac{s+1}{n}e^{-np_i}\sum_{j=0}^{s}\frac{(np_i)^{j+1}}{(j+1)!}\\ &= \frac{s+1}{n}\mathbb{P}(1 \leq N_x' \leq s+1)\\ &\leq \frac{s+1}{n}.\end{aligned}$$

To bound the second term, let $\hat{N}_i$ be an i.i.d. copy of N_i for each i,

$$\begin{aligned}2\mathrm{Var}\left(h\left(\frac{N_i}{n}\right)\right) &= \mathrm{Var}\left(h\left(\frac{N_i}{n}\right) - h\left(\frac{\hat{N}_i}{n}\right)\right)\\ &= \mathbb{E}\left(h\left(\frac{N_i}{n}\right) - h\left(\frac{\hat{N}_i}{n}\right)\right)^2\\ &\leq (\log n)^2\mathbb{E}\left(\frac{N_i}{n} - \frac{\hat{N}_i}{n}\right)^2\\ &= 2(\log n)^2\cdot\frac{p_i}{n}.\end{aligned}$$

A simple combination of these bounds yields the lemma. □

Setting $s = \varepsilon^{-1}$ in Lemma 8 and assuming $\varepsilon\log n \geq 1$, we obtain

$$\mathrm{Var}_L = \sum_{i\in[k]} \mathrm{Var}\left(h\left(\frac{N_i}{n}\right)\cdot \mathbb{1}_{N_i'>\varepsilon^{-1}}\right) \leq \frac{4(\log n)^3}{n}.$$

8 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.

Experiment settings We experimented with nine distributions:

- uniform distribution;
- a two-steps distribution with probability values $0.5k^{-1}$ and $1.5k^{-1}$;
- 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 $0.3k$;
- a distribution drawn from Dirichlet prior with parameter 1;
- a distribution drawn from Dirichlet prior with parameter 1/2.

All distributions have *support size* $k = 10{,}000$. The geometric, Poisson, and Zipf distributions were truncated at k and re-normalized. The horizontal axis shows the number of samples, n, ranging from $k^{0.2}$ to k. 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 k). 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.

Experimental 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 clearly outperformed the state-of-the-art estimators in terms of accuracy and stability regarding distribution structures.

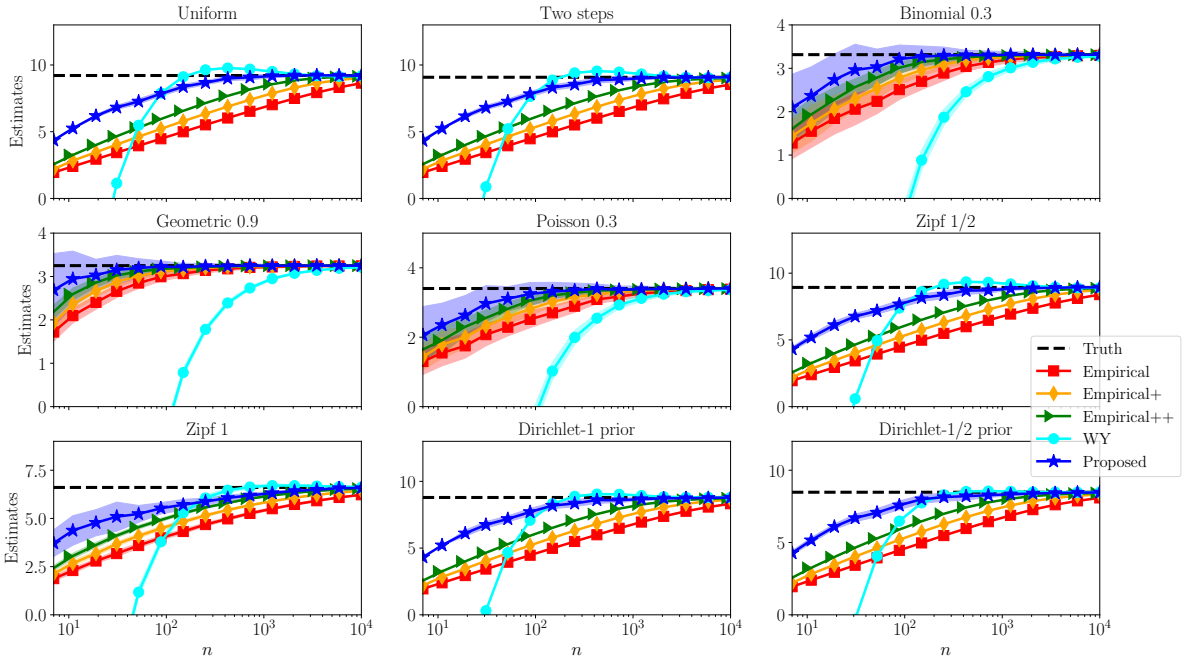

Figure 1: Shannon entropy estimation. For clarity, the horizontal axis is in logarithmic scale. The WY curve is flipped vertically around Truth for all the curves to have similar trends. Besides the samples, the WY estimator takes as input an upper bound of the support size, which is set to be the actual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

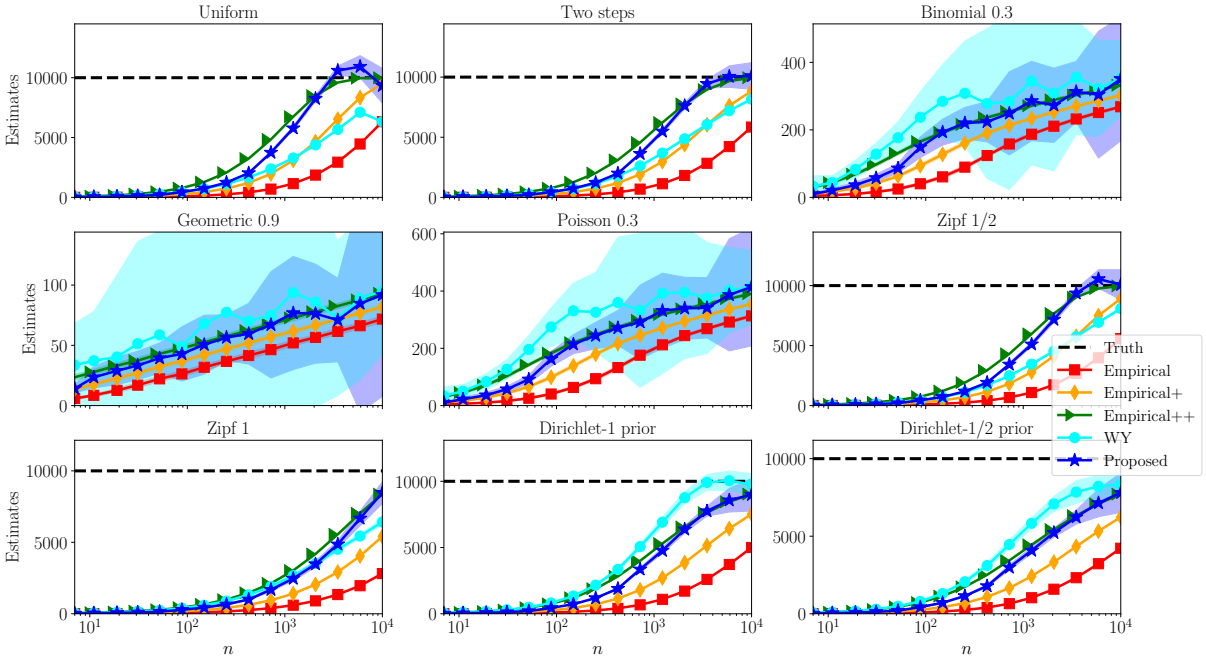

Figure 2: Support size estimation. For clarity, the horizontal axis is in logarithmic scale. Besides the samples, the WY estimator takes as input a lower bound of the smallest positive probability $p^+_{\min}$, which is set to be $\max\{1/(10k), 4p^+_{\min}\}$ in the experiments. Here, $1/(10k)$ is used 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.

9 Computational Complexity

The dominant computation step is finding the min-max polynomial of $B'_m(h,x)$, in which we use the well-known Remez algorithm (Pachón & Trefethen, 2009; Trefethen, 2013). Below, we shall argue that the algorithm takes only $\tilde{O}(n)$ time (*number of bit operations*) to well approximate $B'_m(h,x)$.

9.1 Remez Algorithm

The algorithm named after Remez (1934) is an efficient iterative algorithm that numerically computes the minimax polynomial. For a valid domain $[a,b]$, set our objective to well approximating the function $f(x):[a,b]\to\mathbb{R}$ by a degree-d real polynomial $P(x)$, in the min-max sense. We briefly illustrate a simple version of the algorithm below.

1. There are several different ways to initialize the algorithm. A popular initialization is to use the *Chebyshev nodes*. Specifically, we compute $d+2$ points $x_0,x_1,\dots x_{d+1}$ as
$$x_i := \frac{1}{2}(a+b)+\frac{1}{2}(b-a)\cos\left(\frac{2i+1}{2(d+2)}\pi\right), i=0,1,\dots,d+1.$$
2. For $x_0,x_1,...x_{d+1}$, solve the linear system of $d+2$ equations
$$b_0+b_1\cdot x_i+\dots+b_d\cdot x_i^d+E\cdot(-1)^i=f(x_i)\ \ (\text{where } i=0,1,\dots d+1),$$
for the unknowns $b_0,b_1,...b_d,$ and E.
3. (Re)form the polynomial $P(x)$ as
$$P(x):=b_0+b_1\cdot x+\dots+b_d\cdot x^d.$$
4. Compute the $d+2$ local extrema of the error function
$$\mathcal{E}(x):=P(x)-f(x)$$
over the sign-invariant regions, and denote them by $x_0^*,\dots x_{d+1}^*$, sorted in descending order.
5. Replace x_i by x_i^* for $i=0,1,\dots d+1$ and go back to Step 2 until quantity E converges.

Next, we analyze the time complexity of the Remez algorithm when applied to our setting.

9.2 Complexity of Evaluating $f(x)$

To compute our estimator, the function to approximate is the degree-$\tilde{\Theta}(n)$ polynomial $f(x):=B_m(h_m,\tau_n\cdot x)$ with $m=na-1$ (different from the prior version to simply the notation), $a\in[1,\log n]$, and $\tau_n=c_l(\log n)/n$ for a properly chosen absolute constant $c_l\geq 1$. The degree and interval for the approximation are $d=d_n=\Theta(\log n)$ and $[0,1]$, respectively.

For our purpose, it suffices to approximate $f(x)$ to an order-$1/n$ error.

First-level truncation of $f(x)$ First, we show that only the lower-order part of $f(x)$ matters in the computation. By the definition of Bernstein polynomials and $|h_{m+1}(y)|\lesssim 1, \forall y\in[0,1]$,

$$\begin{aligned}B_m(h_{m+1},\tau_n\cdot x)&=\mathop{\mathbb{E}}_{Y\sim\text{bin}(m,\tau_n\cdot x)}\left[h_{m+1}\left(\frac{Y}{m}\right)\right]\\&=\sum_{t=0}^{m}h_{m+1}\left(\frac{t}{m}\right)\cdot\Pr\left(\text{bin}(m,\tau_n\cdot x)=t\right)\\&=\left[\sum_{t=0}^{4c_l\log^2 n}h_{m+1}\left(\frac{t}{m}\right)\cdot\Pr\left(\text{bin}(m,\tau_n x)=t\right)\right]+\mathcal{O}(\Pr(\text{bin}(m,\tau_n x)>4c_l\log^2 n)).\end{aligned}$$

Note that $m\tau_n\cdot x\leq c_l\log^2 n$ for $x\in[0,1]$. Then, by standard binomial tail bounds, e.g, Lemma 2,

$$\Pr(\text{bin}(m,\tau_n\cdot x)>4c_l\log^2 n)\leq e^{-c_l(\log^2 n)}\leq\frac{1}{n^{\log n}}\leq\frac{1}{n}.$$

Hence, we can *redefine* the function to approximate as

$$f(x) = \sum_{t=0}^{4c_l \log^2 n} h_{m+1}\left(\frac{t}{m}\right) \cdot \Pr\left(\mathrm{bin}(m, \tau_n x) = t\right) = \sum_{t=0}^{4c_l \log^2 n} h_{m+1}\left(\frac{t}{m}\right) \cdot \binom{m}{t} (\tau_n x)^t (1 - \tau_n x)^{m-t}.$$

A natural step to take is expending the polynomial function into its standard form.

$$\begin{aligned}
f(x) &= \sum_{t=0}^{4c_l \log^2 n} h_{m+1}\left(\frac{t}{m}\right) \cdot \binom{m}{t} (\tau_n x)^t (1 - \tau_n x)^{m-t} \\
&= \sum_{t=0}^{4c_l \log^2 n} h_{m+1}\left(\frac{t}{m}\right) \cdot \binom{m}{t} (\tau_n x)^t \sum_{j=0}^{m-t} \binom{m-t}{j} (-\tau_n x)^{m-t-j} \\
&= \sum_{s=0}^{m} x^s \cdot \left(\tau_n^s \sum_{t=0}^{\min\{s, 4c_l \log^2 n\}} h_{m+1}\left(\frac{t}{m}\right) \cdot \binom{m}{t} \binom{m-t}{s-t} (-1)^{s-t} \right).
\end{aligned}$$

For simplicity, let us denote the coefficient of x^s in $f(x)$ by C_s. Below, we bound the magnitude of C_s for $s = 0, 1, \ldots, m$. Recall that $a \lesssim b$ represents $a = \mathcal{O}(b)$ which hides only absolute constants, $|h_{m+1}(y)| \lesssim 1$ for all $y \in [0, 1]$, and $\tau_n = c_l(\log n)/n$ for an absolute constant c_l. Then,

$$\begin{aligned}
|C_s| &= \left| \tau_n^s \sum_{t=0}^{\min\{s, 4c_l \log^2 n\}} h_{m+1}\left(\frac{t}{m}\right) \cdot \binom{m}{t} \binom{m-t}{s-t} (-1)^{s-t} \right| \\
&\lesssim \left(\frac{c_l \log n}{n}\right)^s \sum_{t=0}^{s} \binom{m}{t} \binom{m-t}{s-t} \\
&\le \left(\frac{c_l \log n}{n}\right)^s (2m)^s \\
&\le \left(\frac{c_l \log n}{n}\right)^s (2n \log n)^s \\
&= \exp(\Theta(s \log \log n)).
\end{aligned}$$

Second-level truncation of $f(x)$ Following the above derivations, we can derive an alternative upper bound on C_s. This bound basically shows that for large s, the term corresponding to C_s is negligible. Specifically, consider any $s \ge 2(c_l e)^2 \log^4 n \ge (c_l e)^2 \log^4 n + 8c_l \log^2 n$ where $c_l > 1$,

$$\begin{aligned}
|C_s| &= \left| \tau_n^s \sum_{t=0}^{\min\{s, 4c_l \log^2 n\}} h_{m+1}\left(\frac{t}{m}\right) \cdot \binom{m}{t} \binom{m-t}{s-t} (-1)^{s-t} \right| \\
&\lesssim \left(\frac{c_l \log n}{n}\right)^s \sum_{t=0}^{4c_l \log^2 n} \binom{m}{t} \binom{m-t}{s-t} \\
&\le \left(\frac{c_l \log n}{n}\right)^s \sum_{t=0}^{4c_l \log^2 n} m^t \cdot \frac{m^{s-t}}{(s-t)!} \\
&\le \left(\frac{c_l \log n}{n}\right)^s (n \log n)^s \sum_{t=0}^{4c_l \log^2 n} \frac{1}{(s-t)!} \\
&\lesssim \frac{(c_l \log^2 n)^s}{(s - 4c_l \log^2 n)!} \lesssim \frac{(c_l e \log^2 n)^s}{(s - 4c_l \log^2 n)^{s - 4c_l \log^2 n}} \\
&\le \frac{((c_l e)^2 \log^4 n)^{s/2}}{(s - 4c_l \log^2 n)^{s - 4c_l \log^2 n}} \\
&\le \frac{1}{((c_l e)^2 \log^4 n)^{3 \log^4 n}} \\
&\le \frac{1}{n^2 \log n} \le \frac{1}{mn}.
\end{aligned}$$

Since $x \in [0,1]$, we can truncate $f(x)$ at degree $d_n^\star := 2(c_l e)^2 \log^4 n$ and *redefine* it as

$$f(x) = \sum_{s=0}^{d_n^\star} x^s \cdot C_s,$$

where C_s, as specified above, satisfies $|C_s| \lesssim \exp(\tilde{\Theta}(\log^4 n))$ and

$$C_s = \tau_n^s \sum_{t=0}^{\min\{s, 4c_l \log^2 n\}} h_{m+1}\left(\frac{t}{m}\right) \cdot \binom{m}{t}\binom{m-t}{s-t}(-1)^{s-t}.$$

This modification changes the value of $f(x)$ by at most $1/n$, for all $x \in [0,1]$.

Third-level truncation of $f(x)$ Now we evaluate each coefficient C_s to an error of $1/(nd_n^\star)$, so that we can compute $f(x)$ to an error of $1/n$, for all $x \in [0,1]$. This can be accomplished by computing every

$$C_{s,t} := h_{m+1}\left(\frac{t}{m}\right) \cdot \tau_n^s \binom{m}{t}\binom{m-t}{s-t}(-1)^{s-t}$$

to an $\mathcal{O}(1/(nsd_n^\star))$ absolute error. Note that $C_{s,t}$ is a product of five terms, with each of them bounded by $m^s \le \exp(\Theta(\log^5 n))$ in magnitude. Simple algebra further reduces our objective to approximating every term in the product to an $\exp(-\Theta(\log^5 n))$ error.

We analyze each term as follows: 1) computing $(-1)^{s-t}$ takes $\mathcal{O}(\max\{\log s, \log t\}) = \mathcal{O}(\log\log n)$ time; 2) computing the product of A integers of magnitude $\le B$ takes $\mathcal{O}((A \log B)^2)$ time, which can be achieved by recursively calculating the pairwise products [2]; 3) point 2) shows that we can compute $\binom{m}{t}$, $\binom{m-t}{s-t}$, and n^s *exactly* in $\mathrm{ploylog}(n)$ time; 4) now consider evaluating $(n\tau_n)^s = (c_l \log n)^s$: since $|a^s - b^s| \le |a-b| \cdot s\max\{|a|,|b|\}^{s-1} \le |a-b| \cdot \mathcal{O}(\log^5 n)$ if $|a|, |b| \le \mathcal{O}(\log n)$, it suffices to compute $c_l \log n$ to an $\exp(-\Theta(\log^5 n))$ error, which can be performed in $\mathrm{ploylog}(n)$ time; 5) it remains to compute

$$h_{m+1}\left(\frac{t}{m}\right) = \log(m+1) - (t+1)\log(t+1) + t\log t,$$

to an $\exp(-\Theta(\log^5 n))$ error, which again takes $\mathrm{ploylog}(n)$ time.

Therefore, we can evaluate each $C_{s,t}$, and their sum C_s, to an error of $1/(nd_n^\star)$ in time $\mathrm{ploylog}(n)$. We can further define $C_s^\star$ as the closest integer multiple of $1/(nd_n^\star)$ [3] to C_s, and *redefine*

$$f(x) = \sum_{s=0}^{d_n^\star} x^s \cdot C_s^\star.$$

9.3 Lagrange Interpolation with Chebyshev Nodes

Recall that the degree of the min-max approximation polynomial is $d = d_n = \Theta(\log n)$. We initialize the Remez Algorithm by the Chebyshev nodes:

$$x_i := \frac{1}{2} + \frac{1}{2}\cos\left(\frac{2i+1}{2(d+2)}\pi\right), i = 0, 1, \ldots, d_n + 1.$$

Then, for any integers $i \ne j \in [0, d+1]$,

$$\begin{aligned}
|x_i - x_j| &= \frac{1}{2}\left|\cos\left(\frac{2i+1}{2(d+2)}\pi\right) - \cos\left(\frac{2j+1}{2(d+2)}\pi\right)\right| \\
&= \left|\sin\left(\frac{i+j+1}{2(d+2)}\pi\right) \cdot \sin\left(\frac{i-j}{2(d+2)}\pi\right)\right| \\
&\ge \sin^2\left(\frac{\pi}{2(d+2)}\right) \ge \frac{1}{(d+2)^2}.
\end{aligned}$$

[2] We assume that computing the product two integers $\le B$ takes $\mathcal{O}(\log^2 B)$ time, achievable through the standard schoolbook "long multiplication". A more efficient integer-multiplication algorithm is the Harvey-Hoeven that takes only $\tilde{\mathcal{O}}(\log B)$ time, yielding an $\tilde{\mathcal{O}}(A \log B)$ complexity for the problem considered here.

[3] Assume that $d_n^\star$ is an integer. Otherwise, replace it by $\lceil d_n^\star \rceil$.

Now, consider the following function relating to the i-th Lagrange basis polynomial:

$$\ell_i(x) := \prod_{j\neq i}(x - x_j).$$

For any $\tau > 0$ and approximation sequence $\{x'_j\}_{j=0}^{d+1}$ in $[0,1]$ satisfying $|x_j - x'_j| \le \tau$, denote by $\tilde{\ell}_i(x)$ the corresponding product $\prod_{j\neq i}(x - x'_j)$. Then, for any $x \in [0,1]$,

$$\begin{aligned}
|\ell_i(x) - \tilde{\ell}_i(x)| &\le |\prod_{j\neq i}(x - x_j) - \prod_{j\neq i}(x - x'_j)| \\
&\le \sum_{j\neq i} |(x - x_j) - (x - x'_j)| \prod_{j'<j, j'\neq i} |x - x_{j'}| \prod_{j'>j, j'\neq i} |x - x'_{j'}| \\
&\le (d+1)\tau.
\end{aligned}$$

Under the same setting with $\tau < 1/(4(d+2)^2)$, the i-th Lagrange basis polynomial $L_i(x) := \ell_i(x)/\ell_i(x_i)$ and its approximation $\tilde{L}_i(x) := \tilde{\ell}_i(x)/\tilde{\ell}_i(x_i)$ differ by

$$\begin{aligned}
\left|L_i(x) - \tilde{L}_i(x)\right| &\le \left|\frac{\ell_i(x)}{\ell_i(x_i)} - \frac{\tilde{\ell}_i(x)}{\tilde{\ell}_i(x_i)}\right| \\
&= \left|\frac{\ell_i(x)\tilde{\ell}_i(x_i) - \tilde{\ell}_i(x)\ell_i(x_i)}{\ell_i(x_i)\tilde{\ell}_i(x_i)}\right| \\
&\le \left|(\tilde{\ell}_i(x_i) - \ell_i(x_i))\frac{\ell_i(x)}{\ell_i(x_i)\tilde{\ell}_i(x_i)}\right| + \left|(\ell_i(x) - \tilde{\ell}_i(x))\frac{\ell_i(x_i)}{\ell_i(x_i)\tilde{\ell}_i(x_i)}\right| \\
&\le \tau \cdot \exp(\tilde{\Theta}(\log n)).
\end{aligned}$$

Denote by $\mathcal{L}$ and $\tilde{\mathcal{L}}$ the Lagrange interpolation operator associated with $\{x_j\}_{j=0}^{d+1}$ and $\{x'_j\}_{j=0}^{d+1}$, respectively. Then for any $x \in [0,1]$, the interpolation polynomials of f differ by

$$\begin{aligned}
|\mathcal{L}[f](x) - \tilde{\mathcal{L}}[f](x)| &\le \sum_i |f(x_i)L_i(x) - f(x'_i)\tilde{L}_i(x)| \\
&\le \sum_i |(f(x_i) - f(x'_i))L_i(x) + f(x'_i)(L_i(x) - \tilde{L}_i(x))| \\
&\le \sum_i |L_i(x) \cdot \sum_{s=0}^{d_n^\star}(x_i^s - x_i'^s) \cdot C_s^\star| + \sum_i |f(x'_i)(L_i(x) - \tilde{L}_i(x))| \\
&\le \tau \cdot \exp(\tilde{\Theta}(\log^4 n)).
\end{aligned}$$

Set $\tau = \exp(-\tilde{\Theta}(\log^4 n))/n$ and recall that $E_d[g]$ denotes the best approximation error of the degree-d min-max polynomial over $[0,1]$. By the previous derivations and result of Ehlich & Zeller (1966), for $T_d := 2 + \frac{2}{\pi}\log(d+1)$ and any $x \in [0,1]$,

$$\begin{aligned}
|\tilde{\mathcal{L}}[f](x) - B'_m(h,x)| &\le \frac{1}{n} + |\mathcal{L}[f](x) - B'_m(h,x)| \\
&\le \frac{1}{n} + |\mathcal{L}[f](x) - \mathcal{L}[B'_m(h,\cdot)](x) + \mathcal{L}[B'_m(h,\cdot)](x) - B'_m(h,x)| \\
&\le \frac{1}{n} + T_d \cdot (E_d[B'_m(h,\cdot)] + E_d[f] + E_d[B'_m(h,\cdot)]) + |f(x) - B'_m(h,x)| \\
&\le \frac{1}{n} + 3T_d \cdot E_d[B'_m(h,\cdot)] + (T_d+1)\max_{x\in[0,1]} |f(x) - B'_m(h,x)| \\
&\lesssim T_d\left(\frac{1}{n} + E_d[B'_m(h,\cdot)]\right) \\
&\lesssim \varepsilon \cdot \log\log n.
\end{aligned}$$

Therefore, if we compute each x_j to an $\exp(-\tilde{\Theta}(\log^4 n))$ error, the resulting polynomial $\tilde{\mathcal{L}}[f](x)$ approximates $B'_m(h,x)$ to an error of $\mathcal{O}(\varepsilon \cdot \log\log n)$, for any $x \in [0,1]$. This yields a result only slightly weaker than that in Theorem 1, with the inequality being

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

Choose the approximation nodes $x'_j \in [0,1]$ to be integer multiples of $\exp(-\tilde{\Theta}(\log^4 n))$. Finally, we consider the time complexity of expanding $\tilde{\mathcal{L}}[f](x)$ into its standard form, which basically characterizes the time required for constructing the estimator. Note that

$$\tilde{\mathcal{L}}[f](x) = \sum_i f(x'_i) \cdot \frac{\prod_{j\neq i}(x - x'_j)}{\prod_{j\neq i}(x'_i - x'_j)}.$$

Since $x'_j \exp(\tilde{\Theta}(\log^4 n)) \in \mathbb{N}$ for any j and $f(x) = \sum_{s=0}^{d_n^\star} x^s \cdot C_s^\star$ with $C_s^\star$ being multiples of $1/(nd_n^\star)$, it takes $\mathrm{polylog}(n)$ time to evaluate $f(x'_i)$ and $\prod_{j\neq i}(x'_i - x'_j)$ exactly, with results expressed as rational numbers. In addition, computing each coefficient in the standard form of $\prod_{j\neq i}(x - x'_j)$ takes $\mathcal{O}(2^d \cdot s^2) = \tilde{\mathcal{O}}(\sqrt{n})$ [4] time. Hence, finding the explicit expression of the standard form of $\tilde{\mathcal{L}}[f](x)$ takes $\tilde{\mathcal{O}}(\sqrt{n}\log^2 n) = \tilde{\mathcal{O}}(\sqrt{n})$ time. Let us denote this standard form by

$$\tilde{\mathcal{L}}[f](x) := \sum_{t=0}^{d+1} b_t \cdot x^t.$$

The small probability estimator is thus

$$\hat{\mathcal{V}}_S := \sum_{i\in[k]} \left(\sum_{t=1}^{d+2} \frac{b_{t-1}}{t} \cdot \frac{N_i^{\underline{t}}}{n^{\underline{t}}} \right) \cdot \mathbb{1}_{N'_i \leq \frac{1}{\varepsilon}} \cdot \mathbb{1}_{N_i \lesssim \log n},$$

where N_i and N'_i are sample symbol counts in $[0,n]$. Note that computing each $N_i^{\underline{t}}$ or $n^{\underline{t}}$ takes $\mathcal{O}(\log^2 n)$ time, and there are at most $\mathcal{O}(\sqrt{n})$ distinct $(N_i, N'_i \lesssim 1/\varepsilon)$ pairs. Hence, we can evaluate the small-probability estimator in $\tilde{\mathcal{O}}(n)$ time. In addition, the evaluation of the large-probability estimator is essentially the same as that of the empirical plug-in estimator. Consolidating these facts yields the desired near-linear-time computability.

9.4 Remez Algorithm with High Precision

Note that the first step of the Remez algorithm is initialization and will be executed only once. The last step of the algorithm serves as the initialization step for the next round of iteration. Exact evaluation of the initial nodes is not required in each round for convergence.

As shown by our previous discussion, it suffices to approximate the initial nodes to an accuracy of $\exp(-\mathrm{polylog}(n))$, which takes $\mathrm{polylog}(n)$ time for the first step. Denote by $x'_0, \ldots x'_{d+1} \in [0,1]$ the initial nodes for a particular iteration and assume that $x'_i/\delta_n \in \mathbb{N}, i = 0, \ldots, d+1$.

We proceed to analyzing the second step of the Remez algorithm. According to Section 9.2, we will approximate the polynomial

$$f(x) = \sum_{s=0}^{d_n^\star} x^s \cdot C_s^\star,$$

where $d_n^\star = \Theta(\log^4 n)$ and $C_s^\star$'s are integer multiples of $1/(nd_n^\star)$ satisfying $|C_s^\star| \leq \exp(\tilde{\Theta}(\log^4 n))$. Computing the sequence of $f(x)$ values exactly for x'_i's takes $\mathrm{polylog}(n)$ time. We can express each $f(x'_i)$ as a rational number with both its nominator and denominator being at most $\exp(\mathrm{polylog}(n))$. These claims clearly also hold for the evaluation of x^t at each x'_i with $t, j < d+2 = \Theta(\log n)$. Denote by $V_{b,E} := (b_0, \ldots, b_d, E)^T$ the vector of unknown variables. Multiplying both sides of each equation

$$b_0 + b_1 \cdot x'_i + \ldots + b_d \cdot x_i'^d + E \cdot (-1)^i = f(x'_i)$$

by the least common multiple of the denominators of $x_i'^d$ and $f(x'_i)$, we transform the second step to solving a system of linear equations in the form $AV_{b,E} = y$, where $A \in \mathbb{Z}_+^{(d+2)\times(d+2)}$ and $y \in \mathbb{Z}_+^{(d+2)\times 1}$ are matrices with entries bounded by $\exp(\mathrm{polylog}(n))$. If the initial nodes x'_j's are distinct and sorted accordingly, the system $AV_{b,E} = y$ has a unique solution. Utilizing the algorithm proposed by Dixon (1982), we can solve this system in time $\tilde{\mathcal{O}}((d+2)^3 \log(\|A\| + \|y\|)) = \mathrm{polylog}(n)$ where $\|\cdot\|$ represents the maximum entry in absolute value.

[4] Recall that $d = c_s \log n$. Here we choose $c_s \leq 1/2$.

Once we obtain the coefficient vector $V_{b,E}$, Step 3 of the algorithm takes $\mathrm{polylog}(n)$ time to form the approximation polynomial

$$P(x) \coloneqq b_0 + b_1 \cdot x + ... + b_d \cdot x^d.$$

The fourth step of the Remez algorithm calls for computing the local extrema of the error function

$$\mathcal{E}(x) \coloneqq P(x) - f(x)$$

over the $d + 2$ sign-invariant regions. Noting that $\mathcal{E}(x)$ is a degree $d_n^\star$ polynomial, it suffices to approximate all the real roots of its derivative $\mathcal{E}'(x)$ to an $\exp(-\mathrm{polylog}(n))$ accuracy.

To do this, we first transform $\mathcal{E}'(x)$ to a polynomial with integer coefficients of size $\exp(\mathrm{polylog}(n))$. Then, we apply the quadratic interval refinement algorithm (Abbott, 2014) to approximate the real roots of the transformed polynomial. Shown in the paper of Kerber (2009), for a degree-d square-free polynomial with integer coefficients bounded by 2^σ in absolute value, an ε-accuracy approximation of the real roots using this algorithm requires a time complexity of $\tilde{\mathcal{O}}(d^4\sigma^2 + d^3 \log(1/\varepsilon))$. For the task considered here, this again converts to a time complexity of $\mathrm{polylog}(n)$.

Finally, we can view Step 5 as the initialization step in the next iteration, implying a per-iteration complexity of $\mathrm{polylog}(n)$ for the Remez algorithm. Note that quantity E corresponds to a lower bound on the max approximation error of each iteration. As for the number of iterations, Veidinger (1960) essentially shows that under differentiability, this process has a quadratic convergence. More specifically, let E_ν denote the error bound E of the the ν-th iteration, then $\{E_\nu\}_{\nu \geq 1}$ converges to the optimal degree-d approximation error $E_d[f]$ with

$$|E_d[f] - E_\nu| \lesssim (E_d[f] - E_{\nu-1})^2.$$

It takes only $\mathrm{polylog}(n)$ iterations for E to converge to the $\exp(-\mathrm{polylog}(n))$-neighborhood of its limit $E_d[f]$. Therefore, the total time required for computing the approximation polynomial with Remez algorithm is $\mathcal{O}(\mathrm{polylog}(n))$. Consolidating this with the reasoning in the last section shows that our estimator can be evaluated in time near-linear in n. On the practical side, see Pachón & Trefethen (2009); Trefethen (2013) for an optimized Matlab implementation of the Remez algorithm.

A A Refined Estimator for Shannon Entropy

In this section, we replacing the function $h_n(x)$ employed in Section 4 by a much finer approximation of $B_n(h,x)$. Through this refinement, we establish the full version of Theorem 1. To begin with, we define the following two *f-functions* for $z \in [0,\infty]$:

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

and

$$f_2(z) := \mathop{\mathbb{E}}_{X\sim\mathrm{Poi}(z)}[h(X+1)] = -e^{-z}\sum_{j=1}^{\infty}\frac{z^j}{j!}(j+1)\log(j+1).$$

A.1 Relating f-functions to Bernstein Approximation Errors

For $x \in [0,1]$, set $z = z(x) := nx$. The following lemma relates $f_1(z)$ and $f_2(z)$ to the Bernstein approximation error of h_{n+1}, that is, $h_{n+1}(x) - B_n(h_{n+1},x)$.

Lemma 9. *For any $x \in [0,\log^4 n/n]$,*

$$h_{n+1}(x) - B_n(h_{n+1},x) = (h(z+1) - f_2(z)) - (h(z) - f_1(z)) + \tilde{\mathcal{O}}\left(\frac{1}{n}\right).$$

As a corollary, for any sufficiently large n and $x \in I_n = [0,\tau_n := c_l(\log n)/n]$,

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

Since $1/(na-1) \le \min\{1/\log n, S_p/n\}$, 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 desired inequality

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

to bounds in the form of

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

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

Proof. Let $h_{-1}(x) := h(x+n^{-1})$. By the linearity of expectation,

$$\begin{aligned} h_{n+1}(x) - B_n(h_{n+1},x) &= n\left(h_{-1}(x) - h(x) - B_n(h_{-1},x) + B_n(h,x)\right) \\ &= n\left(h_{-1}(x) - B_n(h_{-1},x)\right) - n\left(h(x) - B_n(h,x)\right). \end{aligned}$$

Note that $z = nx$ implies $z \in [0,\log^4 n]$. Hence, we have

$$\begin{aligned} n\left(h_{-1}(x) - B_n(h_1,x)\right) &= -(nx+1)\log\left(\frac{nx+1}{n}\right) + \sum_{j=0}^{n}(j+1)\log\left(\frac{j+1}{n}\right)\binom{n}{j}x^j(1-x)^{n-j} \\ &= -(z+1)\log\left(\frac{z+1}{n}\right) + \sum_{j=0}^{n}(j+1)\log\left(\frac{j+1}{n}\right)\binom{n}{j}z^j\frac{(n-z)^{n-j}}{n^n} \\ &= -(z+1)\log(z+1) + \left(1-\frac{z}{n}\right)^n\sum_{j=0}^{n}(j+1)\log(j+1)\binom{n}{j}z^j(n-z)^{-j} \\ &= -(z+1)\log(z+1) + \left(1-\frac{z}{n}\right)^n\sum_{j=0}^{n}(j+1)\log(j+1)\frac{n^{\underline{j}}}{n^j}\frac{z^j}{j!}\left(1-\frac{z}{n}\right)^{-j} \\ &= -(z+1)\log(z+1) + e^{-z}\sum_{j=0}^{\infty}\frac{z^j}{j!}(j+1)\log(j+1) + \tilde{\mathcal{O}}\left(\frac{1}{n}\right) \\ &= h(z+1) - f_2(z) + \tilde{\mathcal{O}}\left(\frac{1}{n}\right). \end{aligned}$$

The second last equality is the most non-trivial step. In order to establish this equality, we will need the following three inequalities (assume $z \in [0,\log^4 n]$ and $n \gg 1$).

Inequality 1:

$$\begin{aligned}
0 &\le \left(1-\frac{z}{n}\right)^n \sum_{j=\log^5 n+1}^{n} (j+1)\log(j+1)\,\frac{n^{\underline{j}}}{n^j}\frac{z^j}{j!}\left(1-\frac{z}{n}\right)^{-j} \\
&= \left(1-\frac{z}{n}\right)^n \sum_{j=\log^5 n+1}^{n} (j+1)\log(j+1)\,\frac{n^{\underline{j}}}{2^j(n-z)^j}\frac{(2z)^j}{j!} \\
&\le e^{-z} \sum_{j=\log^5 n+1}^{n} (j+1)\log(j+1)\,\frac{(2z)^j}{j!} \\
&\le e^{-z} \sum_{j=\log^5 n+1}^{n} 2j(j-1)\frac{(2z)^j}{j!} \\
&\le 8z^2 e^{-z} \sum_{j=\log^5 n-1}^{n} \frac{(2z)^j}{j!} \\
&\le 8(\log^8 n)\Pr(\mathrm{Poi}(2z)\ge \log^5 n-1) \\
&\le \frac{1}{n}.
\end{aligned}$$

Inequality 2:

$$0 \le e^{-z} \sum_{j=\log^5 n+1}^{\infty} \frac{z^j}{j!}(j+1)\log(j+1) = 2(\log^8 n)\Pr(\mathrm{Poi}(2z)\ge \log^5 n-1) \le \frac{1}{n}.$$

Inequality 3: For any $j \le \log^5 n$,

$$\begin{aligned}
\left|e^{-z}-\left(1-\frac{z}{n}\right)^n \frac{n^{\underline{j}}}{n^j}\left(1-\frac{z}{n}\right)^{-j}\right| &= \left|e^{-z}-\left(1-\frac{z}{n}\right)^n \frac{n^{\underline{j}}}{(n-z)^j}\right| \\
&\le \left|e^{-z}-\left(1-\frac{z}{n}\right)^n\right| + \left(1-\frac{z}{n}\right)^n \left|1-\frac{n^{\underline{j}}}{(n-z)^j}\right| \\
&\le e^{-z}\frac{z^2}{n} + e^{-z}\left|1-\frac{n^{\underline{j}}}{(n-z)^j}\right| \\
&\le e^{-z}\frac{z^2}{n} + e^{-z}\left(\left|1-\frac{n^j}{(n-z)^j}\right| \vee \left|1-\frac{(n-\log^5 n)^j}{(n-z)^j}\right|\right) \\
&\le e^{-z}\frac{z^2}{n} + e^{-z}\left(\left|\exp\left(\frac{zj}{n-z}\right)-1\right| \vee \left|\frac{(\log^5 n - z)j}{n-z}\right|\right) \\
&\le e^{-z}\frac{z^2}{n} + e^{-z}\left(\left|\frac{zj}{n-z(j+1)}\right| \vee \left|\frac{(\log^5 n)j}{n-z}\right|\right) \\
&\le e^{-z}\frac{2\log^{10} n}{n}.
\end{aligned}$$

Note that Inequality 3 further implies

$$\begin{aligned}
&\left|e^{-z}\sum_{j=0}^{\log^5 n}\frac{z^j}{j!}(j+1)\log(j+1) - \left(1-\frac{z}{n}\right)^n \sum_{j=0}^{\log^5 n}(j+1)\log(j+1)\,\frac{n^{\underline{j}}}{n^j}\frac{z^j}{j!}\left(1-\frac{z}{n}\right)^{-j}\right| \\
&\le \frac{2\log^{10} n}{n}\cdot e^{-z}\sum_{j=0}^{\log^5 n}\frac{z^j}{j!}(2j(j-1)) \\
&\le \frac{2\log^{10} n}{n}\cdot 2z^2 \\
&\le \frac{4\log^{18} n}{n}.
\end{aligned}$$

This, together with Inequality 1 and 2, proves the desired equality. The same reasoning also gives

$$n\left(h(x) - B_n(h,x)\right) = -z\log z + e^{-z}\sum_{j=1}^{\infty}\frac{z^j}{j!}j\log j + \tilde{\mathcal{O}}\left(\frac{1}{n}\right),$$

which completes the proof. □

For any $x \in I_n$, let $z_1 = (na-1)x$, then $z_1 \in I_n' := [0, ac_l \log n]$. Therefore, by Lemma 9,

$$h_{na}(x) - B_{na-1}(h_{na}, x) = \left(h(z_1+1) - f_2(z_1)\right) - \left(h(z_1) - f_1(z_1)\right) + \tilde{\mathcal{O}}\left(\frac{1}{n}\right).$$

In the next section, we approximate function $f_1(z)$ over I_n' with a degree-d polynomial.

A.2 Approximating $f_1(z)$

Consider the first function

$$f_1(z) = -e^{-z}\sum_{j=1}^{\infty}\frac{z^j}{j!}j\log j.$$

We want to approximate f_1 with a low-degree polynomial and bound the corresponding error. For this purpose, we establish some basic properties of $f_1(z)$ as follows.

A.2.1 Properties of $f_1(z)$

Property 1: The function $f_1(z)$ is a continuous function over $[0,\infty)$, and $f_1(0) = 0$.

Property 2: For all $z \geq 0$, the value of $f_1(z)$ is non-negative.

Property 3: Denote $u(y) := (y+2)\log(y+2) + y\log y - 2(y+1)\log(y+1)$. Then, for any $z \geq 0$,

$$f_1{}''(z) = -e^{-z}\sum_{t=0}^{\infty}\frac{z^t}{t!}\cdot u(t) \text{ and } -\log 4 \leq f_1{}''(z) < 0.$$

Proof. We begin by establishing the equality.

$$\begin{aligned}
-f_1{}''(z) &= e^{-z}\sum_{t=1}^{\infty}\frac{(t-1)t^2z^{t-2}\log(t)}{t!} - 2e^{-z}\sum_{t=1}^{\infty}\frac{t^2z^{t-1}\log(t)}{t!} + e^{-z}\sum_{t=1}^{\infty}\frac{tz^t\log(t)}{t!} \\
&= e^{-z}\sum_{t=0}^{\infty}\frac{z^t(t+2)\log(t+2)}{t!} - 2e^{-z}\sum_{t=0}^{\infty}\frac{z^t(t+1)\log(t+1)}{t!} + e^{-z}\sum_{t=0}^{\infty}\frac{tz^t\log(t)}{t!} \\
&= e^{-z}\sum_{t=0}^{\infty}\frac{z^t}{t!}\cdot u(t).
\end{aligned}$$

To prove the inequality, we utilize the following lemma.

Lemma 10. *For any $t \geq 0$,*

$$\frac{\log 4}{t+1} \geq u(t) \geq \frac{1}{t+1}.$$

By Lemma 10, we obtain

$$0 < e^{-z}\sum_{t=0}^{\infty}\frac{z^t}{t!}\cdot\frac{1}{t+1} \leq e^{-z}\sum_{t=0}^{\infty}\frac{z^t}{t!}\cdot u(t) = -f_1{}''(z) \leq e^{-z}\sum_{t=0}^{\infty}\frac{z^t}{t!}\cdot\frac{\log 4}{t+1} = (\log 4)\frac{1-e^{-z}}{z} \leq \log 4.$$

The proof of the lemma follows by standard algebraic calculations and is omitted. □

Property 4: For $z > 0$,

$$0 \le \frac{f_1''(z)}{h''(z)} \le \log 4.$$

Proof. Recall that $h(z) = -z \log z$. Therefore, $h''(z) = -1/z$ and

$$\begin{aligned} 0 &\le \frac{f_1''(z)}{h''(z)} \\ &= e^{-z} \sum_{t=0}^{\infty} \frac{z^{t+1}}{t!} \cdot u(t) \\ &\le e^{-z} \sum_{t=0}^{\infty} \frac{z^{t+1}}{t!} \cdot \frac{\log 4}{t+1} \\ &\le (\log 4)(1 - e^{-z}) \\ &\le \log 4, \end{aligned}$$

where the third step follows by Lemma 10. □

A.2.2 Moduli of Smoothness

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

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

and

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

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

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

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 $E_d[f] \coloneqq E_d[f, [0,1]]$.

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 11. *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 $E_\cdot[f]$, 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.

A.2.3 Bounding Errors in Approximating $f_1(x)$

For simplicity, define $x' := (ac_l \log n) \cdot x$ and consider the function

$$f_{1'}(x) := f_1((ac_l \log n) \cdot x).$$

Under proper scaling, approximating $f_1(x')$ over $I_n' = [0, ac_l \log n]$ is equivalent to approximating $f_{1'}(x)$ over $[0,1]$. By Lemma 11, it suffices to bound $w_\varphi^2(f_{1'}, \cdot)$ for our purpose.

In particular, we know that

$$\min_{g \in \boldsymbol{P}_d} \max_{x \in I_n'} |f_1(x) - g(x)| = E_d[f_{1'}] \le C_1 w_\varphi^2(f_{1'}, d^{-1}).$$

By definition, $w_\varphi^2(f_{1'}, d^{-1})$ is the solution to the following optimization problem.

$$\sup_{u,v} \left| f_{1'}(u) + f_{1'}(v) - 2f_{1'}\left(\frac{u+v}{2}\right)\right|$$

subject to

$$0 \le u, v \le 1, |u - v| \le \frac{2}{d} \cdot \varphi\left(\frac{u+v}{2}\right).$$

First, consider the optimization constraints. Analogous to the arguments in Jiao et al. (2015), we define $M := (u+v)/2$ and $\delta := d^{-1}\sqrt{1/M - 1}$. The feasible region can be expressed as

$$[M - d^{-1}\sqrt{M(1-M)}, M + d^{-1}\sqrt{M(1-M)}] \cap [0,1] = [M - \delta M, M + \delta M] \cap [0,1].$$

By Property 3 in Section A.2.1, $f_1(x')$, or equivalently $f_{1'}(x)$, is a strictly concave function. Therefore, the maximum of $|f(u) + f(v) - 2f(u+v/2)|$ is attained at the boundary of the feasible region.

Note that

$$M - d^{-1}\sqrt{M(1-M)} \ge 0 \iff M \ge \frac{1}{d^2+1}$$

and

$$M + d^{-1}\sqrt{M(1-M)} \le 1 \iff M \le \frac{d^2}{d^2+1}.$$

We need to consider only three cases:

Case 1:

$$u = 0, v = 2M, M \in [0, 1/(d^2+1)].$$

Case 2:

$$u = 2M - 1, v = 1, M \in [d^2/(d^2+1), 1].$$

Case 3:

$$u = M - \delta M, v = M + \delta M, M \in [1/(d^2+1), d^2/(d^2+1)].$$

To facilitate the discussions, we utilize the following lemma.

Lemma 12. *Let $f \in C^1([a,b])$ have second order derivative in (a,b). There exists $c \in (a,b)$ such that*

$$f(a) + f(b) - 2f\left(\frac{a+b}{2}\right) = \frac{1}{4}(b-a)^2 \cdot f''(c).$$

We begin with Case 1. By the Lemma 12, there exists $c \in (0, 2/(d^2+1))$ satisfying

$$\left| f_{1'}(0) + f_{1'}\left(\frac{2}{d^2+1}\right) - 2f_{1'}\left(\frac{1}{d^2+1}\right)\right| \le \frac{1}{4} \cdot \left(\frac{2}{d^2+1}\right)^2 |f_{1'}''(c)| = \left(\frac{1}{d^2+1}\right)^2 |f_{1'}''(c)|.$$

By the definition of function $f_{1'}$,

$$|f_{1'}''(x)| = |(ac_l \log n)^2 g_1''((ac_l \log n) \cdot x)| \le (\log 4)(ac_l \log n)^2.$$

Therefore, we obtain

$$\left(\frac{1}{d^2+1}\right)^2 |f_{1'}''(c)| \lesssim \varepsilon^2.$$

This, together with an analogous argument on Case 2, implies that the objective value is bounded by $\mathcal{O}(\varepsilon^2)$ in both cases. It remains to analyze Case 3. We proceed by considering two regimes:

Regime 1: If $M \le 4/(d^2+1)$, then $|u-v| = 2d^{-1}\sqrt{M(1-M)} \le 4/d^2$. The above reasoning again shows that

$$\left| f_{1'}(u) + f_{1'}(v) - 2f_{1'}\left(\frac{u+v}{2}\right)\right| \lesssim \varepsilon^2.$$

Regime 2: If $4/(d^2+1) \le M \le d^2/(d^2+1)$,

$$M - \delta M = M\left(1 - \frac{\sqrt{M^{-1}-1}}{d}\right) \ge M\left(1 - \frac{\sqrt{(d^2+1)-4}}{2d}\right) \ge \frac{M}{2}.$$

By Lemma 12, there exists $c \in (M-\delta M, M+\delta M) \subseteq (M/2, 3M/2)$ satisfying

$$\left| f_{1'}(u) + f_{1'}(v) - 2f_{1'}\left(\frac{u+v}{2}\right)\right| \le \frac{1}{4}\cdot\left(2\frac{1}{d}\sqrt{M(1-M)}\right)^2 \left|f_{1'}{}''(c)\right|.$$

Then, by Property 4 in Section A.2.1,

$$|f_{1'}{}''(c)| = |(ac_l \log n)^2 f_1''((ac_l \log n)\cdot c)| \le (ac_l \log n)^2 (\log 4)\cdot \frac{1}{(ac_l \log n)\cdot c} \le (\log 8)\cdot \frac{ac_l \log n}{M}.$$

This bound immediately implies

$$\frac{1}{4}\cdot\left(\frac{2}{d}\sqrt{M(1-M)}\right)^2 \cdot \left|f_{1'}{}''(c)\right| \le \frac{1}{d^2}M(1-M)\cdot(\log 8)\cdot\frac{ac_l \log n}{M} \le (\log 8)\cdot\frac{c_l \varepsilon}{c_s^2}.$$

Consolidating the previous results yields

$$\min_{g\in \boldsymbol{P}_d}\max_{x\in I_n'} |f_1(x) - g(x)| \lesssim \varepsilon.$$

For function f_2, an analogous argument shows that

$$\min_{g\in \boldsymbol{P}_d}\max_{x\in I_n'} |f_2(x) - g(x)| \lesssim \varepsilon.$$

In the next section, we apply these inequalities to analyze our refined entropy estimator.

A.3 Proving Theorem 1: A Refined Entropy Estimator

We aim to approximate $B_{na-1}(h_{na}, x) - h_{na}(x)$ over $I_n = [0, c_l \log n/n]$ by a degree-d polynomial. By Lemma 9, for any $x \in I_n$ and $z_1 := (na-1)x \in I_n' = [0, ac_l \log n]$,

$$h_{na}(x) - B_{na-1}(h_{na}, x) = (h(z_1+1) - f_2(z_1)) - (h(z_1) - f_1(z_1)) + \tilde{\mathcal{O}}\left(\frac{1}{n}\right).$$

By the results in Korněichuk (1991),

$$\min_{g\in \boldsymbol{P}_d}\max_{x\in I_n'} |h(x) - g(x)| = (ac_l \log n)\min_{g\in \boldsymbol{P}_d}\max_{x\in[0,1]} |h(x) - g(x)| \lesssim \frac{ac_l \log n}{(c_s \log n)^2} \lesssim \varepsilon$$

and

$$\min_{g\in \boldsymbol{P}_d}\max_{x\in I_n'} |h(x+1) - g(x)| \lesssim \varepsilon.$$

Combining these bounds with the last two inequalities in the previous section, we obtain

$$\min_{g\in \boldsymbol{P}_{d-1}}\max_{x\in I_n} |(h_{na}(x) - B_{na-1}(h_{na}, x)) - g(x)| \lesssim \varepsilon.$$

Denote by $\tilde{g}(x)$ the min-max polynomial that achieves this minimal error. By the derivations in Section 4.2, the degree-$(d-1)$ polynomial $\tilde{h}_{na}(x)$ satisfies

$$\max_{x\in I_n} |h_{na}(x) - \tilde{h}_{na}(x)| \lesssim \varepsilon.$$

Denote $\tilde{h}^*(x) := -\tilde{g}(x) + \tilde{h}_{na}(x)$, and note that by definition, $B'_{na}(h, x) = B_{na-1}(h_{na}, x)$. Then, the triangle inequality implies

$$\max_{x\in I_n} |B'_{na}(h, x) - \tilde{h}^*(x)| = \max_{x\in I_n} |B_{na-1}(h_{na}, x) - \tilde{h}^*(x)| \lesssim \varepsilon.$$

By the triangle inequality of integrals, the degree-d polynomial

$$\tilde{H}^*(x) := \int_0^x \tilde{h}^*(t)dt$$

approximating $B_{na}(h, x)$ possesses the following pointwise error guarantee.

Lemma 13. *For any $x \in I_n$,*

$$|B_{na}(h, x) - \tilde{H}^*(x)| \lesssim x\varepsilon.$$

Hence, $\tilde{H}^*(x)$ is a degree-d polynomial that well approximates $B_{na}(h, x)$ pointwisely.

Next, we argue that the coefficients of $\tilde{H}^*(x)$ can not be too large. For notational convenience, write $\tilde{h}^*(x) := \sum_{v=0}^{d-1} a_v x^v$. By Corollary 2, for any $x \in I_n$,

$$|h_{na}(x) - B_{na-1}(h_{na}, x)| \le 1.$$

Furthermore, $h_{na}(x)$ is an increasing function over I_n, and thus

$$|h_{na}(x)| = \max\left\{|h_{na}(0)|, h_{na}\left(\frac{c_l(\log n)}{n}\right)\right\} \lesssim \log n.$$

Therefore, for any $x \in I_n$,

$$|\tilde{h}^*(x)| \lesssim \log n.$$

The boundedness of $\tilde{h}^*(x)$ implies that its coefficients cannot be too large:

$$|a_v| \lesssim \left(2^{4.5d} \log n\right)\left(\frac{n}{c_l \log n}\right)^v.$$

Write $\tilde{H}^*(x)$ as $\tilde{H}^*(x) = \sum_{t=1}^{d} a'_t x^t$. Then, by $\tilde{H}^*(x) = \int_0^x \tilde{h}^*(t)dt$ and the bound on $|a_v|$,

$$|a'_t| \lesssim 2^{4.5d}\left(\frac{n}{c_l \log n}\right)^{t-1}.$$

The construction of the new entropy estimator follows by replacing $\tilde{H}_{na}(x)$ by $\tilde{H}^*(x)$ in Section 5. The rest of the proof is also similar to that in the main paper and thus omitted.

B Competitive Estimators for General Additive Properties

Consider an arbitrary real function $f : [0,1] \to \mathbb{R}$. Without loss of generality, we will assume that $f(0) = 0$. According to the derivations in Section 4, we can write $B'_n(f,x)$ as

$$B'_n(f,x) := \sum_{j=0}^{n-1} n\left(f\left(\frac{j+1}{n}\right) - f\left(\frac{j}{n}\right)\right)\binom{n-1}{j}x^j(1-x)^{(n-1)-j}.$$

Our aim to approximate $B'_{na}(f,x)$ with a low degree polynomial. For simplicity, we assume that f is a 1-Lipschitz function. For $x \in [0,1]$, set $z = nx$, and define $g_{n+1}(j) := (n+1)f\left(\frac{j}{n+1}\right)$,

$$f_{1,n+1}(z) := e^{-z}\sum_{j=0}^{\infty} g_{n+1}(j+1)\frac{z^j}{j!},$$

and

$$f_{2,n+1}(z) := e^{-z}\sum_{j=0}^{\infty} g_{n+1}(j)\frac{z^j}{j!}.$$

The following lemma relates $f_{1,n+1}(z)$ and $f_{2,n+1}(z)$ to $B'_{n+1}(f,x)$.

Lemma 14. *For any $x \in [0, \log^4 n/n]$ and $z = nx$,*

$$B'_{n+1}(f,x) = f_{1,n+1}(z) - f_{2,n+1}(z) + \tilde{\mathcal{O}}\left(\frac{1}{n}\right).$$

Proof. Note that $z = nx$ implies $z \in [0, \log^4 n]$. Hence, we have

$$\begin{aligned}
\sum_{j=0}^{n}(n+1)f\left(\frac{j+1}{n+1}\right)\binom{n}{j}x^j(1-x)^{n-j} &= \sum_{j=0}^{n} g_{n+1}(j+1)\binom{n}{j}z^j\frac{(n-z)^{n-j}}{n^n} \\
&= \left(1-\frac{z}{n}\right)^n\sum_{j=0}^{n} g_{n+1}(j+1)\binom{n}{j}z^j(n-z)^{-j} \\
&= \left(1-\frac{z}{n}\right)^n\sum_{j=0}^{n} g_{n+1}(j+1)\frac{n^{\underline{j}}}{n^j}\frac{z^j}{j!}\left(1-\frac{z}{n}\right)^{-j} \\
&= e^{-z}\sum_{j=0}^{\infty} g_{n+1}(j+1)\frac{z^j}{j!} + \tilde{\mathcal{O}}\left(\frac{1}{n}\right) \\
&= f_{1,n+1}(z) + \tilde{\mathcal{O}}\left(\frac{1}{n}\right).
\end{aligned}$$

The second last equality is the most non-trivial step. In order to establish this equality, we will need the following three inequalities (assume $z \in [0, \log^4 n]$ and $n \gg 1$).

Inequality 1:

$$\begin{aligned}
0 &\le \left(1-\frac{z}{n}\right)^n \sum_{j=\log^5 n+1}^{n} |g_{n+1}(j+1)|\frac{n^{\underline{j}}}{n^j}\frac{z^j}{j!}\left(1-\frac{z}{n}\right)^{-j} \\
&= \left(1-\frac{z}{n}\right)^n \sum_{j=\log^5 n+1}^{n} (j+1)\frac{n^{\underline{j}}}{2^j(n-z)^j}\frac{(2z)^j}{j!} \\
&\le e^{-z}\sum_{j=\log^5 n+1}^{n} (j+1)\frac{(2z)^j}{j!} \\
&\le e^{-z}\sum_{j=\log^5 n+1}^{n} 2j(j-1)\frac{(2z)^j}{j!} \\
&\le 8z^2e^{-z}\sum_{j=\log^5 n-1}^{n}\frac{(2z)^j}{j!} \\
&\le 8(\log^8 n)\Pr(\mathrm{Poi}(2z) \ge \log^5 n - 1) \\
&\le \frac{1}{n}.
\end{aligned}$$

Inequality 2:

$$0 \le e^{-z} \sum_{j=\log^5 n+1}^{\infty} |g_{n+1}(j+1)| \frac{z^j}{j!} \le e^{-z} \sum_{j=\log^5 n+1}^{\infty} (j+1) \frac{z^j}{j!} \le \frac{1}{n}.$$

Inequality 3: For any $j \le \log^5 n$,

$$\begin{aligned}
\left| e^{-z} - \left(1 - \frac{z}{n}\right)^n \frac{n^{\underline{j}}}{n^j} \left(1 - \frac{z}{n}\right)^{-j} \right| &= \left| e^{-z} - \left(1 - \frac{z}{n}\right)^n \frac{n^{\underline{j}}}{(n-z)^j} \right| \\
&\le \left| e^{-z} - \left(1 - \frac{z}{n}\right)^n \right| + \left(1 - \frac{z}{n}\right)^n \left| 1 - \frac{n^{\underline{j}}}{(n-z)^j} \right| \\
&\le e^{-z} \frac{z^2}{n} + e^{-z} \left| 1 - \frac{n^{\underline{j}}}{(n-z)^j} \right| \\
&\le e^{-z} \frac{z^2}{n} + e^{-z} \left(\left| 1 - \frac{n^j}{(n-z)^j} \right| \vee \left| 1 - \frac{(n - \log^5 n)^j}{(n-z)^j} \right| \right) \\
&\le e^{-z} \frac{z^2}{n} + e^{-z} \left(\left| \exp\left(\frac{zj}{n-z}\right) - 1 \right| \vee \left| \frac{(\log^5 n - z)j}{n-z} \right| \right) \\
&\le e^{-z} \frac{z^2}{n} + e^{-z} \left(\left| \frac{zj}{n - z(j+1)} \right| \vee \left| \frac{(\log^5 n) j}{n-z} \right| \right) \\
&\le e^{-z} \frac{2 \log^{10} n}{n}.
\end{aligned}$$

Note that Inequality 3 further implies

$$\begin{aligned}
&\left| e^{-z} \sum_{j=0}^{\log^5 n} \frac{z^j}{j!} g_{n+1}(j+1) - \left(1 - \frac{z}{n}\right)^n \sum_{j=0}^{\log^5 n} g_{n+1}(j+1) \frac{n^{\underline{j}}}{n^j} \frac{z^j}{j!} \left(1 - \frac{z}{n}\right)^{-j} \right| \\
&\le \frac{2 \log^{10} n}{n} \cdot e^{-z} \sum_{j=0}^{\log^5 n} \frac{z^j}{j!} (j+1) \\
&\le \frac{2 \log^{10} n}{n} \cdot (1 + 2z) \\
&\le \frac{5 \log^{14} n}{n}.
\end{aligned}$$

This, together with Inequality 1 and 2, proves the desired equality. The same reasoning also gives

$$\sum_{j=0}^{n} (n+1) f\left(\frac{j}{n+1}\right) \binom{n}{j} x^j (1-x)^{n-j} = f_{2,n+1}(z) + \tilde{\mathcal{O}}\left(\frac{1}{n}\right),$$

which completes the proof. □

By slightly abusing the notation, we redefine $z := (na - 1)x$. Lemma 14 immediately implies that for any $x \in I_n = [0, c_l(\log n)/n] \subseteq [0, (\log^4(na-1))/(na-1)]$,

$$B'_{na}(f, x) = f_{1,na}(z) - f_{2,na}(z) + \tilde{\mathcal{O}}\left(\frac{1}{na}\right).$$

Note that $z \in I'_n = [0, ac_l \log n]$ in this case. Define $t_{na}(z) := f_{1,na}(z) - f_{2,na}(z)$ and $r_{na}(j) := g_{na}(j+2) + g_{na}(j) - 2g_{na}(j+1)$. Then, direct calculation yields

$$\begin{aligned}
t''_{na}(z) &= e^{-z} \sum_{j=0}^{\infty} r_{na}(j+1) \frac{z^j}{j!} - e^{-z} \sum_{j=0}^{\infty} r_{na}(j) \frac{z^j}{j!} \\
&= e^{-z} \sum_{j=0}^{\infty} r_{na}(j+1) \frac{z^j}{j!} - e^{-z} r_{na}(0) - \sum_{j=0}^{\infty} r_{na}(j+1) \frac{z^{j+1}}{(j+1)!} \\
&= e^{-z} \sum_{j=0}^{\infty} r_{na}(j+1) \left(\frac{z^j}{j!} - \frac{z^{j+1}}{(j+1)!} \right) - e^{-z} r_{na}(0).
\end{aligned}$$

Since f is 1-Lipschitz, we obtain $|r_{na}(j)| \le 2$. Therefore, for any $z \in I_n'$,

$$|t_{na}''(z)| \le e^{-z} \sum_{j=0}^{\infty} |r_{na}(j+1)| \left(\frac{z^j}{j!} + \frac{z^{j+1}}{(j+1)!} \right) + e^{-z} |r_{na}(0)| \le 6.$$

We can bound each summand in the expression of t_{na}'' by the following lemma.

Lemma 15. *For any $j \ge 1$ and $z \ge 0$, we have*

$$\left| e^{-z} \left(\frac{z^j}{j!} - \frac{z^{j+1}}{(j+1)!} \right) \right| \le \frac{1}{\sqrt{2\pi}((j+1) - \sqrt{j+1})}$$

and

$$\left| e^{-z} \left(\frac{z^j}{j!} - \frac{z^{j+1}}{(j+1)!} \right) \right| \le \frac{5}{z}.$$

Proof. For the ease of exposition, denote

$$q_1(z) := e^{-z} \left(\frac{z^j}{j!} - \frac{z^{j+1}}{(j+1)!} \right).$$

Then, the derivative of $q_1(z)$ is

$$\begin{aligned} q_1'(z) &= -e^{-z} \frac{z^j}{j!} + e^{-z} \frac{z^{j-1}}{(j-1)!} + e^{-z} \frac{z^{j+1}}{(j+1)!} - e^{-z} \frac{z^j}{j!} \\ &= e^{-z} \frac{z^{j-1}}{(j+1)!} \left(-2(j+1)z + j(j+1) + z^2 \right). \end{aligned}$$

Set $q_1'(z) = 0$ and note that $q_1(0) = \lim_{z \to \infty} q_1(z) = 0$. Hence, the maximum of $|q_1(z)|$ is attained at either $z_1 := (j+1) - \sqrt{j+1}$ or $z_2 := (j+1) + \sqrt{j+1}$. We first consider the function value at z_1:

$$\begin{aligned} |q_1(z_1)| &= e^{-z_1} \frac{z_1^{j+1}}{(j+1)!} \left| \frac{j+1}{z_1} - 1 \right| \\ &\le e^{-(j+1)+\sqrt{j+1}} ((j+1) - \sqrt{j+1})^{j+1} \frac{e^{j+1}}{\sqrt{2\pi}(j+1)^{j+1+1/2}} \frac{1}{\sqrt{j+1} - 1} \\ &\le e^{\sqrt{j+1}} \left(1 - \frac{1}{\sqrt{j+1}} \right)^{j+1} \frac{1}{\sqrt{2\pi}\sqrt{j+1}} \frac{1}{\sqrt{j+1} - 1} \\ &\le \frac{1}{\sqrt{2\pi}((j+1) - \sqrt{j+1})}. \end{aligned}$$

By the same reasoning, we also have $|q_1(z_2)| \le 1/(\sqrt{2\pi}((j+1) + \sqrt{j+1}))$ for z_2. Analogously, to establish the second inequality, we first denote

$$q_2(z) := e^{-z} \left(\frac{z^{j+1}}{j!} - \frac{z^{j+2}}{(j+1)!} \right).$$

Then, the derivative of $q_2(z)$ is

$$q_2'(z) = e^{-z} \frac{z^j}{(j+1)!} \left(-(2j+3)z + (j+1)^2 + z^2 \right).$$

Set $q_2'(z) = 0$ and note that $q_2(0) = \lim_{z \to \infty} q_2(z) = 0$. Hence, the maximum of $|q_2(z)|$ is attained at either $z_3 := ((2j+3) - \sqrt{4j+5})/2$ or $z_4 := ((2j+3) + \sqrt{4j+5})/2$. Furthermore, note that both $|z_3|, |z_4| \le 2(j+2)$. Therefore, we obtain

$$|q_2(z_3)| = |z_3||q_1(z_3)| \le 2(j+2) \max_z |q_1(z)| \le \frac{2(j+2)}{\sqrt{2\pi}((j+1) - \sqrt{j+1})} \le 5, \forall j \ge 1.$$

Finally, the same proof also shows that $|q_2(z_4)| \le 5$. □

B.1 Proving Theorem 2: The L_1 Distance

Now, let us focus on the problem of estimating the L_1 distance between the unknown distribution $p \in \Delta_k$ and a given distribution $q \in \Delta_k$. Since our estimator is constructed symbol by symbol, it suffices to consider the problem of approximating $\ell_q(x) := |x - q| - q$.

Let $g_{n+1}(j) := (n+1)\ell_q\left(\frac{j}{n+1}\right)$. We note that $r_{na}(j)$ equals 0 for all but at most two different values of j. Therefore, by Lemma 15, for all $z \in I'_n$, we have $|t''_{na}(z)| \lesssim 1$, and $|t''_{na}(z)| \lesssim z^{-1}$, where the first and second inequalities resemble Property 3 and 4 in Section A.2.1, respectively. Using arguments similar to those in Section A.2.3 and A.3, we can construct an estimator for $D_q(p)$ that provides the guarantees stated in Theorem 2. Note that concavity/convexity is actually not crucial for establishing the final result in Section A.2.3. Also note that we need to replace our analysis in Section 6.2 and 7.2 for the corresponding large-probability estimator by that in Hao et al. (2018).

B.2 Proving Theorem 3: General Additive Properties

More generally, our result on L_1 distance extends to 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. Without loss of generality, assume that all functions f_i's are 1-Lipschitz and satisfy $f_i(0) = 0$. By the previous derivations, we immediately have $|t''_{na}(z)| \le 6$, which recovers Property 3 in Section A.2.3. Again, concavity/convexity is actually unnecessary for establishing the final result in Section A.2.3. The proof will be complete if we also recover Property 4 in that section. In other words, we only need to show $|t''_{na}(z)z| \lesssim 1$, where

$$t''_{na}(z)z = e^{-z}\sum_{j=0}^{\infty} r_{na}(j+1)\left(\frac{z^{j+1}}{j!} - \frac{z^{j+2}}{(j+1)!}\right) - e^{-z}z \cdot r_{na}(0).$$

Fix $z \in I'_n$ and treat it as a constant. Let $b_j := r_{na}(j+1)$ and $a_j := e^{-z}\left(\frac{z^{j+1}}{j!} - \frac{z^{j+2}}{(j+1)!}\right)$. By Lemma 15, we have $|a_j| \le 5, \forall j \ge 1$. Note that there is need to worry about the slack term $e^{-z}zr_{na}(0)$ and the first term in the sum which corresponds to $j = 0$, because both terms contribute at most $\mathcal{O}(1)$ in absolute value to the above expression for any $z \ge 0$. The key observation is that any consecutive partial sum of sequence $\{b_j\}_{j\ge1}$ is also bounded by $\mathcal{O}(1)$ in magnitude. Specifically, for any $n_1, n_2 \in \mathbb{Z}^+$ satisfying the inequality $n_1 + 2 \le n_2$,

$$\begin{aligned}
\left|\sum_{j=n_1}^{n_2} b_j\right| &= \left|\sum_{j=n_1}^{n_2} r_{na}(j+1)\right| \\
&= \left|\sum_{j=n_1}^{n_2} \left(g_{na}(j+3) + g_{na}(j+1) - 2g_{na}(j+2)\right)\right| \\
&= \left|\sum_{j=n_1+3}^{n_2+3} g_{na}(j) + \sum_{j=n_1+1}^{n_2+1} g_{na}(j) - 2\sum_{j=n_1+2}^{n_2+2} g_{na}(j)\right| \\
&= \left|\left(g_{na}(n_2+3) - g_{na}(n_2+2)\right) + \left(g_{na}(n_1+1) - g_{na}(n_1+2)\right)\right| \\
&\le 2.
\end{aligned}$$

Furthermore, the sequence $\{a_j\}_{j\ge1}$ can change its monotonicity at most two times, which can be proved by considering the sign of $a_j - a_{j-1}$. More concretely,

$$\begin{aligned}
\operatorname{sign}(a_j - a_{j-1}) &= \operatorname{sign}\left(e^{-z}\left(\frac{z^{j+1}}{j!} - \frac{z^{j+2}}{(j+1)!}\right) - e^{-z}\left(\frac{z^{j}}{(j-1)!} - \frac{z^{j+1}}{j!}\right)\right) \\
&= \operatorname{sign}\left(2(j+1)z - z^2 - (j+1)j\right) \\
&= \operatorname{sign}\left(-j^2 + j(2z-1) + (2z - z^2)\right).
\end{aligned}$$

Since z is fixed, the last expression can change its value at most two times as j increases from 0 to infinity. The last piece of the proof is the following corollary of the well-known Abel's inequality.

Lemma 16. *Let $\{a'_j\}_{j=1}^m$ be a sequence of real numbers that is either increasing or decreasing, and let $\{b'_j\}_{j=1}^m$ be a sequence of real or complex numbers. Then, for $B'_t := \sum_{j=1}^t b'_t$,*

$$\left|\sum_{j=1}^{m} a'_j b'_j\right| \le \max_{t=1,\ldots,m} |B'_t|(2|a'_n| + |a'_1|).$$

By the previous discussions, we can find two indices j_1 and j_2, such that $\{a_j\}_{j=1}^{j_1}$, $\{a_j\}_{j=j_1+1}^{j_2}$, and $\{a_j\}_{j\geq j_2+1}$ are all monotone subsequences.

Then, we apply Lemma 16 to each subsequence and further bound the resulting quantity by the inequalities established above: $\left|\sum_{j=n_1}^{n_2} b_j\right| \lesssim 1$ and $|a_j| \leq 6, \forall j \geq 1$. This concludes the proof.

Finally, we point out that the above argument applies to a much broader class of additive properties beyond the Lipschitz ones, which is not addressed here for the sake of clarity and simplicity.

C Summary: Estimator Construction and Analysis

This section is essentially the same as Section 4 of the main paper (but with hyperlinks added) and serves as a summary of the previous derivations and our techniques.

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 $\varepsilon \leq 1$, define the *amplification factor* as $a := \varepsilon \log n$. Without loss of generality, assume that $\varepsilon \geq 1/\log n$ and hence $a \geq 1$. 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.

C.1 Bernstein Polynomial

Drawing i.i.d. samples Y^m from any distribution p, the expected value of the empirical estimator 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 $B_m(h,p_i)$.

C.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 C.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 them.

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

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

into two parts by thresholding the empirical counts N_i' at level $1/\varepsilon$. The first part this operation induces is $\mathcal{V}_L := \sum_{i\in[k]} B_m(h,p_i)\mathbb{1}_{N_i'>1/\varepsilon}$, corresponding to the contribution of symbols with potentially large probabilities. By Appendix C.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 Appendix C.4 and C.5 to learn each summand adaptively (to the magnitude of the probability) and compute the summation.

Concretely, recall that $\tau_n = \Theta(\log n/n)$ 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, we 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* for $\mathcal{V}_S$ is

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

where for each 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 fundamental 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, 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 (Trefethen, 2013).

Turning back to our estimator $\hat{V}_S$, by the reasoning in Appendix C.6 and C.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 the pointwise bound

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

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, we proved that the absolute bias is also at most $\mathcal{O}(S_p/n)$. 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 (Trefethen, 2013). As shown in Section 9, the algorithm takes just $\tilde{O}(n)$ time to well approximate $B_m'(h, x)$.

C.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 first-half samples' empirical 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 bound the estimation bias, we leverage the next lemma, stating that the Bernstein polynomial of h closely approximates the function over $[0, 1]$.

Lemma 17. *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 upper bound of $2S_p/n$.

We note that for Shannon entropy, 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 polynomials (operators) presents us the following pointwise bound.

Lemma 18 (Bustamante (2017) Proposition 4.9)**.** *For any $t \ge 1$, $x \in [0, 1]$, 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 inequality completes the bias analysis of the large-probability estimator, while Section 6.2 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.

C.4 Choice of Parameters and Sample Splitting

Section 4 calls for estimating $B_m(h, x)$. Applying the method of *differential smoothing* in Appendix C.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 $A^{\underline{B}}$ represents *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 the expression of $\hat{\mathcal{V}}_S$. Next, we briefly illustrate the intuitions behind our choices of parameter τ and d.

For $X \sim \mathrm{bin}(n, x)$, the variance of $\hat{H}_m(X)$ generally gets larger as the degree parameter d increases. On the other hand, a higher-degree polynomial can 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 C.2, we set $\tau = \tau_n = \Theta(\log n/n)$ since below this threshold, sample statistics are not sufficient 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.

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 N_i' at $1/\varepsilon$.

Section 5 and 6.2 present relevant technical details.

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.

C.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 Appendix C.2 and C.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 approximation*. 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 4. 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 closed 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 identity on h_m and a more straightforward argument leading to a weaker result, see Section 4 and 5.

C.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. Let $\varphi(x) := \sqrt{x(1-x)}$. 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 19. *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 $E_\cdot[f]$, 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.

C.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 20. *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 Appendix C.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 proved these bounds in Appendix A.2 and A.3. In Appendix B, a similar yet more involved argument extended 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.

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.

D A Competitive Estimator for Support Size

D.1 Estimator Construction

Denote by p and S_p an unknown distribution and its support size, respectively. For $\varepsilon \le e^{-2}$, redefine the amplification parameter as $a := |\log^{-2}\varepsilon| \cdot \log S_p$. Let X^{na} be an i.i.d. sample sequence drawn from p, and N_i'' be the number of times symbol i appears empirically.

The na-sample empirical estimator approximates the support size $S_p = \sum_{i\in[k]} \mathbb{1}_{p_i>0}$ by

$$\hat{S}^E(X^{na}) := \sum_{i\in[k]} \mathbb{1}_{N_i''>0}.$$

Taking expectation, we have

$$\mathbb{E}[\hat{S}^E(X^{na})] := \sum_{i\in[k]} \mathbb{E}[\mathbb{1}_{N_i''>0}] = \sum_{i\in[k]} (1-(1-p_i)^{na}).$$

For a length-$\mathrm{Poi}(n)$ sample sequence X^N, denote by ϕ_j the number of symbols that appear j times. Following Acharya et al. (2017a); Orlitsky et al. (2016), we can estimate $\mathbb{E}[\hat{S}^E(X^{na})]$ by

$$\hat{S}(X^N) := \sum_{j=1}^{\infty} \phi_j (1 - (-(a-1))^j \Pr(Z \ge j)),$$

where $Z \sim \mathrm{Poi}(r)$ for some *smoothing parameter* r. Similar to the previous notation, we define N_i as the number of times symbol i appears in X^N. Then, all the N_i's are mutually independent.

D.2 Bounding the Bias

The following lemma bounds the bias of $\hat{S}(X^N)$ in estimating $\mathbb{E}[\hat{S}^E(X^{na})]$.

Lemma 21. *For any $a \geq 1$,*

$$|\mathbb{E}[\hat{S}(X^N)] - \mathbb{E}[\hat{S}^E(X^{na})]| \leq \min\{na, S_p\}\, e^{-r} + 2.$$

Proof. Note that for any $m \geq 0$ and $p \in [0,1]$,

$$0 \leq e^{-mp} - (1-p)^m \leq 2p.$$

Hence, we obtain

$$\begin{aligned}
&|\mathbb{E}[\hat{S}(X^N)] - \mathbb{E}[\hat{S}^E(X^{na})]| \\
&= \left| \mathbb{E}\left[\sum_j \phi_j\right] - \mathbb{E}\left[\sum_j^{\infty} \phi_j(-(a-1))^j \Pr(Z \geq j)\right] - \sum_{i\in[k]} (1-(1-p_i)^{na}) \right| \\
&= \left| \sum_{i\in[k]} (1-e^{-np_i}) - \mathbb{E}\left[\sum_j^{\infty} \phi_j(-(a-1))^j \Pr(Z \geq j)\right] - \sum_{i\in[k]} (1-(1-p_i)^{na}) \right| \\
&\leq \left| \sum_{i\in[k]} (-e^{-np_i}) - \mathbb{E}\left[\sum_j^{\infty} \phi_j(-(a-1))^j \Pr(Z \geq j)\right] - \sum_{i\in[k]} (-e^{-nap_i}) \right| + 2\sum_{i\in[k]} p_i \\
&= \left| \sum_{i\in[k]} e^{-np_i}(e^{-n(a-1)p_i} - 1) - \mathbb{E}\left[\sum_j^{\infty} \phi_j(-(a-1))^j \Pr(Z \geq j)\right] \right| + 2 \\
&\leq \min\{na, S_p\}\, e^{-r} + 2,
\end{aligned}$$

where the last step follows by Lemma 7 and Corollary 2 in Orlitsky et al. (2016). □

D.3 Bounding the Mean Absolute Deviation

D.3.1 Bounds for $\hat{S}(X^N)$

In this section, we analyze the mean absolute deviation of $\hat{S}(X^N)$. To do this, we need the following two lemmas. The first lemma bounds the coefficients of this estimator.

Lemma 22 (Acharya et al. (2017a))**.** *For any $j \geq 1$ and $a \geq 1$,*

$$|1 - (-(a-1))^j \Pr(Z \geq j)| \leq 1 + e^{r(a-1)}.$$

The second lemma is the well-known McDiarmid's inequality.

Lemma 23. *Let $Y_1, \ldots, Y_m$ be independent random variables taking values in ranges $R_1, \ldots, R_m$, and let $F : R_1 \times \ldots \times R_m \to C$ with the property that if one freezes all but the w^{th} coordinate of $F(y_1, \ldots, y_m)$ for some $1 \leq w \leq m$, then F fluctuates by only most $c_w > 0$, thus $|F(y_1, \ldots, y_{w-1}, y_w, y_{w+1}, \ldots, y_m) - F(y_1, \ldots, y_{w-1}, y'_w, y_{w+1}, \ldots, y_m)| \leq c_w$ for all $y_j \in R_j$ and $y'_w \in R_w$ for $1 \leq j \leq m$. Then for any $\lambda > 0$, one has $\Pr(|F(Y) - \mathbb{E}[F(Y)]| \geq \lambda\sigma) \leq C\exp(-c\lambda^2)$ for some absolute constants $C, c > 0$, where $\sigma^2 := \sum_{j=1}^m c_j^2$.*

Note that $\hat{S}(X^N)$, viewed as a function of N_i's with indexes i satisfying $p_i \neq 0$, fulfills the conditions described in Lemma 23, with parameter $m = S_p$ and $c_w = 2 + 2e^{r(a-1)}$ for all $1 \leq w \leq m$. Therefore, for $\sigma^2 := 4S_p(1 + e^{r(a-1)})^2$,

$$\Pr(|\hat{S}(X^N) - \mathbb{E}[\hat{S}(X^N)]| \geq \lambda\sigma) \leq C\exp(-c\lambda^2).$$

This inequality further implies

$$\begin{aligned}
\mathbb{E}\left|\hat{S}(X^N) - \mathbb{E}[\hat{S}(X^N)]\right| &= \int_0^{\infty} \Pr(|\hat{S}(X^N) - \mathbb{E}[\hat{S}(X^N)]| \geq t)\, dt \\
&= \sigma \int_0^{\infty} \Pr(|\hat{S}(X^N) - \mathbb{E}[\hat{S}(X^N)]| \geq \lambda\sigma)\, d\lambda \\
&\leq C\sigma \int_0^{\infty} \exp(-c\lambda^2) d\lambda \\
&\lesssim \sqrt{S_p}(1 + e^{r(a-1)}).
\end{aligned}$$

Analogously, treating $\hat{S}(X^N)$ as a function of X_i's yields

$$\mathbb{E}\left|\hat{S}(X^N) - \mathbb{E}[\hat{S}(X^N)]\right| \lesssim \sqrt{n}(1+e^{r(a-1)}).$$

Consolidating the previous results, we obtain

$$\mathbb{E}\left|\hat{S}(X^N) - \mathbb{E}[\hat{S}(X^N)]\right| \lesssim \sqrt{\min\{S_p, n\}}(1+e^{r(a-1)}).$$

D.3.2 Bounds for $\hat{S}^E(X^{na})$

The following lemma bounds the variance of $\hat{S}^E(X^{na})$ in terms of S_p.

Lemma 24. *For $m \geq 1$ and $X^m \sim p$,*

$$\mathrm{Var}(\hat{S}^E(X^m)) \lesssim S_p.$$

Proof. In this proof, we slightly abuse the notation and denote by N_i the number of times symbol i appears in X^m. Incorporating the definition,

$$\mathrm{Var}(\hat{S}^E(X^m)) = \mathrm{Var}\left(\sum_{i:p_i>0} \mathbb{1}_{N_i>0}\right) = \mathbb{E}\left(\sum_{i:p_i>0} \mathbb{1}_{N_i>0}\right)^2 - \left(\mathbb{E}\left[\sum_{i:p_i>0} \mathbb{1}_{N_i>0}\right]\right)^2.$$

Let Y^M be an independent length-$\mathrm{Poi}(m)$ sample sequence from p, and N_i' be the number of times symbol i appearing in X^M. Then,

$$\begin{aligned}
\mathbb{E}\left(\sum_{i:p_i>0} \mathbb{1}_{N_i>0}\right)^2 &= \mathbb{E}\left[\sum_{i:p_i>0} \mathbb{1}_{N_i>0} + \sum_{i\neq j:p_i>0,p_j>0} \mathbb{1}_{N_i>0}\mathbb{1}_{N_j>0}\right] \\
&= \sum_{i:p_i>0} (1 - \mathbb{E}[\mathbb{1}_{N_i=0}]) + \sum_{i\neq j:p_i>0,p_j>0} \mathbb{E}[(1-\mathbb{1}_{N_i=0})(1-\mathbb{1}_{N_j=0})] \\
&= \sum_{i:p_i>0} (1-(1-p_i)^m) + \sum_{i\neq j:p_i>0,p_j>0} (1-(1-p_i)^m-(1-p_j)^m+(1-p_i-p_j)^m).
\end{aligned}$$

Note that for any $m \geq 0$ and $p \in [0,1]$,

$$0 \leq e^{-mp} - (1-p)^m \leq 2p.$$

Then, we must have both

$$|(1-(1-p_i)^m) - (1-e^{-mp_i})| \leq 2p_i$$

and

$$|(1-(1-p_i)^m-(1-p_j)^m+(1-p_i-p_j)^m) - (1-e^{-mp_i}-e^{-mp_j}+e^{-m(p_i+p_j)})| \leq 4(p_i+p_j).$$

Therefore,

$$\begin{aligned}
\left|\mathbb{E}\left(\sum_{i:p_i>0} \mathbb{1}_{N_i>0}\right)^2 - \mathbb{E}\left(\sum_{i:p_i>0} \mathbb{1}_{N_i'>0}\right)^2\right| &\leq \sum_{i:p_i>0} 2p_i + \sum_{i\neq j:p_i>0,p_j>0} 4(p_i+p_j) \\
&\leq 4\sum_{i:p_i>0}\sum_{j:p_j>0}(p_i+p_j) \\
&\leq 8S_p.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\left|\left(\mathbb{E}\left[\sum_{i:p_i>0} \mathbb{1}_{N_i>0}\right]\right)^2 - \left(\mathbb{E}\left[\sum_{i:p_i>0} \mathbb{1}_{N_i'>0}\right]\right)^2\right| \\
&= \left|\mathbb{E}\left[\sum_{i:p_i>0} \mathbb{1}_{N_i>0}\right] - \mathbb{E}\left[\sum_{i:p_i>0} \mathbb{1}_{N_i'>0}\right]\right| \left|\mathbb{E}\left[\sum_{i:p_i>0} \mathbb{1}_{N_i>0}\right] + \mathbb{E}\left[\sum_{i:p_i>0} \mathbb{1}_{N_i'>0}\right]\right| \\
&\leq \left|\sum_{i:p_i>0} \mathbb{E}\left[\mathbb{1}_{N_i>0}\right] - \sum_{i:p_i>0} \mathbb{E}\left[\mathbb{1}_{N_i'>0}\right]\right| \cdot 2S_p \\
&\leq \left(\sum_{i:p_i>0} 2p_i\right) \cdot 2S_p \\
&\leq 4S_p.
\end{aligned}$$

Finally, note that changing the value of a single observation changes the value of $\sum_{i:p_i>0} \mathbb{1}_{N_i'>0}$ by at most one. Hence, by McDiarmid's inequality,

$$\mathrm{Var}\left(\sum_{i:p_i>0} \mathbb{1}_{N_i'>0}\right) \lesssim S_p.$$

The triangle inequality combines the previous inequalities and yields

$$\mathrm{Var}\left(\sum_{i:p_i>0} \mathbb{1}_{N_i>0}\right) \lesssim S_p. \qquad \square$$

By Jensen's inequality, the above lemma implies that

$$\mathbb{E}\left|\hat{S}^E(X^{na}) - \mathbb{E}[\hat{S}^E(X^{na})]\right| \le \sqrt{\mathrm{Var}(\hat{S}^E(X^{na}))} \lesssim \sqrt{S_p}.$$

D.4 Proving Theorem 4

Setting $r = |\log \varepsilon|$, we obtain

$$e^{r(a-1)} \le S_p^{|\log^{-1}\varepsilon|}$$

and

$$e^{-r} = e^{-|\log\varepsilon|} = \varepsilon.$$

Therefore, by the previous results,

$$\begin{aligned}\mathbb{E}\left|\hat{S}(X^N) - \hat{S}^E(X^{na})\right| &\le \mathbb{E}\left|\hat{S}(X^N) - \mathbb{E}[\hat{S}^E(X^{na})]\right| + \mathbb{E}\left|\mathbb{E}[\hat{S}^E(X^{na})] - \hat{S}^E(X^{na})\right| \\ &\lesssim S_p^{|\log^{-1}\varepsilon|+\frac{1}{2}} + S_p \cdot \varepsilon.\end{aligned}$$

Normalize both sides by S_p. Then,

$$\mathbb{E}\left|\frac{\hat{S}(X^N)}{S_p} - \frac{\hat{S}^E(X^{na})}{S_p}\right| \lesssim S_p^{|\log^{-1}\varepsilon|-\frac{1}{2}} + \varepsilon.$$

E A Competitive Estimator for Support Coverage

E.1 Estimator Construction

Recall that $c(p) = 1-(1-p_i)^m$, where m is a *given parameter*. For $\varepsilon \le e^{-2}$, redefine the amplification parameter as $a := |\log^{-2}\varepsilon| \cdot \log C_p$. Similar to the last section, let X^{na} be an independent length-na sample sequence drawn from p, and N_i'' be the number of times symbol i appears empirically.

The na-sample empirical estimator estimates the m-sample support coverage $C_p = \sum_{i\in[k]} c(p_i)$ by

$$\hat{C}^E(X^{na}) := \sum_{i\in[k]} c\left(\frac{N_i''}{na}\right) = \sum_{i\in[k]} \left(1-\left(1-\frac{N_i''}{na}\right)^m\right).$$

Taking expectation, we obtain

$$\mathbb{E}[\hat{C}^E(X^{na})] = \sum_{i\in[k]} \mathbb{E}\left[1-\left(1-\frac{N_i''}{na}\right)^m\right].$$

For the ease of exposition, let us denote

$$T(p) := \sum_{i\in[k]} \mathbb{E}\left[1-e^{-m\frac{N_i''}{na}}\right].$$

Noting that for any $t \ge 1$ and $p \in [0,1]$,

$$|e^{-tp} - (1-p)^t| \le 2p,$$

hence, we have

$$|\mathbb{E}[\hat{C}^E(X^{na})] - T(p)| \le \sum_{i\in[k]} \mathbb{E}\left[2\cdot\frac{N_i''}{na}\right] = 2.$$

Then, it suffices to estimate $T(p)$, which satisfies

$$\begin{aligned}
T(p) &= \sum_{i\in[k]} \left(1-\mathbb{E}\left[e^{-m\frac{N_i''}{na}}\right]\right) \\
&= \sum_{i\in[k]} \left(1-\sum_{j=0}^{na}\binom{na}{j} p_i^j (1-p_i)^{na-j} e^{-m\frac{j}{na}}\right) \\
&= \sum_{i\in[k]} \left(1-\sum_{j=0}^{na}\binom{na}{j} \left(p_i\cdot e^{-\frac{m}{na}}\right)^j (1-p_i)^{na-j}\right) \\
&= \sum_{i\in[k]} \left(1-\left(1-p_i(1-e^{-\frac{m}{na}})\right)^{na}\right).
\end{aligned}$$

Analogous to the definition of $T(p)$, let us denote

$$T_1(p) := \sum_{i\in[k]} \left(1-\exp\left(-na(1-e^{-\frac{m}{na}})p_i\right)\right).$$

Since $(1-e^{-\frac{m}{na}})\cdot p_i \in [0,1]$, we must have

$$|T(p) - T_1(p)| \le \sum_{i\in[k]} 2(1-e^{-\frac{m}{na}})p_i \le 2.$$

Define a new amplification parameter $a' := a(1-e^{-\frac{m}{na}})$. Then, we can express $T_1(p)$ as

$$T_1(p) := \sum_{i\in[k]} \left(1-\exp\left(-na'p_i\right)\right).$$

For simplicity, we will assume that $m \ge 1.5n$ and $a > 1.8$, ensuring

$$a' = a(1-e^{-\frac{m}{na}}) \ge a(1-e^{-\frac{1.5}{a}}) > 1.$$

Analogous to case of support size estimation, we draw a length-$\mathrm{Poi}(n)$ sample sequence X^N and estimate $\mathbb{E}[\hat{C}^E(X^{na})]$ by the estimator

$$\hat{C}(X^N) := \sum_{j=1}^{\infty} \phi_j (1-(-(a'-1))^j \Pr(\mathrm{Poi}(r)\ge j)),$$

where ϕ_j denotes the number of symbols appearing j times.

E.2 Bounding the Bias

We bound the bias of $\hat{C}(X^N)$ in estimating $\mathbb{E}[\hat{C}^E(X^{na})]$ as follows.

$$\begin{aligned}
|\mathbb{E}[\hat{C}(X^N)] - \mathbb{E}[\hat{C}^E(X^{na})]| &\le |\mathbb{E}[\hat{C}(X^N)] - T_1(p)| + |T_1(p) - \mathbb{E}[\hat{C}^E(X^{na})]| \\
&\le |\mathbb{E}[\hat{C}(X^N)] - T_1(p)| + 4 \\
&= \left| \sum_{i\in[k]} e^{-np_i}(e^{-n(a'-1)p_i} - 1) \right. \\
&\quad \left. - \sum_{i\in[k]} e^{-np_i} \sum_{j=1}^{\infty} \frac{(-(a'-1)np_i)^j}{j!} \Pr(\mathrm{Poi}(r) \ge j) \right| + 4 \\
&\le \left| \sum_{i\in[k]} e^{-np_i} \left(\sum_{j=1}^{\infty} \frac{(-(a'-1)np_i)^j}{j!} \Pr(\mathrm{Poi}(r) < j) \right) \right| + 4.
\end{aligned}$$

To bound the last sum, we need the following lemma.

Lemma 25. *For any $y, r \ge 0$,*

$$\left| \sum_{j=1}^{\infty} \frac{(-y)^j}{j!} \Pr(\mathrm{Poi}(r) < j) \right| \le e^{-r}(1 - e^{-y}).$$

Proof. By Lemma 6 of Orlitsky et al. (2016),

$$\begin{aligned}
\left| \sum_{j=1}^{\infty} \frac{(-y)^j}{j!} \Pr(\mathrm{Poi}(r) < j) \right| &\le \max_{s\le y} \left| \mathbb{E}_{L\sim\mathrm{Poi}(r)} \left[\frac{(-s)^L}{L!} \right] \right| (1 - e^{-y}) \\
&= \max_{s\le y} \left| J_0(2\sqrt{sr}) \right| e^{-r}(1 - e^{-y}) \\
&\le e^{-r}(1 - e^{-y}),
\end{aligned}$$

where J_0 is the first-order Bessel function of the first kind, and satisfies the elegant inequality $|J_0(x)| \le 1, \forall x \ge 0$ (Abramowitz & Stegun, 1965). □

Leveraging the above lemma, we obtain

$$\begin{aligned}
|\mathbb{E}[\hat{C}(X^N)] - \mathbb{E}[\hat{C}^E(X^{na})]| &\le \left| \sum_{i\in[k]} e^{-np_i} \left(\sum_{j=1}^{\infty} \frac{(-(a'-1)np_i)^j}{j!} \Pr(\mathrm{Poi}(r) < j) \right) \right| + 4 \\
&\le e^{-r} \sum_{i\in[k]} e^{-np_i}(1 - e^{-(a'-1)np_i}) + 4 \\
&\le e^{-r} \sum_{i\in[k]} (1 - e^{-na'p_i}) + 4.
\end{aligned}$$

Note that $na' = na(1 - e^{-\frac{m}{na}}) \le m$. Therefore,

$$|\mathbb{E}[\hat{C}(X^N)] - \mathbb{E}[\hat{C}^E(X^{na})]| \le e^{-r} \sum_{i\in[k]} (1 - e^{-mp_i}) + 4 = e^{-r} C_p + 4.$$

E.3 Bounding the Mean Absolute Deviation

E.3.1 Bounds for $\hat{C}(X^N)$

First, we bound the mean absolute deviation of $\hat{C}(X^N)$ in terms of C_p. By Jensen's inequality,

$$\begin{aligned}
\mathbb{E}\left|\hat{C}(X^N) - \mathbb{E}[\hat{C}(X^N)]\right| &\le \sqrt{\mathrm{Var}\left(\hat{C}(X^N)\right)} \\
&= \sqrt{\sum_{i\in k} \mathrm{Var}\left(\sum_{j=1}^{\infty} \mathbb{1}_{N_i=j}(1-(-(a'-1))^j \Pr(\mathrm{Poi}(r)\ge j))\right)} \\
&\le \sqrt{\sum_{i\in k} \mathbb{E}\left[\left(\sum_{j=1}^{\infty} \mathbb{1}_{N_i=j}(1-(-(a'-1))^j \Pr(\mathrm{Poi}(r)\ge j))\right)^2\right]} \\
&= \sqrt{\sum_{i\in k}\sum_{j=1}^{\infty} \mathbb{E}\left[\mathbb{1}_{N_i=j}\right](1-(-(a'-1))^j \Pr(\mathrm{Poi}(r)\ge j))^2} \\
&\le (1+e^{r(a'-1)})\sqrt{\sum_{i\in k}(1-e^{-np_i})}.
\end{aligned}$$

By our assumption that $m \ge 1.5n$,

$$\begin{aligned}
\mathbb{E}[|\hat{C}(X^N) - \mathbb{E}[\hat{C}(X^N)]|] &\le (1+e^{r(a'-1)})\sqrt{\sum_{i\in k}(1-e^{-np_i})} \\
&\le (1+e^{r(a'-1)})\sqrt{\sum_{i\in k}(1-e^{-mp_i})} \\
&\le (1+e^{r(a'-1)})\sqrt{\sum_{i\in k}(1-(1-p_i)^m)} \\
&= (1+e^{r(a'-1)})\sqrt{C_p}.
\end{aligned}$$

E.3.2 Bounds for $\hat{C}^E(X^{na})$

Next, we bound the mean absolute deviation of the na-sample empirical estimator. To deal with the dependence among the counts N_i'''s, we need the following lemma (Joag-Dev & Proschan, 1983).

Definition 1. *Random variables $X_1, \ldots, X_S$ are said to be* negatively associated *if for any pair of disjoint subsets A_1, A_2 of $1, 2, \ldots, S$, and any component-wise increasing functions f_1, f_2,*

$$Cov(f_1(X_i, i \in A_1), f_2(X_j, j \in A_2)) \le 0.$$

The following result can be used to check whether random variables are negatively associated or not.

Lemma 26. *Let $X_1, \ldots, X_S$ be S independent random variables with log-concave densities. Then the joint conditional distribution of $X_1, \ldots, X_S$ given $\sum_{i=1}^{S} X_i$ is negatively associated.*

Lemma 26 shows that N_i'''s are negatively correlated. Furthermore, note that

$$c^*(x) := 1 - \left(1 - \frac{x}{na}\right)^m$$

is an increasing function, and we can write the quantity of interest as

$$\hat{C}^E(X^{na}) := \sum_{i\in[k]} c^*(N_i'').$$

Hence, for any $i, j \in [k]$ such that $i \ne j$,

$$\mathrm{Cov}(c^*(N_i''), c^*(N_j'')) \le 0.$$

Consequently,

$$\begin{aligned}
\mathrm{Var}\left(\hat{C}^E(X^{na})\right) &= \sum_{i\in[k]} \mathrm{Var}(c^*(N_i'')) + 2 \sum_{i,j\in[k], i\neq j} \mathrm{Cov}(c^*(N_i''), c^*(N_j'')) \\
&\leq \sum_{i\in[k]} \mathrm{Var}(c^*(N_i'')) \\
&\leq \sum_{i\in[k]} \mathbb{E}(c^*(N_i''))^2 \\
&= \sum_{i\in[k]} \mathbb{E}\left[\sum_{j=0}^{na} \mathbb{1}_{N_i=j} (C^*(j))^2\right] \\
&\leq \sum_{i\in[k]} \sum_{j=1}^{na} \mathbb{E}[\mathbb{1}_{N_i=j}] \\
&= \sum_{i\in[k]} (1-(1-p_i)^{na}).
\end{aligned}$$

Without loss of generality, we will assume that a is a positive integer. Then,

$$\begin{aligned}
\sum_{i\in[k]} (1-(1-p_i)^{na}) &= \sum_{i\in[k]} (1-(1-p_i)^n)(\sum_{j=0}^{a-1}(1-p_i)^{nj}) \\
&\leq a \sum_{i\in[k]} (1-(1-p_i)^n) \\
&\leq a \sum_{i\in[k]} (1-(1-p_i)^m) \\
&= aC_p.
\end{aligned}$$

Finally, Jensen's inequality implies

$$\mathbb{E}\left|\hat{C}^E(X^{na}) - \mathbb{E}[\hat{C}^E(X^{na})]\right| \leq \sqrt{\mathrm{Var}(\hat{C}^E(X^{na}))} \leq \sqrt{aC_p}.$$

E.4 Proving Theorem 5

The triangle inequality consolidates the major inequalities in the previous sections and yields

$$\mathbb{E}\left|\hat{C}(X^N) - \hat{C}^E(X^{na})\right| \lesssim e^{-r}C_p + 4 + \sqrt{aC_p} + (1+e^{r(a'-1)})\sqrt{C_p}.$$

By the fact that $a' < a = |\log^{-2}\varepsilon| \cdot \log C_p$, we set $r = |\log\varepsilon|$ and obtain

$$\mathbb{E}\left|\hat{C}(X^N) - \hat{C}^E(X^{na})\right| \lesssim \varepsilon C_p + 4 + (1 + C_p^{|\log^{-1}\varepsilon|} + \sqrt{\log C_p})\sqrt{C_p}.$$

Then, normalizing both sides by C_p gives

$$\mathbb{E}\left|\frac{\hat{C}(X^N)}{C_p} - \frac{\hat{C}^E(X^{na})}{C_p}\right| \lesssim C_p^{|\log^{-1}\varepsilon|-\frac{1}{2}} + \varepsilon.$$

References

Abbott, J. Quadratic interval refinement for real roots. *ACM Communications in Computer Algebra*, 48(1/2):3–12, 2014.

Abramowitz, M. and Stegun, I. A. *Handbook of mathematical functions with formulas, graphs, and mathematical table*. National Bureau of Standards Applied Mathematics Series 55, 1965.

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.

Berens, H., Lorentz, G. G., and MacKenzie, R. E. Inverse theorems for Bernstein polynomials. *Indiana University Mathematics Journal*, 21(8):693–708, 1972.

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.

Chung, F. R. and Lu, L. *Complex graphs and networks (No. 107)*. American Mathematical Soc., 2006.

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.

Dixon, J. D. Exact solution of linear equations using P-adic expansions. *Numerische Mathematik*, 40 (1):137–141, 1982.

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.

Ehlich, H. and Zeller, K. Auswertung der normen von interpolationsoperatoren. *Mathematische Annalen*, 164(2):105–112, 1966.

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.

Joag-Dev, K. and Proschan, F. Negative association of random variables with applications. *The Annals of Statistics*, pp. 286–295, 1983.

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.

Kerber, M. On the complexity of reliable root approximation. In *Proceedings of the International Workshop on Computer Algebra in Scientific Computing*, pp. 155–167. Springer, 2009.

Korněichuk, N. P. *Exact constants in approximation theory*, volume 38. Cambridge University Press, 1991.

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.

Remez, E. Y. Sur la détermination des polynômes d'approximation de degré donnée. *Comm. Soc. Math. Kharkov*, 10(196):41–63, 1934.

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.

Veidinger, L. On the numerical determination of the best approximations in the Chebyshev sense. *Numerische Mathematik*, 2(1):99–105, 1960.

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.