A Unified Maximum Likelihood Approach for Estimating Symmetric Properties of Discrete Distributions

Jayadev Acharya [1] **Hirakendu Das** [2] **Alon Orlitsky** [3] **Ananda Theertha Suresh** [4]

Abstract

Symmetric distribution properties such as support size, support coverage, entropy, and proximity to uniformity, arise in many applications. Recently, researchers applied different estimators and analysis tools to derive asymptotically sample-optimal approximations for each of these properties. We show that a single, simple, plug-in estimator—*profile maximum likelihood (PML)*–is sample competitive for all symmetric properties, and in particular is asymptotically sample-optimal for all the above properties.

1. Introduction

1.1. Symmetric distribution properties

Let $\Delta \stackrel{\text{def}}{=} \{(p_1,\ldots,p_k)\colon\ p_i \geq 0,\ \sum_{i=1}^{k} p_i = 1,\ 1 \leq k \leq \infty\}$ denote the collection of all discrete distributions over finite or infinite support. A distribution *property* is a mapping $f : \Delta \to \mathbb{R}$. It is *symmetric* if it remains unchanged under relabeling of domain symbols, namely if it is determined by just the probability multiset $\{p_1, p_2, \ldots, p_k\}$. Many important properties are symmetric. For example:

Support size $S(p) = |\{x : p(x) > 0\}|$, plays an important role in population and vocabulary estimation.

Support coverage $S_m(p) = \sum_x (1-(1-p(x))^m)$, the expected number of elements observed in m samples, arises in ecological and biological studies, *e.g.*, (Colwell et al., 2012).

Shannon entropy $H(p) = \sum_x p(x) \log \frac{1}{p(x)}$, central to information theory (Cover & Thomas, 2006), has numerous applications.

Distance to uniform $\|p-u\|_1 = \sum_x |p(x)-1/|\mathcal{X}||$, where u is the uniform distribution over the domain $\mathcal{X}$ of p. This distance measure appears in the error of hypothesis testing, and the uniform distribution is arguably one of the commonest discrete distributions.

1.2. Distribution estimation

Considerable research, over many years, has focused on estimating distribution properties. In the common setting, an unknown underlying distribution $p \in \Delta$ generates n independent samples $X^n \stackrel{\text{def}}{=} X_1, , \ldots, X_n$, and the objective is to estimate a given property $f(p)$ as accurately as possible.

Specifically, an *estimator* for a distribution p over $\mathcal{X}$ is a function $\hat{f} : \mathcal{X}^n \to \mathbb{R}$ mapping observed samples to a property estimate. The *sample complexity* of $\hat{f}$ is the smallest number of samples it requires to estimate a property f with accuracy ε and confidence probability δ, for all distributions in a collection $\mathcal{P} \subseteq \Delta$,

$$C^{\hat{f}}(f, \mathcal{P}, \delta, \varepsilon) \stackrel{\text{def}}{=} \min\left\{n\colon p(|f(p) - \hat{f}(X^n)| \geq \varepsilon) \leq \delta\ \forall p \in \mathcal{P}\right\}.$$

The sample complexity of estimating f is the lowest sample complexity of any estimator,

$$C^*(f, \mathcal{P}, \delta, \varepsilon) = \min_{\hat{f}} C^{\hat{f}}(f, \mathcal{P}, \delta, \varepsilon).$$

By taking the median of about $\log \frac{1}{\delta}$ independent estimators, the error rate can be driven down from a constant to δ. Therefore, the sample complexity depends on δ only through a factor of *at most* $\log \frac{1}{\delta}$. For simplicity, we therefore abbreviate $C^{\hat{f}}(f, \mathcal{P}, 1/3, \varepsilon)$ by $C^{\hat{f}}(f, \mathcal{P}, \varepsilon)$.

1.3. Result summary

Recent research has shown that while simple estimators for the aforementioned properties require sample size n proportional to the support size k, more sophisticated techniques need only a sub-linear sample size $n = \Theta(k/\log k)$.

[*]Equal contribution [1]Cornell University, Ithaca, NY [2]Yahoo Inc!, Sunnyvale, CA [3]University of California, San Diego [4]Google Research. Correspondence to: Jayadev Acharya <acharya@cornell.edu>, Hirakendu Das <hdas@yahoo-inc.com>, Alon Orlitsky <alon@ucsd.edu>, Ananda Theertha Suresh <theertha@google.com>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

However, each of the problems was approximated via different estimators and analysis techniques, that for some properties were rather complex.

Motivated by the principle of maximum likelihood, we show that a single, simple, plug-in estimator—profile maximum likelihood (PML) (Orlitsky et al., 2004b)— is competitive for estimating any symmetric property. Its sample complexity is at most quadratically worse than that of any estimator.

Specifically, we show that if a symmetric property can be estimated using n samples with confidence δ, then the PML plug-in estimator can estimate it using as many samples with confidence $\delta \cdot e^{\sqrt{n}}$. While this increase may seem high, note that it is sub-exponential. We show that if a property has an estimator that has a small bounded difference constant (how much the estimator changes when we change one sample), then the error probability reduces exponentially with n (Please see Section 7.1). Combined, these two facts imply that for properties with locally-smooth estimators, the PML plug-in estimator is optimal up to a constant: $C^{\text{PML}} = \Theta(C^*)$. We then show that all the above properties have locally-smooth estimators, hence they can be estimated by the PML plug-in estimator with up to a constant factor more than the optimal number of samples.

1.4. Outline

The rest of the paper is organized as follows. In Section 2 we describe existing results and those shown in this paper. In Section 3 we formally define the quantities involved and state the results. In Section 4 we define profiles and PML. In Section 5, we outline the new approach. In Section 6, we demonstrate auxiliary results for maximum likelihood estimators. In Section 7, we outline how we apply maximum likelihood to support, support coverage, entropy, and uniformity. In Section 8, we provide the details for support, and support coverage and in the appendix we outline results for distance to uniformity and entropy.

2. Previous and New Results

2.1. Previous Results

Plug-in estimation is a general approach for estimating distribution properties. It uses the samples X^n to find an approximation $\hat{p}$ of p, and lets $f(\hat{p})$ estimate $f(p)$.

One of the most common distribution estimators, dating back to Fisher is *maximum likelihood*, that for clarity we call *sequence maximum likelihood (SML)* (Aldrich, 1997). To any sample x^n it assigns the distribution p that maximizes $p(x^n)$. The SML estimate is exceedingly simple to derive. The *multiplicity* $N_x \stackrel{\text{def}}{=} N_x(x^n)$ of symbol x is the number of times it appears in the sequence x^n. The *empirical frequency* estimator assigns to each symbol x, the fraction $\hat{p}(x) \stackrel{\text{def}}{=} N_x/n$ of times it appears in the sample x^n. For example, if $x^7 =$*bananas*, empirical frequency would assign $\hat{p}(a) = 3/7$, $\hat{p}(n) = 2/7$, and $\hat{p}(b) = \hat{p}(s) = 1/7$. It can be readily shown that SML is exactly the empirical frequency estimator.

While the SML plug-in estimator performs well in the limit of many samples, sophisticated techniques have recently yielded more accurate estimators for several important symmetric properties.

Support size. With finitely many samples, $S(p)$ cannot be estimated to any accuracy as many symbols with arbitrarily small probability may not be observed. Motivated by databases, where each entry appears at least once, (Raskhodnikova et al., 2009) considered distributions whose non-zero probabilities are at least $\frac{1}{k}$,

$$\Delta_{\ge \frac{1}{k}} \stackrel{\text{def}}{=} \{p \in \Delta : p(x) \in \{0\} \cup [1/k, 1]\},$$

and estimated the normalized support $\tilde{S}(p) \stackrel{\text{def}}{=} S(p)/k$. It can be shown that $C^{\text{SML}}(\tilde{S}(p), \Delta_{\ge \frac{1}{k}}, \varepsilon) = \Theta(k \log \frac{1}{\varepsilon})$. Yet (Valiant & Valiant, 2011a; Wu & Yang, 2015) showed that $C^*(\tilde{S}(p), \Delta_{\ge \frac{1}{k}}, \varepsilon) = \Theta\left(\frac{k}{\log k} \cdot \log^2 \frac{1}{\varepsilon}\right)$.

Support coverage. Here too we consider the normalized coverage $\tilde{S}_m(p) \stackrel{\text{def}}{=} S_m(p)/m$. (Good & Toulmin, 1956) proposed the Good Toulmin (GT) estimator that achieves $C^{\text{GT}}(\tilde{S}_m(p), \Delta, \varepsilon) = m/2$. Recently, (Orlitsky et al., 2016) derived a simple estimator showing that $C^*(\tilde{S}_m(p), \Delta, \varepsilon) = \Theta(\frac{m}{\log m} \cdot \log \frac{1}{\varepsilon})$. (Zou et al., 2016) derived a more complex estimator with similar dependence on m but worse dependence on ε.

Shannon entropy. Since elements with arbitrarily small probability can contribute to an arbitrarily high entropy, $H(p)$ cannot be estimated over aribtrary support with finitely many samples. Therefore researchers are mostly interested in estimating entropy of distributions with support size at most k.

$$\Delta_k \stackrel{\text{def}}{=} \{p \in \Delta : S(p) \le k\}.$$

It can be shown that $C^{\text{SML}}(H(p), \Delta_k, \varepsilon) = \Theta(\frac{k}{\varepsilon})$ (Paninski, 2003). Moreover, (Paninski, 2003) showed that $C^*(H(p), \Delta_k, \varepsilon)$ is sublinear in k, (Valiant & Valiant, 2011a) showed that the optimal dependence on k is $k/\log k$ and (Wu & Yang, 2016; Jiao et al., 2015) obtained the optimal dependence on both k, and ε, and showed that $C^*(H(p), \Delta_k, \varepsilon) = \Theta\left(\frac{k}{\log k} \cdot \frac{1}{\varepsilon}\right)$.

Distance to uniform. (Valiant & Valiant, 2011b) showed that $C^*(\|p - u\|_1, \Delta_k, \varepsilon) = \mathcal{O}\left(\frac{k}{\log k} \cdot \frac{1}{\varepsilon^2}\right)$, and (Jiao et al., 2016) showed that this bound is tight.

These results are summarized in Table 1.

Other properties were considered as well. (Bar-Yossef et al., 2001; Acharya et al., 2015; Caferov et al., 2015; Obremski & Skorski, 2017) estimated Rényi entropy and (Bu et al., 2016) estimated KL divergence. (Canonne, 2015) surveyed testing whether distributions have certain properties, and (Jiao et al., 2014) studied the performance of SML estimators for several properties. Closest to this work in terms of approach and techniques are (Acharya et al., 2011; 2012; 2013a;b; Valiant & Valiant, 2013; Orlitsky & Suresh, 2015) that design algorithms whose sample complexity is provably close to the best possible regardless of the domain size.

2.2. Profile Maximum Likelihood

Symmetric distribution properties do not depend on the symbol labels. They are determined by a simple sufficient statistic: the number of elements appearing any given number of times. The *profile* of a sequence X^n, denoted $\varphi(X^n)$ is the multiset of the multiplicities of all the symbols appearing in X^n. For example, $\varphi(a\ b\ r\ a\ c\ a\ d\ a\ b\ r\ a) = \{1, 1, 2, 2, 5\}$, as two symbols appearing once, two appearing twice, and one symbol appearing five times, removing the association of the individual symbols with the multiplicities. Profiles are also referred to as histograms of histograms (Batu et al., 2000), histogram order statistics (Paninski, 2003), and fingerprints (Valiant & Valiant, 2011a).

Motivated by the *principle of maximum likelihood*, (Orlitsky et al., 2004b; 2017b) discarded the symbol labels, and considered the *profile maximum likelihood (PML)* distribution that maximizes the probability of the observed profile.

A number of PML properties were established. (Orlitsky et al., 2004b; 2005) proved PML's existence, consistency, and some of its properties. (Orlitsky et al., 2004d; 2005; Orlitsky & Pan, 2009; Pan et al., 2009) described additional properties and derived the PML distributions of several short and simple profiles. (Orlitsky et al., 2017b;c) provide a unified review of several of these results. (Anevski et al., 2013) contains a combination of previously-known and new results. A related distribution-estimation approach is described in (Orlitsky et al., 2004c; 2003).

Several approaches were taken to computing the PML distribution. Algebraic computation was considered in (Acharya et al., 2010). A combination of the EM and MCMC algorithms have shown excellent results for calculating the PML distribution and applying it to support-size estimation (Orlitsky et al., 2004a; 2006; Pan, 2012) and a summary of some of the results appears in (Orlitsky et al., 2017a). (Vontobel, 2012; 2014) derived the Bethe approximation of these algorithms.

Following the first draft of this work, (Vatedka & Vontobel, 2016) showed that both theoretically and empirically plug-in estimators obtained from the PML estimate yield good estimates for symmetric functionals of Markov distributions.

2.3. New Results

We show that replacing the SML plug-in estimator by PML yields a unified estimator that, like the best results shown via specialized techniques developed, is optimal.

Theorem 1. *There is a unified approach based on PML distribution that achieves the optimal sample complexity for the problems of estimating the entropy, support, support coverage, and distance to uniformity.*

We prove in Corollary 1 that the PML approach is *competitive* with respect to *any symmetric property*. For symmetric properties, these results are perhaps a justification of Fisher's thoughts on Maximum Likelihood:

"*Of course nobody has been able to prove that maximum likelihood estimates are best under all circumstances. Maximum likelihood estimates computed with all the information available may turn out to be inconsistent. Throwing away a substantial part of the information may render them consistent.*"

R. A. Fisher's thoughts on Maximum Likelihood (Le Cam, 1979).

To prove these PML guarantees, we establish two results that are of interest on their own right.

- With n samples, PML estimates any symmetric property of p with essentially the same accuracy, and at most $e^{3\sqrt{n}}$ times the error, of any other estimator. This follows by combining Theorem 3 with Lemma 1.

- For a large class of symmetric properties, including all those mentioned above, if there is an estimator that uses n samples, and has an error probability $1/3$, we design an estimator using $O(n)$ samples, whose error probability is nearly exponential in n. We remark that this decay is much faster than applying the median trick. This result follows by combining McDiarmid's inequality with Lemma 2.

Combined, these results prove that PML plug-in estimators are sample-optimal.

We also introduce the notion of β-*approximate ML* distributions, described in Definition 1. These distributions are more relaxed version of PML, hence may be more easily computed, yet they provide essentially the same performance guarantees.

Property name	$f(p)$	$\mathcal{P}$	C^{SML}	C^*	PML
Entropy	$H(p)$	Δ_k	$\frac{k}{\varepsilon}$	$\frac{k}{\log k}\frac{1}{\varepsilon}$	Theorem 5 and Section 8.1
Support size	$\tilde{S}(p)$	$\Delta_{\geq \frac{1}{k}}$	$k \log \frac{1}{\varepsilon}$	$\frac{k}{\log k} \log^2 \frac{1}{\varepsilon}$	Theorem 5 and Section 8.2
Support coverage	$\tilde{S}_m(p)$	Δ	m	$\frac{m}{\log m} \log \frac{1}{\varepsilon}$	Theorem 5 and Section A
Distance to u	$\|p-u\|_1$	$\Delta_{\mathcal{X}}$	$\frac{k}{\varepsilon^2}$	$\frac{k}{\log k}\frac{1}{\varepsilon^2}$	Theorem 5 and Section A

Table 1. Estimation complexity for various properties up to a constant factor. For all properties shown, PML achieves the best known results up to a constant factor. The details of where the optimal sample complexity was derived for each problem is discussed in Section 2.1.

3. Formal Definitions and Results

In the past, different sophisticated estimators were used for every property in Table 1. We show that the simple plug-in estimator that uses any PML approximation $\tilde{p}$, has optimal performance guarantees for all these properties.

In the next theorem, assume n is at least the optimal sample complexity of estimating entropy, support, support coverage, and distance to uniformity (given in Table 1) respectively.

Theorem 2. *For all $\varepsilon > c/n^{0.2}$, any plug-in $\exp(-\sqrt{n})$-approximate PML $\tilde{p}$ satisfies,*

Entropy

$$C^{\tilde{p}}(H(p), \Delta_k, \varepsilon) \asymp C^*(H(p), \Delta_k, \varepsilon),$$

Support size

$$C^{\tilde{p}}(S(p)/k, \Delta_{\geq \frac{1}{k}}, \varepsilon) \asymp C^*(S(p)/k, \Delta_{\geq \frac{1}{k}}, \varepsilon),$$

Support coverage

$$C^{\tilde{p}}(S_m(p)/m, \Delta, \varepsilon) \asymp C^*(S_m(p)/m, \Delta, \varepsilon),$$

Distance to uniformity

$$C^{\tilde{p}}(\|p-u\|_1, \Delta_{\mathcal{X}}, \varepsilon) \asymp C^*(\|p-u\|_1, \Delta_k, \varepsilon).$$

4. PML: Profile Maximum Likelihood

4.1. Preliminaries

For a sequence X^n, recall that the *multilplicity* N_x is the number of times x appears in X^n. Discarding, the labels, profile of a sequence (Orlitsky et al., 2004b) is defined below. Let Φ^n be all profiles of length-n sequences. Then, $\Phi^4 = \{\{1,1,1,1\},\{1,1,2\},\{1,3\},\{2,2\},\{4\}\}$. In particular, a profile of a length-n sequence is an unordered partition of n. Therefore, $|\Phi^n|$, the number of profiles of length-n sequences is equal to the partition number of n. Then, by the Hardy-Ramanujan bounds on the partition number,

Lemma 1 ((Hardy & Ramanujan, 1918)). $|\Phi^n| \leq \exp(3\sqrt{n})$.

For a distribution p, the probability of a profile φ is defined as

$$p(\varphi) \stackrel{\text{def}}{=} \sum_{X^n:\varphi(X^n)=\varphi} p(X^n),$$

the probability of observing a sequence with profile φ. Under *i.i.d.* sampling, $p(\varphi) = \sum_{X^n:\varphi(X^n)=\varphi} \prod_{i=1}^n p(X_i)$. For example, the probability of observing a sequence with profile $\varphi = \{1,2\}$ is the probability of observing a sequence with one symbol appearing once, and one symbol appearing twice. A sequence with a symbol x appearing twice and y appearing once (*e.g.*, $x\ y\ x$) has probability $p(x)^2p(y)$. Appropriately normalized, for any p, the probability of the profile $\{1,2\}$ is

$$p(\{1,2\}) = \sum_{\varphi(X^n)=\{1,2\}} \prod_{i=1}^n p(X_i) = \binom{3}{1} \sum_{a\neq b\in\mathcal{X}} p(a)^2 p(b), \tag{1}$$

where the normalization factor is independent of p. The summation is a monomial symmetric polynomial in the probability values. See (Pan, 2012) for more examples.

4.2. PML Estimation Scheme

Recall that p_{X^n} is the distribution maximizing the probability of X^n. Similarly, define (Orlitsky et al., 2004b):

$$p_\varphi \stackrel{\text{def}}{=} \max_{p\in\mathcal{P}} p(\varphi)$$

as the distribution in $\mathcal{P}$ that maximizes the probability of observing a sequence with profile φ.

For example, for $\varphi = \{1,2\}$. For $\mathcal{P} = \Delta_k$, from (1),

$$p_\varphi = \arg\max_{p\in\Delta_k} \sum_{a\neq b} p(a)^2 p(b).$$

Note that in contrast, SML only maximizes one term of this expression.

We give two examples from the table in (Orlitsky et al., 2004b) to distinguish between SML and PML distributions,

For $a, b > 0$, denote $a \lesssim b$ or $b \gtrsim a$ if for some universal constant c, $a/b \leq c$. Denote $a \asymp b$ if both $a \lesssim b$ and $a \gtrsim b$.

and also show an instance where PML outputs distributions with a larger support than those appearing in the sample.

Example 1. *Let $\mathcal{X} = \{a, b, \ldots, z\}$. Suppose $X^n = x\ y\ x$, then the SML distribution is $(2/3, 1/3)$. However, the distribution in Δ that maximizes the probability of the profile $\varphi(x\ y\ x) = \{1, 2\}$ is $(1/2, 1/2)$. Another example, illustrating the power of PML to predict new symbols is $X^n = a\ b\ a\ c$, with profile $\varphi(a\ b\ a\ c) = \{1, 1, 2\}$. The SML distribution is $(1/2, 1/4, 1/4)$, but the PML is a uniform distribution over 5 elements, namely $(1/5, 1/5, 1/5, 1/5, 1/5)$.*

Suppose we want to estimate a symmetric property $f(p)$ of an unknown distribution $p \in \mathcal{P}$ given n independent samples. Our high level approach using PML is described below.

Input: Class of distributions $\mathcal{P}$, symmetric function $f(\cdot)$, sample X^n

1. Compute p_{φ} : $\arg\max_{p \in \mathcal{P}} p(\varphi(X^n))$.
2. Output $f(p_{\varphi})$.

There are a few advantages of this approach (as is true with any plug-in approach): (i) the computation of PML is agnostic to the function f at hand, (ii) there are no parameters to be tuned, (iii) techniques such as Poisson sampling or median tricks are not necessary, (iv) well motivated by the maximum-likelihood principle.

Comparison to the linear-programming plug-in estimator (Valiant & Valiant, 2011a). Our approach is perhaps closest in flavor to the plug-in estimator of (Valiant & Valiant, 2011a). Indeed, as mentioned in (Valiant, 2012), their linear-programming estimator is motivated by the question of estimating the PML. Their result was the first estimator to provide sample complexity bounds in terms of the alphabet size, and accuracy the problems of entropy and support estimation. Before we explain the differences of the two approaches, we briefly explain their approach.

Define, $\varphi_\mu(X^n)$ to be the number of elements that appear μ times. For example, when $X^n = a\ b\ r\ a\ c\ a\ d\ a\ b\ r\ a$, $\varphi_1 = 2, \varphi_2 = 2$, and $\varphi_5 = 1$. (Valiant & Valiant, 2011a) design a linear program that uses SML for high values of μ, and formulate a linear program to find a distribution for which $\mathbb{E}[\varphi_\mu]$'s are *close* to the observed φ_μ's. They then plug-in this estimate to estimate the property. On the other hand, our approach, by the nature of ML principle, tries to find the distribution that best explains the entire profile of the observed data, not just some partial characteristics. It therefore has the potential to estimate any symmetric property and estimate the distribution closely in any distance measures, competitive with the best possible. For example, the guarantees of the linear program approach are suboptimal in terms of the desired accuracy ε. For entropy estimation the optimal dependence is $\frac{1}{\varepsilon}$, whereas (Valiant & Valiant, 2011a) yields $\frac{1}{\varepsilon^2}$. This is more prominent for support size and support coverage, which have optimal dependence of $\mathrm{polylog}(\frac{1}{\varepsilon})$, whereas (Valiant & Valiant, 2011a) gives a $\frac{1}{\varepsilon^2}$ dependence. Besides, we analyze the first method proposed for estimating symmetric properties, designed from the first principles, and show that in fact it is competitive with the optimal estimators for various problems.

5. Proof Outline

Our arguments have two components. In Section 6 we prove a general result for the performance of plug-in estimation via maximum likelihood approaches.

Let $\mathcal{P}$ be a class of distributions over $\mathcal{Z}$, and $f : \mathcal{P} \to \mathbb{R}$ be a function. For $z \in \mathcal{Z}$, let

$$p_z \stackrel{\text{def}}{=} \arg\max_{p \in \mathcal{P}} p(z)$$

be the maximum-likelihood estimator of z in $\mathcal{P}$. Upon observing z, $f(p_z)$ is the ML estimator of f. In Theorem 4, we show that if there is an estimator that achieves error probability δ, then the ML estimator has an error probability at most $\delta|\mathcal{Z}|$. We note that variations of this result in the asymptotic statistics were studied before (see (Lehmann & Casella, 1998)). Our contribution is to use these results in the context of symmetric properties and show sample complexity bounds in the non-asymptotic regime.

We *emphasize that*, throughout this paper $\mathcal{Z}$ will be the set of profiles of length n, and $\mathcal{P}$ will be distributions induced over profiles by length-n $i.i.d.$ samples. Therefore, we have $|\mathcal{Z}| = |\Phi^n|$. By Lemma 1, if there is a *profile based* estimator with error probability δ, then the PML approach will have error probability at most $\delta \exp(3\sqrt{n})$. Such arguments were used in hypothesis testing to show the existence of competitive testing algorithms for fundamental statistical problems (Acharya et al., 2011; 2012).

At its face value this seems like a weak result. Our second key step is to prove that for the properties we are interested, it is possible to obtain very sharp guarantees. For example, we show that if we can estimate the entropy to an accuracy $\pm\varepsilon$ with error probability $1/3$ using n samples, then we can estimate the entropy to accuracy $\pm 2\varepsilon$ with error probability $\exp(-n^{0.9})$ using only $2n$ samples. Using this sharp concentration, the new error probability term dominates $|\Phi^n|$, and we obtain our results. The arguments for sharp concentration are based on modifications to existing estimators and a new analysis. Most of these results are technical and are in the appendix.

6. Maximum Likelihood Property Estimation

We establish performance guarantees of ML property estimation in a general set-up. Recall that $\mathcal{P}$ is a collection of distributions over $\mathcal{Z}$, and $f : \mathcal{P} \to \mathbb{R}$. Given a sample Z from an unknown $p \in \mathcal{P}$, we want to estimate $f(p)$. The maximum likelihood approach is the following two-step procedure.

1. Find $p_{_Z} = \arg\max_{p\in\mathcal{P}} p(Z)$.
2. Output $f(p_{_Z})$.

We bound the performance of this approach in the following theorem.

Theorem 3. *Suppose there is an estimator $\hat{f} : \mathcal{Z} \to \mathbb{R}$, such that for any p, and $Z \sim p$,*

$$\Pr\left(\left|f(p) - \hat{f}(Z)\right| > \varepsilon\right) < \delta, \tag{2}$$

then

$$\Pr\left(|f(p) - f(p_{_Z})| > 2\varepsilon\right) \le \delta \cdot |\mathcal{Z}|. \tag{3}$$

Proof. Consider symbols with $p(z) \ge \delta$ and $p(z) < \delta$ separately. A distribution p with $p(z) \ge \delta$ outputs z with probability at least δ. For (2) to hold, we must have, $\left|f(p) - \hat{f}(z)\right| < \varepsilon$. By the definition of ML, $p_{_z}(z) \ge p(z) \ge \delta$, and for (2) to hold for $p_{_z}$, $\left|f(p_{_z}) - \hat{f}(z)\right| < \varepsilon$. By the triangle inequality, for all such z,

$$|f(p) - f(p_{_z})| \le \left|f(p) - \hat{f}(z)\right| + \left|f(p_{_z}) - \hat{f}(z)\right| \le 2\varepsilon.$$

Thus if $p(z) \ge \delta$, then PML satisfies the required guarantee with zero probability of error, and any error occurs only when $p(z) < \delta$. We bound this probability as follows. When $Z \sim p$,

$$\Pr\left(p(Z) < \delta\right) \le \sum_{z\in\mathcal{Z}:p(z)<\delta} p(z) < \delta \cdot |\mathcal{Z}|. \qquad \square$$

For some problems, it might be easier to just approximate the ML, instead of finding it exactly. We define an approximation ML as follows:

Definition 1 (β-approximate ML)**.** *Let $\beta \le 1$. For $Z \in \mathcal{Z}$, $\tilde{p}_Z \in \mathcal{P}$ is a β-approximate ML distribution if $\tilde{p}_z(z) \ge \beta \cdot p_{_z}(z)$. When $\mathcal{Z}$ is profiles of length-n, a β-approximate PML is a distribution $\tilde{p}_\varphi$ such that $\tilde{p}_\varphi(\varphi) \ge \beta \cdot p_\varphi(\varphi)$.*

The next result proves guarantees for any β-approximate ML estimator.

Theorem 4. *Suppose there exists an estimator satisfying (2). For any $p \in \mathcal{P}$ and $Z \sim p$, any β-approximate ML $\tilde{p}_Z$ satisfies:*

$$\Pr\left(|f(p) - f(\tilde{p}_Z)| > 2\varepsilon\right) \le \delta \cdot |\mathcal{Z}|/\beta.$$

The proof is very similar to the previous theorem and is presented in the Appendix B.

6.1. Competitiveness of ML via Median Trick

Suppose for a property $f(p)$, there is an estimator with sample complexity n that achieves an accuracy $\pm\varepsilon$ with probability of error at most $1/3$. The standard method to boost the error probability is the median trick: (i) Obtain $O(\log(1/\delta))$ independent estimates using $O(n\log(1/\delta))$ independent samples. (ii) Output the median of these estimates. This is an ε-accurate estimator of $f(p)$ with error probability at most δ. By definition, estimators are a mapping from the samples to $\mathbb{R}$. However, in many applications the estimators map from a much smaller (some sufficient statistic) of the samples. Denote by Z_n the space consisting of all sufficient statistics that the estimator uses. For example, estimators for symmetric properties, such as entropy typically use the profile of the sequence, and hence $Z_n = \Phi^n$. Using the median-trick, we get the following result.

Corollary 1. *Let $\hat{f} : Z_n \to \mathbb{R}$ be an estimator of $f(p)$ with accuracy ε and error-probability $1/3$. The ML estimator achieves accuracy 2ε with probability at least 2/3 using*

$$\min\left\{n' : \frac{n'}{20\log(3Z_{n'})}\right\} > n \text{ samples.}$$

Proof. Since n is the number of samples to get error probability $1/3$, by the Chernoff bound, the error after n' samples is at most $\exp(-(n'/(20n)))$. Therefore, the error probability of the ML estimator for accuracy 2ε is at most $\exp(-(n'/(20n)))Z_{n'}$, which we desire to be at most $1/3$. $\square$

For estimators that use the profile of sequences, $|\Phi^n| < \exp(3\sqrt{n})$. Plugging this in the previous result shows that the PML based approach has a sample complexity of at most $\mathcal{O}(n^2)$. This result holds for all symmetric properties, independent of ε, and the alphabet size k. For the problems mentioned earlier, something much better in possible, namely the PML approach is optimal up to constant factors.

7. Sample optimality of PML

7.1. Sharp Concentration for Some Properties

To obtain sample-optimality guarantees for PML, we need to drive the error probability down much faster than the median trick. We achieve this by using McDiarmid's inequality stated below. Let $\hat{f} : \mathcal{X}^* \to \mathbb{R}$. Suppose $\hat{f}$ gets n independent samples X^n from an unknown distribution. Moreover, changing one of the X_j to any X'_j changed $\hat{f}$ by

at most c_*. Then McDiarmid's inequality (bounded difference inequality, (Boucheron et al., 2013)) states that,

$$\Pr\left(\left|\hat{f}(X^n) - \mathbb{E}[\hat{f}(X^n)]\right| > t\right) \le 2\exp\left(-\frac{2t^2}{nc_*^2}\right). \quad (4)$$

This inequality can be used to show strong error probability bounds for many problems. We mention a simple application for estimating discrete distributions.

Example 2. *It is well known (Devroye & Lugosi, 2001) that SML requires $\Theta(k/\varepsilon^2)$ samples to estimate p in ℓ_1 distance with probability at least $2/3$. In this case, $\hat{f}(X^n) = \sum_x \left|\frac{N_x}{n} - p(x)\right|$, and therefore c_* is at most $2/n$. Using McDiarmid's inequality, it follows that SML has an error probability of $\delta = 2\exp(-k/2)$, while still using $\Theta(k/\varepsilon^2)$ samples.*

Let B_n be the bias of an estimator $\hat{f}(X^n)$ of $f(p)$, namely $B_n \stackrel{\text{def}}{=} \left|f(p) - \mathbb{E}[\hat{f}(X^n)]\right|$. By the triangle inequality,

$$\begin{aligned}&\left|f(p) - \hat{f}(X^n)\right| \\ &\le \left|f(p) - \mathbb{E}[\hat{f}(X^n)]\right| + \left|\hat{f}(X^n) - \mathbb{E}[\hat{f}(X^n)]\right| \\ &= B_n + \left|\hat{f}(X^n) - \mathbb{E}[\hat{f}(X^n)]\right|.\end{aligned}$$

Plugging this in (4),

$$\Pr\left(\left|f(p) - \hat{f}(X^n)]\right| > t + B_n\right) \le 2\exp\left(-\frac{2t^2}{nc_*^2}\right). \quad (5)$$

With this in hand, we need to show that c_* can be bounded for estimators for the properties we consider. In particular, we will show that

Lemma 2. *Let $\alpha > 0$ be a fixed constant. For entropy, support, support coverage, and distance to uniformity there exist profile based estimators that use the optimal number of samples (given in Table 1), have bias ε and if we change any of the samples, changes by at most $c \cdot \frac{n^\alpha}{n}$, where c is a positive constant.*

We prove this lemma by proposing several modifications to the existing sample-optimal estimators. The modified estimators will preserve the sample complexity up to constant factors and also have a small c_*. The proof details are given in the appendix.

Using (5) with Lemma 2,

Theorem 5. *Let n be the optimal sample complexity of estimating entropy, support, support coverage and distance to uniformity (given in table 1) and c be a large positive constant. Let $\varepsilon \ge c/n^{0.2}$, then any for any $\beta > \exp(-\sqrt{n})$, the β-PML estimator estimates entropy, support, support coverage, and distance to uniformity to an accuracy of 4ε with probability at least $1 - \exp(-\sqrt{n})$.*

Proof. Let $\alpha = 0.05$. By Lemma 2, for each property of interest, there are estimators based on the profiles of the samples such that using near-optimal number of samples, they have bias ε and maximum change if we change any of the samples is at most $c'n^\alpha/n$ for some constant c'. Hence, by McDiarmid's inequality, an accuracy of 2ε is achieved with probability at least $1 - 2\exp\left(-2\varepsilon^2 n^{1-2\alpha}/c'^2\right)$. Now suppose $\tilde{p}$ is any β-approximate PML distribution. Then by Theorem 4

$$\begin{aligned}\Pr\left(|f(p) - f(\tilde{p})| > 4\varepsilon\right) &< \frac{\delta \cdot |\Phi^n|}{\beta} \\ &\le \frac{2\exp(-2\varepsilon^2 n^{1-2\alpha}/c'^2 + 3\sqrt{n})}{\beta} \\ &\le \exp(-\sqrt{n}),\end{aligned}$$

where in the last step we used $\varepsilon^2 n^{1-2\alpha} \gtrsim c'\sqrt{n}$, and $\beta > \exp(-\sqrt{n})$. □

8. Support and Support Coverage

We analyze both support coverage and the support estimation via a single approach. We first start with support coverage. Recall that the goal is to estimate $S_m(p)$, the expected number of distinct symbols that we see after observing m samples from p. By the linearity of expectation,

$$S_m(p) = \sum_{x\in\mathcal{X}} \mathbb{E}[\mathbb{I}_{N_x(X^m)>0}] = \sum_{x\in\mathcal{X}} \left(1 - (1-p(x))^m\right).$$

The problem is closely related to the support coverage problem (Orlitsky et al., 2016), where the goal is to estimate $U_t(X^n)$, the number of new distinct symbols that we observe in $n \cdot t$ additional samples. Hence

$$S_m(p) = \mathbb{E}\left[\sum_{i=1}^n \varphi_i\right] + \mathbb{E}[U_t],$$

where $t = (m-n)/n$. We show that the modification of an estimator in (Orlitsky et al., 2016) is also near-optimal and satisfies conditions in Lemma 2. We propose to use the following estimator

$$\hat{S}_m(p) = \sum_{i=1}^n \varphi_i + U_t^{\text{SGT}}(X^n),$$

where $U_t^{\text{SGT}}(X^n) = \sum_{i=1}^n \varphi_i(-t)^i \Pr(Z \ge i)$ and Z is a Poisson random variable with mean r and $t = (m-n)/n$.

The above theorem also works for any $\varepsilon \gtrsim 1/n^{0.25-\eta}$ for any $\eta > 0$

We remark that the proof also holds for Binomial smoothed random variables as discussed in (Orlitsky et al., 2016).

We need to bound the maximum coefficient and the bias to apply Lemma 2. We first bound the maximum coefficient of this estimator.

Lemma 3. *For all $n \le m/2$, the maximum coefficient of $\hat{S}_m(p)$ is at most $1 + e^{r(t-1)}$.*

Proof. For any i, the coefficient of φ_i is $1 + (-t)^i \Pr(Z \ge i)$. It can be upper bounded as $1 + \sum_{i=0}^{t} \frac{e^{-r}(rt)^i}{i!} = 1 + e^{r(t-1)}$. □

The next lemma bounds the bias of the estimator.

Lemma 4. *For all $n \le m/2$, the bias of the estimator is bounded by*

$$|\mathbb{E}[\hat{S}_m(p)] - S_m(p)| \le 2 + 2e^{r(t-1)} + \min(m, S(p))e^{-r}.$$

Proof. As before let $t = (m-n)/n$.

$$\begin{aligned}
&\mathbb{E}[\hat{S}_m(p)] - S_m(p) \\
&= \sum_{i=1}^{n} \mathbb{E}[\varphi_i] + \mathbb{E}[U_t^{\mathrm{SGT}}(X^n)] - \sum_{x \in \mathcal{X}} (1 - (1-p(x))^m) \\
&= \mathbb{E}[U_t^{\mathrm{SGT}}(X^n)] - \sum_{x \in \mathcal{X}} \left((1-p(x))^n - (1-p(x))^m\right).
\end{aligned}$$

Hence by Lemma 8 and Corollary 2, in (Orlitsky et al., 2016), we get

$$|\mathbb{E}[\hat{S}_m(p)] - S_m(p)| \le 2 + 2e^{r(t-1)} + \min(m, S(p))e^{-r}.$$

□

Using the above two lemmas we prove results for both the observed support coverage and support estimator.

8.1. Support Coverage Estimator

Recall that the quantity of interest in support coverage estimation is $S_m(p)/m$, which we wish to estimate to an accuracy of ε.

Proof of Lemma 2 for observed. If we choose $r = \log \frac{3}{\varepsilon}$, then by Lemma 3, the maximum coefficient of $\hat{S}_m(p)/m$ is at most $\frac{2}{m} e^{\frac{m}{n} \log \frac{3}{\varepsilon}}$, which for $m \le \alpha \frac{n \log(n/2^{1/\alpha})}{\log(3/\varepsilon)}$ is at most $n^\alpha/m < n^\alpha/n$. Similarly, by Lemma 4,

$$\frac{1}{m}|\mathbb{E}[\hat{S}_m(p)] - S_m(p)| \le \frac{1}{m}(2 + 2e^{r(t-1)} + me^{-r}) \le \varepsilon,$$

for all $\varepsilon > 6n^\alpha/n$. □

8.2. Support Estimator

Recall that the quantity of interest in support estimation is $\tilde{S}(p)$, which we wish to estimate to an accuracy of ε.

Proof of Lemma 2 for support. Note that we are interested in distributions with all the non zero probabilities are at least $1/k$. We propose to estimate $\tilde{S}(p)$ using $\hat{S}_m(p)/k$, for $m = k \log \frac{3}{\varepsilon}$. If we choose $r = \log \frac{3}{\varepsilon}$, then by Lemma 3, the maximum coefficient of $\hat{S}_m(p)/k$ is at most $\frac{2}{k} e^{\frac{m}{n} \log \frac{3}{\varepsilon}}$, which for $n \ge \frac{k}{\alpha \log(k/2^{1/\alpha})} \log^2 \frac{3}{\varepsilon}$ is at most $k^\alpha/k < n^\alpha/n$.

To bound the bias, note that for this choice of m

$$\begin{aligned}
0 \le S(p) - S_m(p) &= \sum_x (1-p(x))^m \\
&\le \sum_x e^{-mp(x)} \le k e^{-\log \frac{3}{\varepsilon}} = \frac{k\varepsilon}{3}.
\end{aligned}$$

Similarly, by Lemma 4,

$$\begin{aligned}
&\frac{1}{k}|\mathbb{E}[\hat{S}_m(p)] - S(p)| \\
&\le \frac{1}{k}|\mathbb{E}[\hat{S}_m(p)] - S_m(p)| + \frac{1}{k}|S(p) - S_m(p)| \\
&\le \frac{1}{k}(2 + 2e^{r(t-1)} + ke^{-r}) + \frac{\varepsilon}{3} \le \varepsilon,
\end{aligned}$$

for all $\varepsilon > 12n^\alpha/n$. □

9. Discussion and Future Directions

We studied estimation of symmetric properties of discrete distributions using the principle of maximum likelihood, and proved optimality of this approach for a number of problems. A number of directions are of interest. We believe that the lower bound requirement on ε is perhaps an artifact of our proof technique, and that the PML based approach is indeed optimal for all ranges of ε. Approximation algorithms for estimating the PML distributions would be a fruitful direction to pursue. Given our results, approximations stronger than $\exp(-\varepsilon^2 n)$ would be very interesting. In the particular case when the desired accuracy is a constant, even an exponential approximation would be sufficient for many properties. We plan to apply the heuristics proposed by (Vontobel, 2012) for various problems we consider, and compare with the state of the art provable methods.

Acknowledgements

The authors thank the reviewers for valuable feedback and the NSF for support through grants CIF-1564355, CIF-1619448, CRII-CIF-1657471, and a Cornell University start-up grant. Jayadev Acharya thanks Clement Canonne, Jiantao Jiao, and Pascal Vontobel for interesting discussions.

References

Acharya, Jayadev, Das, Hirakendu, Mohimani, Hosein, Orlitsky, Alon, and Pan, Shengjun. Exact calculation of pattern probabilities. In *ISIT*, pp. 1498 –1502, 2010.

Acharya, Jayadev, Das, Hirakendu, Jafarpour, Ashkan, Orlitsky, Alon, and Pan, Shengjun. Competitive closeness testing. *COLT*, 19:47–68, 2011.

Acharya, Jayadev, Das, Hirakendu, Jafarpour, Ashkan, Orlitsky, Alon, Pan, Shengjun, and Suresh, Ananda Theertha. Competitive classification and closeness testing. In *COLT*, 2012.

Acharya, Jayadev, Jafarpour, Ashkan, Orlitsky, Alon, and Suresh, Ananda Theertha. Optimal probability estimation with applications to prediction and classification. In *COLT*, 2013a.

Acharya, Jayadev, Jafarpour, Ashkan, Orlitsky, Alon, and Suresh, Ananda Theertha. A competitive test for uniformity of monotone distributions. In *AISTATS*, 2013b.

Acharya, Jayadev, Orlitsky, Alon, Suresh, Ananda Theertha, and Tyagi, Himanshu. The complexity of estimating Rényi entropy. In *SODA*, 2015.

Aldrich, John. R.a. fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3):162–176, 09 1997.

Anevski, Dragi, Gill, Richard D, and Zohren, Stefan. Estimating a probability mass function with unknown labels. *arXiv preprint arXiv:1312.1200*, 2013.

Bar-Yossef, Ziv, Kumar, Ravi, and Sivakumar, D. Sampling algorithms: lower bounds and applications. In *Symposium on Theory of computing*, pp. 266–275. ACM, 2001.

Batu, Tugkan, Fortnow, Lance, Rubinfeld, Ronitt, Smith, Warren D., and White, Patrick. Testing that distributions are close. In *FOCS*, pp. 259–269, 2000.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013.

Bu, Yuheng, Zou, Shaofeng, Liang, Yingbin, and Veeravalli, Venugopal V. Estimation of KL divergence between large-alphabet distributions. In *ISIT*, 2016.

Caferov, Cafer, Kaya, Barış, ODonnell, Ryan, and Say, AC Cem. Optimal bounds for estimating entropy with pmf queries. In *International Symposium on Mathematical Foundations of Computer Science*, pp. 187–198. Springer, 2015.

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

Canonne, Clément L. A survey on distribution testing: Your data is big. but is it blue? *Electronic Colloquium on Computational Complexity (ECCC)*, 22:63, 2015.

Colwell, Robert K, Chao, Anne, Gotelli, Nicholas J, Lin, Shang-Yi, Mao, Chang Xuan, Chazdon, Robin L, and Longino, John 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, Thomas M. and Thomas, Joy A. *Elements of information theory (2. ed.)*. Wiley, 2006.

Devroye, Luc and Lugosi, Gábor. *Combinatorial methods in density estimation*. Springer, 2001.

Good, IJ and Toulmin, GH. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

Hardy, Godfrey H and Ramanujan, Srinivasa. Asymptotic formulaæ in combinatory analysis. *Proceedings of the London Mathematical Society*, 2(1):75–115, 1918.

Jiao, Jiantao, Venkat, Kartik, Han, Yanjun, and Weissman, Tsachy. Maximum likelihood estimation of functionals of discrete distributions. *arXiv preprint arXiv:1406.6959*, 2014.

Jiao, Jiantao, Venkat, Kartik, Han, Yanjun, and Weissman, Tsachy. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

Jiao, Jiantao, Han, Yanjun, and Weissman, Tsachy. Minimax estimation of the L1 distance. In *ISIT*, pp. 750–754, 2016.

Le Cam, Lucien Marie. *Maximum likelihood: an introduction*. JSTOR, 1979.

Lehmann, Erich Leo and Casella, George. *Theory of point estimation*, volume 31. Springer Science & Business Media, 1998.

Obremski, Maciej and Skorski, Maciej. Renyi entropy estimation revisited. In *APPROX*, 2017.

Orlitsky, A., Pan, S., Sajama, Santhanam, P., Viswanathan, K., and Zhang, J. Pattern maximum likelihood: Computation and experiments. Arxiv, 2017a.

Orlitsky, Alon and Pan, Shengjun. The maximum likelihood probability of skewed patterns. In *ISIT*, 2009.

Orlitsky, Alon and Suresh, Ananda Theertha. Competitive distribution estimation: Why is good-turing good. In *NIPS*, pp. 2143–2151, 2015.

Orlitsky, Alon, Santhanam, Narayana P., and Zhang, Junan. Always good turing: Asymptotically optimal probability estimation. In *FOCS*, 2003.

Orlitsky, Alon, Sajama, S, Santhanam, NP, Viswanathan, K, and Zhang, Junan. Algorithms for modeling distributions over large alphabets. In *ISIT*, 2004a.

Orlitsky, Alon, Santhanam, Narayana P., Viswanathan, Krishnamurthy, and Zhang, Junan. On modeling profiles instead of values. In *UAI*, 2004b.

Orlitsky, Alon, Santhanam, Narayana P, and Zhang, Junan. Universal compression of memoryless sources over unknown alphabets. *IEEE Transactions on Information Theory*, 50(7):1469–1481, 2004c.

Orlitsky, Alon, Santhanam, Narayana Prasad, Viswanathan, Krishna, and Zhang, Junan. Low (size) and order in distribution modeling. 2004d.

Orlitsky, Alon, Santhanam, Narayana, Viswanathan, Krishnamurthy, and Zhang, Junan. Convergence of profile based estimators. In *Proceedings of the 2005 IEEE International Symposium on Information Theory (ISIT)*, pp. 1843 –1847, 2005.

Orlitsky, Alon, Santhanam, Narayana Prasad, Viswanathan, Krishna, and Zhang, Junan. Theoretical and experimental results on modeling low probabilities. In *Information Theory Workshop*, 2006.

Orlitsky, Alon, Suresh, Ananda Theertha, and Wu, Yihong. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 2016. doi: 10.1073/pnas.1607774113.

Orlitsky, Alon, Santhanam, Narayana, Viswanathan, Krishnamurthy, and Zhang, Junan. On estimating the probability multiset part i: The pattern maximum likelihood approach. Arxiv, 2017b.

Orlitsky, Alon, Santhanam, Narayana, Viswanathan, Krishnamurthy, and Zhang, Junan. On estimating the probability multiset part ii: Properties of the pattern maximum likelihood estimator. Arxiv, 2017c.

Pan, Shengjun. *On the theory and application of pattern maximum likelihood*. PhD thesis, UC San Diego, 2012.

Pan, Shengjun, Acyarya, Jayadev, and Orlitsky, Alon. The maximum likelihood probability of unique-singleton, ternary, and length-7 patterns. pp. 1135–1139, 2009.

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

Raskhodnikova, Sofya, Ron, Dana, Shpilka, Amir, and Smith, Adam. Strong lower bounds for approximating distribution support size and the distinct elements problem. *SIAM Journal on Computing*, 39(3):813–842, 2009.

Timan, A. F. *Theory of Approximation of Functions of a Real Variable*. Pergamon Press, 1963.

Valiant, Gregory and Valiant, Paul. Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new clts. In *STOC*, 2011a.

Valiant, Gregory and Valiant, Paul. The power of linear estimators. In *FOCS*, pp. 403–412. IEEE, 2011b.

Valiant, Gregory and Valiant, Paul. Instance-by-instance optimal identity testing. *Electronic Colloquium on Computational Complexity (ECCC)*, 20:111, 2013.

Valiant, Gregory John. *Algorithmic approaches to statistical questions*. PhD thesis, University of California, Berkeley, 2012.

Vatedka, Shashank and Vontobel, Pascal O. Pattern maximum likelihood estimation of finite-state discrete-time markov chains. In *ISIT*, 2016.

Vontobel, Pascal O. The bethe approximation of the pattern maximum likelihood distribution. In *IEEE ISIT*, pp. 2012–2016, 2012.

Vontobel, Pascal O. The bethe and sinkhorn approximations of the pattern maximum likelihood estimate and their connections to the valiant-valiant estimate. In *Information Theory and Applications Workshop, ITA*, pp. 1–10, 2014.

Wu, Yihong and Yang, Pengkun. Chebyshev polynomials, moment matching, and optimal estimation of the unseen. *CoRR*, abs/1504.01227, 2015.

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

Zou, James, Valiant, Gregory, Valiant, Paul, Karczewski, Konrad, Chan, Siu On, Samocha, Kaitlin, Lek, Monkol, Sunyaev, Shamil, Daly, Mark, and MacArthur, Daniel G. Quantifying unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects. *Nature Communications*, 7:13293 EP, 10, 2016.

A. Entropy and Distance to Uniformity

The known optimal estimators for entropy and distance to uniformity both depend on the best polynomial approximation of the corresponding functions and the splitting trick (Wu & Yang, 2016; Jiao et al., 2015). Building on their techniques, we show that a slight modification of their estimators satisfy conditions in Lemma 2. Both these functions can be written as functionals of the form:

$$f(p) = \sum_x g(p(x)),$$

where $g(y) = -y \log y$ for entropy and $g(y) = \left|y - \frac{1}{k}\right|$ for uniformity.

Both (Wu & Yang, 2016; Jiao et al., 2015) first approximate $g(y)$ with $P_{L,g}(y)$ polynomial of some degree L. Clearly a larger degree implies a smaller bias/approximation error, but estimating a higher degree polynomial also implies a larger statistical estimation error. Therefore, the approach is the following:

- For small values of $p(x)$, we estimate the polynomial $P_{L,g}(p(x)) = \sum_{i=1}^{L} b_i \cdot (p(x))^i$.
- For large values of $p(x)$ we simply use the empirical estimator for $g(p(x))$.

However, it is not a priori known which symbols have high probability and which have low probability. Hence, they both assume that they receive $2n$ samples from p. They then divide them into two set of samples, $X_1^{'}, \ldots, X_n^{'}$, and $X_1, \ldots, X_n$. Let $N_x^{'}$, and N_x be the number of appearances of symbol x in the first and second half respectively. They propose to use the estimator of the following form:

$$\hat{g}(X_1^{2n}) = \max\left\{\min\left\{\sum_x g_x, f_{\max}\right\}, 0\right\}.$$

where $f_{\max}$ is the maximum value of the property f and

$$g_x = \begin{cases} G_{L,g}(N_x), \text{ for } N_x^{'} < c_2 \log n, \text{ and } N_x < c_1 \log n, \\ g\left(\frac{N_x}{n}\right), \text{ for } N_x^{'} < c_2 \log n, \text{ and } N_x \geq c_1 \log n, \\ g\left(\frac{N_x}{n}\right) + g_n, \text{ for } N_x^{'} \geq c_2 \log n, \end{cases}$$

where g_n is the first order bias correction term for g, $G_{L,g}(N_x) = \sum_{i=1}^{L} b_i N_x^{\underline{i}}/n^{\underline{i}}$ is the unbiased estimator for $P_{L,g}$, and c_1 and c_2 are two constants which we decide later. We remark that unlike previous works, we set g_x to 0 for some values of N_x and N_x' to ensure that c^* is bounded. The following lemma bounds c^* for any such estimator $\hat{g}$.

Lemma 5. *For any estimator $\hat{g}$ defined as above, changing any one of the values changes the estimator by at most*

$$8 \max\left(e^{L^2/n} \max|b_i|, \frac{L_g}{n}, g\left(\frac{c_1 \log(n)}{n}\right), g_n\right),$$

where $L_g = n \max_{i \in \mathbb{N}} |g(i/n) - g((i-1)/n)|$.

A.1. Entropy

The following lemma is adapted from Proposition 4 in (Wu & Yang, 2016) where we make the constants explicit.

Lemma 6. *Let $g_n = 1/(2n)$ and $\alpha > 0$. Suppose $c_1 = 2c_2$, and $c_2 > 35$, Further suppose that $n^3\left(\frac{16c_1}{\alpha^2} + \frac{1}{c_2}\right) > \log k \cdot \log n$. There exists a polynomial approximation of $-y \log y$ with degree $L = 0.25\alpha$, over $[0, c_1 \frac{\log n}{n}]$ such that $\max_i |b_i| \leq n^\alpha/n$ and the bias of the entropy estimator is at most $\mathcal{O}\left(\left(\frac{c_1}{\alpha^2} + \frac{1}{c_2} + \frac{1}{n^{3.9}}\right)\frac{k}{n \log n}\right)$.*

Proof. Our estimator is similar to that of (Wu & Yang, 2016; Jiao et al., 2016) except for the case when $N_x^{'} < c_2 \log n$, and $N_x > c_1 \log n$. For any $p(x)$, and $N_x^{'}$ and N_x both distributed $Bin(np(x))$. By the Chernoff bounds for binomial distributions, the probability of this event can be bounded by,

$$\max_{p(x)} \Pr\left(N_x^{'} < c_2 \log n, N_x > 2c_2 \log n\right) \leq \frac{1}{n^{0.1\sqrt{2}c_2}} \leq \frac{1}{n^{4.9}}.$$

Therefore, the additional bias the modification introduces is at most $k \log k/n^{4.9}$ which is smaller than the bias term of (Wu & Yang, 2016; Jiao et al., 2016).

The largest coefficient can be bounded by using that the best polynomial approximation of degree L of $x \log x$ in the interval $[0, 1]$ has all coefficients at most 2^{3L}. Therefore, the largest change we have (after appropriately normalizing) is the largest value of b_i which is

$$\frac{2^{3L} e^{L^2/n}}{n}.$$

For $L = 0.25\alpha \log n$, this is at most $\frac{n^a}{n}$. □

The proof of Lemma 2 for entropy follows from the above lemma and Lemma 5 and by substituting $n = \mathcal{O}\left(\frac{k}{\log k}\frac{1}{\varepsilon}\right)$.

A.2. Distance to Uniformity

We state the following result stated in (Jiao et al., 2016).

Lemma 7. *Let $c_1 > 2c_2$, $c_2 = 35$. There is an estimator for distance to uniformity that changes by at most n^α/n when a sample is changed, and the bias of the estimator is at most $O(\frac{1}{\alpha}\sqrt{\frac{c_1 \log n}{k \cdot n}})$.*

Proof. Estimating the distance to uniformity has two regions based on N_x' and N_x.

Case 1: $\frac{1}{k} < c_2 \log n/n$**.** In this case, we use the estimator defined in the last section for $g(x) = |x - 1/k|$.

Case 2: $\frac{1}{k} > c_2 \log n/n$**.** In this case, we have a slight change to the conditions under which we use various estimators.

- For $\left|N'_x - \frac{1}{k}\right| < \sqrt{\frac{c_2 \log n}{kn}}$, & $\left|N_x - \frac{1}{k}\right| < \sqrt{\frac{c_1 \log n}{kn}}$: $g_x = G_{L,g}(N_x)$,
- For $\left|N'_x - \frac{1}{k}\right| < \sqrt{\frac{c_2 \log n}{kn}}$, & $\left|N_x - \frac{1}{k}\right| < \sqrt{\frac{c_1 \log n}{kn}}$: $g_x = 0$,
- For $\left|N'_x - \frac{1}{k}\right| \geq \sqrt{\frac{c_2 \log n}{kn}}$: $g_x = \left(\frac{N_x}{n}\right)$.

The estimator proposed in (Jiao et al., 2016) is slightly different, assigning $G_{L,g}(N_x)$ for the first two cases. We design the second case to bound the maximum deviation. The bias of their estimator was shown to be at most $\mathcal{O}\left(\frac{1}{L}\sqrt{\frac{\log n}{k \cdot n \log n}}\right)$, which can be shown by using Equation Equation 7.2.2 of (Timan, 1963)

$$E_{|x-\tau|,L,[0,1]} \leq O\left(\frac{\sqrt{\tau(1-\tau)}}{L}\right). \tag{6}$$

By our choice of c_1, c_2, our modification changes the bias by at most $1/n^4 < \varepsilon^2$.

To bound the largest deviation, we use the fact from Lemma 2 in (Cai et al., 2011) that the largest coefficient of the best degree-L polynomial approximation of $|x|$ in $[-1, 1]$ has all coefficients at most 2^{3L}. Similar argument as with entropy yields that after appropriate normalization, the largest difference in estimation will be at most n^α/n. □

The proof of Lemma 2 for entropy follows from the above lemma and Lemma 5 and by substituting $n = \mathcal{O}\left(\frac{k}{\log k}\frac{1}{\varepsilon^2}\right)$.

B. Proof of Approximate ML Performance

Proof. We consider symbols such that $p(z) \geq \delta/\beta$ and $p(z) < \delta/\beta$ separately. For an z with $p(z) \geq \delta/\beta$, by the definition of $f(p_{_Z})$,

$$\tilde{p}_z(z) \geq p_{_z}(z)\beta \geq p(z)\beta \geq \delta.$$

Applying (2) to $\tilde{p}_z$, we have for $Z \sim \tilde{p}_z$,

$$\begin{aligned}\delta &> \Pr\left(\left|f(\tilde{p}_z) - \hat{f}(Z)\right| > \varepsilon\right)\\ &\geq \tilde{p}_z(z) \cdot \mathbb{I}\left\{\left|f(\tilde{p}_z) - \hat{f}(z)\right| > \varepsilon\right\}\\ &\geq \delta \cdot \mathbb{I}\left\{\left|f(\tilde{p}_z) - \hat{f}(z)\right| > \varepsilon\right\},\end{aligned}$$

where $\mathbb{I}$ is the indicator function, and therefore, $\mathbb{I}\left\{\left|f(\tilde{p}_z) - \hat{f}(z)\right| > \varepsilon\right\} = 0$. This implies that $\left|f(\tilde{p}_z) - \hat{f}(z)\right| < \varepsilon$. By an identical reasoning, since $p(z) > \delta/\beta$, we have $\left|f(p) - \hat{f}(z)\right| < \varepsilon$. By the triangle inequality,

$$|f(p) - f(\tilde{p}_z)| \leq \left|f(p) - \hat{f}(z)\right| + \left|f(\tilde{p}_z) - \hat{f}(z)\right| < 2\varepsilon.$$

Thus if $p(z) \geq \delta/\beta$, then PML satisfies the required guarantee with zero probability of error, and any error occurs only when $p(z) < \delta/\beta$. We bound this probability as follows. When $Z \sim p$,

$$\Pr\left(p(Z) \leq \delta/\beta\right) \leq \sum_{z \in \mathcal{Z}: p(z) < \delta/\beta} p(z) \leq \delta \cdot |\mathcal{Z}|/\beta. \qquad \square$$