Electronic Colloquium on Computational Complexity, Report No. 186 (2016)

A Unified Maximum Likelihood Approach for Optimal Distribution Property Estimation

Jayadev Acharya[*]
Cornell University
acharya@cornell.edu

Hirakendu Das
Yahoo!
hdas@yahoo-inc.com

Alon Orlitsky
UC San Diego
alon@ucsd.edu

Ananda Theertha Suresh
Google Research
theertha@google.com

November 19, 2016

Abstract

The advent of data science has spurred interest in estimating properties of discrete distributions over large alphabets. Fundamental symmetric properties such as support size, support coverage, entropy, and proximity to uniformity, received most attention, with each property estimated using a different technique and often intricate analysis tools.

Motivated by the principle of maximum likelihood, we prove that for all these properties, a single, simple, plug-in estimator—profile maximum likelihood (PML) [1]—performs as well as the best specialized techniques. We also show that the PML approach is *competitive* with respect to any symmetric property estimation, raising the possibility that PML may optimally estimate many other symmetric properties.

1 Introduction

1.1 Property estimation

Recent machine-learning and data-science applications have motivated a new set of questions about inferring from data. A large class of these questions concerns estimating properties of the unknown underlying distribution.

Let Δ denote the collection of discrete distributions. A distribution *property* is a mapping $f : \Delta \to \mathbb{R}$. A distribution property is *symmetric* if it remains unchanged under relabeling of the domain symbols. For example, *support size*

$$S(p) = |\{x : p(x) > 0\}|,$$

or *entropy*

$$H(p) = \sum_x p(x) \log \frac{1}{p(x)}$$

[*]Supported by an MIT-Shell Energy Initiative grant, and Cornell University startup grant.

ISSN 1433-8092

are all symmetric properties which only depend on the set of values of $p(x)$'s and not what the symbols actually represent. To illustrate, the two distributions $p = (p(a) = 2/3, p(b) = 1/3)$ and $p' = (p'(a) = 1/3, p'(b) = 2/3)$ have the same entropy. In the common setting for these questions, an unknown underlying distribution $p \in \Delta$ generates n independent samples $X^n \overset{\text{def}}{=} X_1, , \ldots, X_n$, and from this *sample* we would like to estimate a given property $f(p)$.

An age-old universal approach for estimating distribution properties is *plug-in estimation.* It uses the samples X^n to find an approximation $\hat{p}$ of p, and declares $f(\hat{p})$ as the approximation $f(p)$.

Perhaps the simplest approximation for p is the *sequence maximum likelihood (SML).* It assigns to any sample x^n the distribution p that maximizes $p(x^n)$. It can be easily shown that SML is exactly the *empirical frequency* estimator that assigns to each symbol the fraction of times it appears in the sample, $p_{X^n}(x) = \frac{N_x}{n}$, where $N_x \overset{\text{def}}{=} N_x(X^n)$, the *multiplicity* of symbol x, is the number of times it appears in the sequence X^n. We will just write N_x, when X^n is clear from the context . For example, if $n = 11$, and $X^n = a\ b\ r\ a\ c\ a\ d\ a\ b\ r\ a$, $N_a = 5, N_b = 2, N_c = 1, N_d = 1$, and $N_r = 2$, and $p_{X^{11}}(a) = 5/11$, $p_{X^{11}}(b) = 2/11$, $p_{X^{11}}(c) = 1/11$, $p_{X^{11}}(d) = 1/11$, and $p_{X^{11}}(r) = 2/11$.

While the SML plug-in estimator performs well in the limit of many samples and its convergence rate falls short of the best-known property estimates. For example, suppose we sample the uniform distribution over k elements $n = k/2$ times. Since at most n distinct symbols will appear, the empirical distribution will have entropy at most $\log n \le \log k - 1$ bits. However from Table 1.3, for large k, only $n = O(k/\log k)$ samples are required to obtain a 1-bit accurate estimate.

Modern applications where the sample size n could be sub-linear in the domain size k, have motivated many results characterizing the sample complexity of estimating various distribution properties (See *e.g.*, [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]). Complementary to property estimation is the distribution property testing, which aims to design (sub-linear) algorithms to test whether distributions have some specific property (See *e.g.*, [15, 16, 17, 18, 19, 20, 21, 22], and [23] for a survey). A particular line of work is competitive distribution estimation and testing [24, 25, 26, 27, 28, 29], where the objective is to design algorithms independent of the domain size, with complexity close to the *best possible* algorithm. Some of our techniques are motivated by those in competitive testing.

1.2 Prior results

Since SML is suboptimal, several recent papers have used diverse and sophisticated techniques to estimate important symmetric distribution properties.

Support size $S(p) = |\{x : p(x) > 0\}|$, plays an important role in population and vocabulary estimation. However estimating $S(p)$ is hard with *any finite* number of samples due to symbols with *negligible* positive probability that will not appear in our sample, but still contribute to $S(p)$. To circumvent this, [30] considered distributions in Δ with non-zero probabilities at least $\frac{1}{k}$,

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

For $\Delta_{\ge \frac{1}{k}}$, SML requires $k \log\left(\frac{1}{\varepsilon}\right)$ to estimate the support size to an additive accuracy of εk. Over a series of work [30, 5, 8], it was shown that the optimal sample complexity of support estimation is $\Theta\left(\frac{k}{\log k} \cdot \log^2 \frac{1}{\varepsilon}\right)$.

Support coverage $S_m(p) = \sum_x (1 - (1 - p(x))^m)$, the expected number of elements observed when the distribution is sampled m times, arises in many ecological and biological studies [31]. The goal is to estimate $S_m(p)$ to an additive $\pm\varepsilon m$ upon observing as few samples as possible. Good and Toulmin [32] proposed an estimator that for any constant ε, requires $m/2$ samples to estimate $S_m(p)$. Recently, [12, 13] showed that it is possible to estimate $S_m(p)$ after observing only $\mathcal{O}(\frac{m}{\log m})$ samples. In particular, [12] showed that it is possible to estimate $S_m(p)$ after observing only $\mathcal{O}(\frac{m}{\log m} \cdot \log \frac{1}{\varepsilon})$ samples. Moreover, this dependence on m and ε is optimal.

Entropy $H(p) = \sum_x p(x) \log \frac{1}{p(x)}$, the Shannon entropy of p is a central object in information theory [33], and also arises in many fields such as machine learning [34], neuroscience [35, 36], and others. Entropy estimation has been studied for over half a century, and a number of different estimators have been proposed over time. Estimating $H(p)$ is hard with any finite number of samples due to the possibility of infinite support. To circumvent this, similar to previous works we consider distributions in Δ with support size at most k,

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

The goal is to estimate the entropy of a distribution in Δ_k to an additive $\pm\varepsilon$, where Δ_k is all discrete distributions over at most k symbols. In a recent set of papers [5, 7, 11], the min-max sample complexity of estimating entropy to $\pm\varepsilon$ was shown to be $\Theta\left(\frac{k}{\log k} \cdot \frac{1}{\varepsilon}\right)$.

Distance to uniform $\|p - u\|_1 = \sum_x |p(x) - 1/k|$, where u is a uniform distribution over a known set $\mathcal{X}$, with $|\mathcal{X}| = k$. Let $\Delta_{\mathcal{X}}$ be the set of distributions over the set $\mathcal{X}$. For an unknown $p \in \Delta_{\mathcal{X}}$, to estimate $||p - u||_1$ to an additive $\pm\varepsilon$, [6] showed that $\mathcal{O}\left(\frac{k}{\log k} \cdot \frac{1}{\varepsilon^2}\right)$ samples are sufficient. The dependence was later shown to be tight in [37].

[5] also proposed a plug-in approach for estimating symmetric properties. We discuss and compare the approaches in Section 3.

1.3 New results

Each of the above properties was studied in one or more papers and approximated by different sophisticated estimators, often drawing from involved techniques from fields such as approximation theory. By contrast, we show that a single simple plug-in estimator achieves the state of the art performance for all these problems.

As seen in the introduction for entropy, SML is suboptimal in the large alphabet regime, since it over-fits the estimate on only the *observed symbols* (See [38] for detailed performance of SML estimators of entropy, and other properties). However, symmetric properties of distributions do not depend on the labels of the symbols. For all these properties, it makes sense to look at a sufficient statistic, the data's *profile* (Definition 1) that represents the number of elements appearing any given number of times. Again following the *principle of maximum likelihood*, [1, 39] suggested discarding the symbol labels, and finding a distribution that maximizes the probability of the observed profile, which we call as *profile maximum likelihood (PML)*.

We show that replacing the SML plug-in estimator by PML yields a unified estimator that is provably at least as good as the best specialized techniques developed for all of the above properties.

Theorem 1 (Informal)**.** *There is a unified approach based on PML distribution that achieves the optimal sample complexity for all the four problems mentioned above (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.

However, several heuristics for estimating PML has been studied including approaches motivated by algebraic approaches [40], EM-MCMC algorithms [41], [42, Chapter 6], Bethe approximation [43, 44]. As discussed in Section 3, PML estimation reduces to maximizing a monomial-symmetric polynomial over the simplex. We also provide another justification of the PML approach by proving that even approximating a PML can result in sample-optimal estimators for the problems we consider. We hope that these strong sample complexity guarantees will motivate algorithm designers to design efficient algorithms for approximating PML. Table 1.3 summarizes the results in terms of the sample complexity.

Property	Notation	$\mathcal{P}$	SML	Best possible	References	PML
Entropy	$H(p)$	Δ_k	$\frac{k}{\varepsilon}$	$\frac{k}{\log k}\frac{1}{\varepsilon}$	[5, 7, 11]	optimal[1]
Support size	$\frac{S(p)}{k}$	$\Delta_{\geq \frac{1}{k}}$	$k \log \frac{1}{\varepsilon}$	$\frac{k}{\log k} \log^2 \frac{1}{\varepsilon}$	[8]	optimal
Support coverage	$\frac{S_m(p)}{m}$	Δ	m	$\frac{m}{\log m} \log \frac{1}{\varepsilon}$	[12]	optimal
Distance to uniform	$\|p-u\|_1$	$\Delta_{\mathcal{X}}$	$\frac{k}{\varepsilon^2}$	$\frac{k}{\log k}\frac{1}{\varepsilon^2}$	[6, 37]	optimal

Table 1: Estimation complexity for various properties, up to a constant factor. For all properties shown, PML achieves the best known results. Citations are for specialized techniques, PML results are shown in this paper. Support and support coverage results have been normalized for consistency with existing literature.

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.
- 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.

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 2. These distributions are more relaxed version of PML, hence may be more easily computed, yet they provide essentially the same performance guarantees.

The rest of the paper is organized as follows. In Section 2, we formally state our results. In Section 3, we define profiles, and PML. In Section 4, we outline the our approach. In Section 5, we

[1] We call an algorithm optimal if it is optimal up to universal constant factors.

demonstrate auxiliary results for maximum likelihood estimators. In Section 6, we outline how we apply maximum likelihood to support, entropy, and uniformity, and support coverage.

2 Formal definitions and results

Recall that Δ_k is the set of all discrete distributions with support at most k, and $\Delta = \Delta_\infty$ is the set of all discrete distributions. A property estimator is a mapping $\hat{f} : \mathcal{X}^n \to \mathbb{R}$ that converts observed samples over $\mathcal{X}$ to an estimated property value. The *sample complexity* of $\hat{f}$ when estimating a property $f : \Delta \to \mathbb{R}$ for distributions in a collection $\mathcal{P} \subseteq \Delta$, is the number of samples $\hat{f}$ needs to determine f with high accuracy and probability for all distributions in $\mathcal{P}$. Specifically, for approximation accuracy ε and confidence probability δ,

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

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

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

In the past, different sophisticated estimators were used for every property in Table 1.3. We show that the simple plug-in estimator that uses any PML approximation $\tilde{p}$, has optimal performance guarantees for all these properties.

It can be shown that the sample complexity has only moderate dependence on δ, that is typically de-emphasized. For simplicity, we therefore abbreviate $C^{\hat{f}}(f, \mathcal{P}, 1/3, \varepsilon)$ by $C^{\hat{f}}(f, \mathcal{P}, \varepsilon)$.

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 1.3) respectively.

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

Entropy

$$C^{\tilde{p}}(H(p), \Delta_k, \varepsilon) \asymp C^*(H(p), \Delta_k, \varepsilon),^{2}$$

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

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

[2] 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 $a \gtrsim b$.

3 PML: Profile maximum likelihood

3.1 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 [1] is defined below.

Definition 1. 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, $\varphi(a\ b\ r\ a\ c\ a\ d\ a\ b\ r\ a) = \{1, 1, 2, 2, 5\}$, denoting that there are 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 histogram order statistics [2], fingerprints [5], and as histograms of histograms [17].

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-Ramanujam bounds on the partition number,

Lemma 1 ([45, 1])**.** *$|\Phi^n| \le \exp(3\sqrt{n})$.*

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

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

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

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

where the normalization factor is independent of p. The summation is a monomial symmetric polynomial in the probability values. See [42, Section 2.1.2] for more examples and definitions.

3.2 Algorithm

Recall that p_{X^n} is the distribution maximizing the probability of X^n. Similarly, define [1]:

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

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 [1] to distinguish between SML and PML distributions, and also show an instance where PML outputs distributions over a larger domain 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: $\mathcal{P}$, symmetric function $f(\cdot)$, sample $X^n$
    1. Compute $p_\varphi : \arg\max_{p \in \mathcal{P}} p(\varphi(X^n))$.
    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.

We remark that various aspects of PML have been studied. [39] has a comprehensive collection of various results about PML. [46, 47] study universal compression and probability estimation using PML distributions. [39, 48, 49] derive PML distribution for various special, and small length profiles. [39, 50] prove consistency of PML. [51] study PML over Markov Chains.

Comparision to the linear-programming plug-in estimator [5]. Our approach is perhaps closest in flavor to the plug-in estimator of [5]. 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 $\varphi_5 = 1$. [5] 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 sub-optimal in terms of the desired accuracy ε. For entropy estimation the optimal dependence is $\frac{1}{\varepsilon}$, whereas [5] yields $\frac{1}{\varepsilon^2}$. This is more prominent for support size and support coverage, which have optimal dependence of $\text{polylog}(\frac{1}{\varepsilon})$, whereas [5] 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.

4 Proof outline

Our arguments have two components. In Section 5 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 [52]). 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 [24, 25, 53].

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.

5 Estimating properties via maximum likelihood

In this section, we prove the 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

$$\Pr\left(|f(p) - f(p_Z)| > 2\varepsilon\right) \leq \delta \cdot |\mathcal{Z}|\,. \tag{3}$$

Proof. Consider symbols with $p(z) \geq \delta$ and $p(z) < \delta$ separately. A distribution p with $p(z) \geq \delta$ 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, $p_z(z) \geq p(z) \geq \delta$, and again for (2) to hold for p_z, $\left|f(p_z) - \hat{f}(z)\right| < \varepsilon$. By the triangle inequality, for all such z,

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

Thus if $p(z) \geq \delta$, 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 $Z \sim p$,

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

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

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:*

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

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

6 Sample optimality of PML

We first prove that PML is *competitive* with respect to the best estimator for *any symmetric property*. This is a side-result and a direct corollary of the results in the previous section. We then provide much sharper concentration bounds for estimating the properties we are considering, and use it to prove the optimality of PML.

6.1 Median trick and competitiveness of PML

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 using*

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

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

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 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.

6.2 Sharp concentration for some interesting 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, [54, Theorem 6.2]) states that,

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

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 [55] 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

$$B_n \overset{\text{def}}{=} \left|f(p) - \mathbb{E}[\hat{f}(X^n)]\right|.$$

By the triangle inequality,

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

Plugging this in (4),

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

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 1.3), 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 1.3) and c be a large positive constant. Let $\varepsilon \geq c/n^{0.2}$, 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 $1 - \exp(-\sqrt{n})$.*[3]

Proof. Let $\alpha = 0.1$. 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 n^α/n. Hence, by McDiarmid's inequality, an accuracy of 2ε is achieved with probability at least $1 - \exp(-2\varepsilon^2 n^{1-a}/c^2)$. Now suppose $\tilde{p}$ is any β-approximate PML distribution. Then by Theorem 4

$$\Pr\left(|f(p) - f(\tilde{p})| > 4\varepsilon\right) < \frac{\delta \cdot |\Phi^n|}{\beta} \leq \frac{\exp(-2\varepsilon^2 n^{1-a}/c^2)\exp(3\sqrt{n})}{\beta} \leq \exp(-\sqrt{n}),$$

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

7 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 [43] for various problems we consider, and compare with the state of the art provable methods.

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

Acknowledgements

We thank Ashkan Jafarpour, Shashank Vatedka, Krishnamurthy Viswanathan, and Pascal Vontobel for very helpful discussions and suggestions.

References

[1] A. Orlitsky, N. P. Santhanam, K. Viswanathan, and J. Zhang, "On modeling profiles instead of values," in *UAI*, 2004. (document), 1.3, 3.1, 1, 3.2

[2] L. Paninski, "Estimation of entropy and mutual information," *Neural computation*, vol. 15, no. 6, pp. 1191–1253, 2003. 1.1, 3.1

[3] T. Batu, S. Dasgupta, R. Kumar, and R. Rubinfeld, "The complexity of approximating the entropy," *SIAM Journal on Computing*, vol. 35, no. 1, pp. 132–150, 2005. 1.1

[4] Z. Bar-Yossef, R. Kumar, and D. Sivakumar, "Sampling algorithms: lower bounds and applications," in *Proceedings of the thirty-third annual ACM symposium on Theory of computing*. ACM, 2001, pp. 266–275. 1.1

[5] G. Valiant and P. Valiant, "Estimating the unseen: an n/log(n)-sample estimator for entropy and support size, shown optimal via new clts," in *STOC*, 2011. 1.1, 1.2, 1.3, 3.1, 3.2

[6] ——, "The power of linear estimators," in *Foundations of Computer Science (FOCS), 2011 IEEE 52nd Annual Symposium on*. IEEE, 2011, pp. 403–412. 1.1, 1.2, 1.3

[7] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Trans. Information Theory*, vol. 62, no. 6, pp. 3702–3720, 2016. 1.1, 1.2, 1.3, B, B.1, B.1

[8] ——, "Chebyshev polynomials, moment matching, and optimal estimation of the unseen," *CoRR*, vol. abs/1504.01227, 2015. [Online]. Available: http://arXiv.org/abs/1504.01227 1.1, 1.2, 1.3

[9] J. Acharya, A. Orlitsky, A. T. Suresh, and H. Tyagi, "The complexity of estimating Rényi entropy," in *SODA*, 2015. 1.1

[10] C. Caferov, B. Kaya, R. O'Donnell, and A. C. Say, "Optimal bounds for estimating entropy with pmf queries," in *International Symposium on Mathematical Foundations of Computer Science*. Springer, 2015, pp. 187–198. 1.1

[11] J. Jiao, K. Venkat, Y. Han, and T. Weissman, "Minimax estimation of functionals of discrete distributions," *IEEE Transactions on Information Theory*, vol. 61, no. 5, pp. 2835–2885, 2015. 1.1, 1.2, 1.3, B

[12] A. Orlitsky, A. T. Suresh, and Y. Wu, "Optimal prediction of the number of unseen species," *Proceedings of the National Academy of Sciences*, 2016. 1.1, 1.2, 1.3, A, A

[13] J. Zou, G. Valiant, P. Valiant, K. Karczewski, S. O. Chan, K. Samocha, M. Lek, S. Sunyaev, M. Daly, and D. MacArthur, "Quantifying the unobserved protein-coding variants in human populations provides a roadmap for large-scale sequencing projects," *bioRxiv*, 2015. 1.1, 1.2

[14] Y. Bu, S. Zou, Y. Liang, and V. V. Veeravalli, "Estimation of kl divergence between large-alphabet distributions," in *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, 2016, pp. 1118–1122. 1.1

[15] T. Batu, "Testing properties of distributions," Ph.D. dissertation, Cornell University, 2001. 1.1

[16] O. Goldreich and D. Ron, "On testing expansion in bounded-degree graphs," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 7, no. 20, 2000. 1.1

[17] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing that distributions are close," in *Annual Symposium on Foundations of Computer Science (FOCS)*, 2000, pp. 259–269. 1.1, 3.1

[18] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Transactions on Information Theory*, vol. 54, no. 10, pp. 4750–4755, 2008. 1.1

[19] S. O. Chan, I. Diakonikolas, P. Valiant, and G. Valiant, "Optimal algorithms for testing closeness of discrete distributions," in *Symposium on Discrete Algorithms (SODA)*, 2014. 1.1

[20] C. L. Canonne, I. Diakonikolas, T. Gouleakis, and R. Rubinfeld, "Testing shape restrictions of discrete distributions," in *33rd Symposium on Theoretical Aspects of Computer Science*, 2016. 1.1

[21] J. Acharya, C. Daskalakis, and G. C. Kamath, "Optimal testing for properties of distributions," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 3591–3599. [Online]. Available: http://papers.nips.cc/paper/5839-optimal-testing-for-properties-of-distributions.pdf 1.1

[22] I. Diakonikolas and D. M. Kane, "A new approach for testing properties of discrete distributions," *arXiv preprint arXiv:1601.05557*, 2016. 1.1

[23] C. L. Canonne, "A survey on distribution testing: Your data is big. but is it blue?" *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 22, p. 63, 2015. [Online]. Available: http://eccc.hpi-web.de/report/2015/063 1.1

[24] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and S. Pan, "Competitive closeness testing," *Proceedings of the 24th Annual Conference on Learning Theory (COLT)*, vol. 19, pp. 47–68, 2011. 1.1, 4

[25] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, S. Pan, and A. T. Suresh, "Competitive classification and closeness testing," in *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012, pp. 22.1–22.18. 1.1, 4

[26] J. Acharya, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "Optimal probability estimation with applications to prediction and classification," in *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, 2013, pp. 764–796. 1.1

[27] ——, “A competitive test for uniformity of monotone distributions,” in *Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2013. 1.1

[28] G. Valiant and P. Valiant, “Instance-by-instance optimal identity testing,” *Electronic Colloquium on Computational Complexity (ECCC)*, vol. 20, p. 111, 2013. 1.1

[29] A. Orlitsky and A. T. Suresh, “Competitive distribution estimation: Why is good-turing good,” in *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 2015, pp. 2143–2151. 1.1

[30] S. Raskhodnikova, D. Ron, A. Shpilka, and A. Smith, “Strong lower bounds for approximating distribution support size and the distinct elements problem,” *SIAM Journal on Computing*, vol. 39, no. 3, pp. 813–842, 2009. 1.2

[31] R. K. Colwell, A. Chao, N. J. Gotelli, S.-Y. Lin, C. X. Mao, R. L. Chazdon, and J. T. Longino, “Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages,” *Journal of plant ecology*, vol. 5, no. 1, pp. 3–21, 2012. 1.2

[32] I. Good and G. Toulmin, “The number of new species, and the increase in population coverage, when a sample is increased,” *Biometrika*, vol. 43, no. 1-2, pp. 45–63, 1956. 1.2

[33] T. M. Cover and J. A. Thomas, *Elements of information theory (2. ed.)*. Wiley, 2006. 1.2

[34] S. Nowozin, “Improved information gain estimates for decision tree induction,” in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012. 1.2

[35] M. J. Berry, D. K. Warland, and M. Meister, “The structure and precision of retinal spike trains,” *Proceedings of the National Academy of Sciences*, vol. 94, no. 10, pp. 5411–5416, 1997. 1.2

[36] I. Nemenman, W. Bialek, and R. R. de Ruyter van Steveninck, “Entropy and information in neural spike trains: Progress on the sampling problem,” *Physical Review E*, vol. 69, pp. 056 111–056 111, 2004. 1.2

[37] J. Jiao, Y. Han, and T. Weissman, “Minimax estimation of the L1 distance,” in *IEEE International Symposium on Information Theory, ISIT 2016, Barcelona, Spain, July 10-15, 2016*, 2016, pp. 750–754. 1.2, 1.3, B.1, B.2, B.2

[38] J. Jiao, K. Venkat, Y. Han, and T. Weissman, “Maximum likelihood estimation of functionals of discrete distributions,” *arXiv preprint arXiv:1406.6959*, 2014. 1.3

[39] A. Orlitsky, N. Santhanam, K. Viswanathan, and J. Zhang, “On estimating the probability multiset,” Manuscript, 2011. [Online]. Available: http://www-ee.eng.hawaii.edu/~prasadsn/skelnew59.pdf 1.3, 3.2

[40] J. Acharya, H. Das, H. Mohimani, A. Orlitsky, and S. Pan, “Exact calculation of pattern probabilities,” in *Proceedings of the 2010 IEEE International Symposium on Information Theory (ISIT)*, 2010, pp. 1498 –1502. 1.3

[41] A. Orlitsky, S. Sajama, N. Santhanam, K. Viswanathan, and J. Zhang, "Algorithms for modeling distributions over large alphabets," in *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*. IEEE, 2004, pp. 304–304. 1.3

[42] S. Pan, "On the theory and application of pattern maximum likelihood," Ph.D. dissertation, UC San Diego, 2012. 1.3, 3.1

[43] P. O. Vontobel, "The bethe approximation of the pattern maximum likelihood distribution," in *Proceedings of the 2012 IEEE International Symposium on Information Theory, ISIT 2012, Cambridge, MA, USA, July 1-6, 2012*, 2012, pp. 2012–2016. 1.3, 7

[44] ——, "The bethe and sinkhorn approximations of the pattern maximum likelihood estimate and their connections to the valiant-valiant estimate," in *2014 Information Theory and Applications Workshop, ITA 2014, San Diego, CA, USA, February 9-14, 2014*, 2014, pp. 1–10. 1.3

[45] G. H. Hardy and S. Ramanujan, "Asymptotic formulaæ in combinatory analysis," *Proceedings of the London Mathematical Society*, vol. 2, no. 1, pp. 75–115, 1918. 1

[46] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1469–1481, 2004. 3.2

[47] ——, "Always good turing: Asymptotically optimal probability estimation," in *Annual Symposium on Foundations of Computer Science (FOCS)*, 2003. 3.2

[48] J. Acharya, A. Orlitsky, and S. Pan, "Recent results on pattern maximum likelihood," in *Networking and Information Theory, 2009. ITW 2009. IEEE Information Theory Workshop on*. IEEE, pp. 251–255. 3.2

[49] A. Orlitsky and S. Pan, "The maximum likelihood probability of skewed patterns," in *Proceedings of the 2009 IEEE international conference on Symposium on Information Theory-Volume 2*. IEEE Press, 2009, pp. 1130–1134. 3.2

[50] D. Anevski, R. D. Gill, and S. Zohren, "Estimating a probability mass function with unknown labels," *arXiv preprint arXiv:1312.1200*, 2013. 3.2

[51] S. Vatedka and P. O. Vontobel, "Pattern maximum likelihood estimation of finite-state discrete-time markov chains," *extended version*, 2016. 3.2

[52] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer Science & Business Media, 1998, vol. 31. 4

[53] J. Acharya, H. Das, A. Jafarpour, A. Orlitsky, and A. T. Suresh, "Tight bounds for universal compression of large alphabets," in *Proceedings of the 2013 IEEE International Symposium on Information Theory (ISIT)*, 2013. 4

[54] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. [Online]. Available: https://books.google.com/books?id=koNqWRluhP0C 6.2

[55] L. Devroye and G. Lugosi, *Combinatorial methods in density estimation*. Springer, 2001. 2

[56] A. F. Timan, *Theory of Approximation of Functions of a Real Variable.* Pergamon Press, 1963. B.2

[57] T. T. Cai, M. G. Low *et al.*, “Testing composite hypotheses, hermite polynomials and optimal estimation of a nonsmooth functional,” *The Annals of Statistics*, vol. 39, no. 2, pp. 1012–1041, 2011. B.2

A 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 [12], 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 [12] is also near-optimal and satisfies conditions in Lemma 2. We propose to use the following estimator

$$\hat{S}_m(p) = \sum_{i=1}^{n} \varphi_i + \sum_{i=1}^{n} \varphi_i (-t)^i \Pr(Z \geq i),$$

where Z is a Poisson random variable with mean r and $t = (m-n)/n$. We remark that the proof also holds for Binomial smoothed random variables as discussed in [12].

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 $n \leq m/2$, 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

$$1 + (-t)^i \Pr(Z \geq i).$$

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 [12], 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.

A.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 $\varepsilon > 6n^\alpha/n$. □

A.2 Support estimator

Recall that the quantity of interest in support estimation is $S(p)/k$, 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 $S(p)/k$ using

$$\hat{S}_m(p)/k,$$

for $m = k \log \frac{3}{\varepsilon}$. Note that for this choice of m

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

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 $n \geq \alpha \frac{k}{\log(k/2^{1/\alpha})} \log^2 \frac{3}{\varepsilon}$ is at most $k^\alpha/k < n^\alpha/n$. Similarly, by Lemma 4,

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

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

B 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 [7, 11]. 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[7, 11] 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

$$g_x = \begin{cases} G_{L,g}(N_x), & \text{for } N_x^{'} < c_2\log n, \text{ and } N_x < c_1 \log n, \\ 0, & \text{for } N_x^{'} < c_2 \log n, \text{ and } N_x \geq c_1\log n, \\ g\left(\frac{N_x}{n}\right) + g_n, & \text{for } N_x^{'} \geq c_2 \log n, \end{cases}$$

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

B.1 Entropy

The following lemma is adapted from Proposition 4 in [7] 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 $\max_i|b_i| \le n^\alpha/n$ 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 [7, 37] 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,

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

Therefore, the additional bias the modification introduces is at most $k\log k/n^{4.9}$ which is smaller than the bias term of [7, 37].

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 $\frac{n^\alpha}{n}$. □

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

B.2 Distance to uniformity

We state the following result stated in [37].

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:

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

The estimator proposed in [37] 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 [56, Equation 7.2.2]

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

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 ([57, Lemma 2]) 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 $n = \mathcal{O}\left(\frac{k}{\log k}\frac{1}{\varepsilon^2}\right)$.

C 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 $Z \sim \tilde{p}_z$,

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

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 $Z \sim p$,

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