Compressed Maximum Likelihood

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

Abstract

Maximum likelihood (ML) is one of the most fundamental and general statistical estimation techniques. Inspired by recent advances in estimating distribution functionals, we propose *compressed maximum likelihood* (CML) that applies ML to compressed samples. We show that CML is sample-efficient for several fundamental learning tasks over both discrete and continuous domains, including learning structural densities, estimating probability multisets, and inferring symmetric distribution functionals.

1. Introduction

Maximum likelihood (ML) is a strikingly simple yet fundamental statistical inference paradigm for estimating parameters in probabilistic data generalization models. Over the past century, it has been applied to derive numerous important results in mathematics, statistics, and machine learning (Friedman et al., 2001).

At its core, the ML principle favors the model that maximizes the observed data's probability. Consider for example flipping a coin with unknown heads probability p ten times and observing six heads and four tails. The observed sequence's probability is $p^6(1-p)^4$, and ML suggests that the best guess of p is $\hat{p} = 0.6$ that maximizes this probability.

ML is so natural and intuitive that it may be stumbled upon without realizing its depth and significance. Yet, it can be shown to be near-optimal under relatively mild and general regularity conditions (Van der Vaart, 2000). In particular, $\hat{p}$ tends to the actual value of p as the number of observations increases and enjoys both consistency and efficiency.

This paper concerns *functional estimation*, where $\mathcal{P}$ is a distribution collection over a domain space $\mathcal{Z}$, and $f : \mathcal{P} \to \mathcal{Q}$ is a functional, mapping distributions in $\mathcal{P}$ into a space $\mathcal{Q}$ equipped with a pseudo-metric d.

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

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

An *estimator* for f is a function $\hat{f} : \mathcal{Z} \to \mathcal{Q}$ such that when $Z \sim p$, the loss $d(\hat{f}(Z), f(p))$ is small. The ML plug-in estimator of f first finds the ML distribution estimate

$$p_z := \arg\max_{p \in \mathcal{P}} p(z)$$

and then estimates the functional's value as $f(p_z)$.

For example, if $\mathcal{P}$ is a parametric family indexed by Θ, and f maps each $p_\theta \in \mathcal{P}$ to its index $\theta \in \Theta$, the task becomes the classical *parameter estimation* problem, and ML maps sample Z to the parameter θ maximizing $p_\theta(Z)$.

1.1. Statistical Guarantees of ML Methods

ML is particularly significant in modern scientific applications that often involve complex structures in high dimensions. Deriving a bespoke estimator for each application may be an arduous task. Yet ML provides a simple universal methodology that in principle can be applied to any problem.

Unfortunately, it is well-known that for finite samples, ML estimators may be suboptimal, and other estimators often concentrate faster around the true parameters. Hence, establishing *finite-sample guarantees* for ML-based methods is of fundamental importance.

While the ML principle is quite natural, showing its finite-sample efficiency is often not easy. Recent work Acharya et al. (2017a) provided a simple argument for bypassing the difficulty in analyzing ML methods. Their primary lemma, stated next, relies only on the fact that $p_z(z) \geq p(z)$ for any $z \in \mathcal{Z}$. For simplicity, we write $d_f(p, q)$ for $d(f(p), f(q))$.

Lemma 1. *For any $\mathcal{Z}$, $\mathcal{P}$, and accuracy $\varepsilon > 0$,*

$$\max_{p \in \mathcal{P}} \Pr_{Z \sim p}(d_f(p, p_Z) > 2\varepsilon) \leq |\mathcal{Z}| \max_{p \in \mathcal{P}} \Pr_{Z \sim p}\Big(d(f(p), \hat{f}(Z)) > \varepsilon\Big).$$

Namely, the ML plug-in estimator is competitive with all other estimators in that if any other estimator achieves loss at most ε with probability $1 - \delta$ then the ML plug-in estimator will achieve a 2ε loss with probability at least $1 - |\mathcal{Z}| \cdot \delta$.

Leveraging this lemma, the paper showed that the ML-based *profile maximum likelihood (PML)* probability-multiset estimator in Orlitsky et al. (2004) is sample-optimal for several symmetric functionals, including entropy and support size.

The *profile* of a sample represents the number of elements appearing any given number of times, and hence is a sufficient statistic for symmetric functionals. Therefore instead of maximizing the sample's probability, Acharya et al. (2017a) maximized the probability of the sample's profile.

1.2. Compressed Maximum Likelihood

Motivated by the above rationale, our paper develops a general framework for applying and analyzing ML for several statistical and machine-learning problems. It does so by adding a *compressor* between the sample and ML learner.

Specifically, given domain $\mathcal{Z}$, a *compressor* is a function φ that maps each $z \in \mathcal{Z}$ to an element $\varphi(z)$ in some co-domain Φ. In general, we will allow φ to possess randomness independent of the samples.

Given compressor φ, we define the *φ-compressed maximum likelihood (CML)* estimator as

$$p_\varphi := \arg\max_{p\in\mathcal{P}} p(\varphi(Z)),$$

where for notational brevity, we suppress Z in $p_{\varphi(Z)}$.

2. Main Results

2.1. Compression for Learning

Consider an arbitrary compressor φ, mapping elements in $\mathcal{Z}$ to a co-domain Φ. We evaluate the compressor's quality for our learning objective via the following two criteria.

Typicality A compressor is (m,γ)-*typical* for an integer m and probability threshold $\gamma \in (0,1)$, if for every $p \in \mathcal{P}$, there is an m-element subset $\mathcal{T} \subseteq \Phi$ with probability

$$p(\mathcal{T}) \geq 1-\gamma.$$

Intuitively, smaller m corresponds to a better compressor.

Learnability Given error parameters ε, δ, we say that the compressor *enables* (ε,δ)-*learning* under pseudo-metric d if there is an algorithm $\mathcal{A} : \Phi \to \mathcal{Q}$ satisfying

$$\Pr_{Z\sim p}\left(d(f(p),\, \mathcal{A}(\varphi(Z))) > \varepsilon\right) \leq \delta,\ \forall p \in \mathcal{P}.$$

Intuitively, the smaller the error parameters, the better the compressor is for our learning tasks.

The theorem below shows that for any "good quality" compressor, the respective CML plug-in estimate will also be accurate, with high probability.

Theorem 1. *For any compressor φ that is (m,γ)-typical and enables (ε,δ)-learning, distribution $p \in \mathcal{P}$, and $Z \sim p$,*

$$\Pr\left(d_f(p, p_{\varphi(Z)}) > 2\varepsilon\right) \leq \gamma + m\cdot\delta.$$

Proof. We prove the theorem by classifying all possible *patterns* $\phi \in \Phi$ into three categories.

Given any distribution $p \in \mathcal{P}$, let $\mathcal{T}$ be the smallest set with $p(\mathcal{T}) \geq 1-\gamma$. By definition, $|\mathcal{T}| \leq m$.

For any pattern $\phi \in \mathcal{T}$ that satisfies $p(\phi) > \delta$, since the compressor enables (ε,δ)-learning with an algorithm $\mathcal{A}$, we must have $d(f(p), \mathcal{A}(\phi)) \leq \varepsilon$. By the definition of CML, $p_\varphi(\phi) \geq p(\phi) > \delta$, hence, $d(f(p_\varphi), \mathcal{A}(\phi)) \leq \varepsilon$. The triangle inequality combines both and yields $d_f(p_\varphi, p) \leq 2\varepsilon$.

Consider $\phi \in \mathcal{T}$ satisfying $p(\phi) \leq \delta$. By $|\mathcal{T}| \leq m$, the total probability of all such patterns is at most $m\cdot\delta$. In addition, the total probability of patterns $\phi \notin \mathcal{T}$ is at most γ since $p(\mathcal{T}) \geq 1-\gamma$. Therefore, the probability that $d_f(p_\varphi, p) > 2\varepsilon$ is at most $m\cdot\delta+\gamma$, which completes the proof. □

Similar to Acharya et al. (2017a), it suffices to obtain a CML approximation for the competitive guarantees to hold.

Definition 1 (Approximate CML (ACML))**.** *For any $\beta \leq 1$, $z \in \mathcal{Z}$, and compressor φ, a distribution $\tilde{p}_{\varphi(z)} \in \mathcal{P}$ is a* β-approximate CML *if $\tilde{p}_{\varphi(z)}(\varphi(z)) \geq \beta \cdot p_{\varphi(z)}(\varphi(z))$.*

Corollary 1. *For any (m,γ)-typical compressor φ that enables (ε,δ)-learning, distribution $p \in \mathcal{P}$, $Z \sim p$, and a β-approximate CML $\tilde{p}_{\varphi(Z)}$,*

$$\Pr\left(d_f(p, \tilde{p}_{\varphi(Z)}) > 2\varepsilon\right) \leq \gamma + m\cdot\delta/\beta.$$

Building on this framework, we design and analyze CML estimators for various applications. For each, we add a compressor between the samples and the estimator. The main challenges are finding a good compressor φ that works well with ML and an effective algorithm $\mathcal{A}$ for the task.

2.2. CML Estimators

The previous sections presented the general CML framework, and Theorem 1 demonstrated its statistical competitiveness. The remaining sections describe concrete CML estimators for four statistical inference tasks, over both continuous and discrete domains, described below.

Continuous distributions Section 4 applies the CML method to learn structured continuous *i.i.d.* distributions.

We consider a broad distribution class $\mathcal{P}$ where the difference $p-q$ between any two distributions $p, q \in \mathcal{P}$ has essentially at most s sign changes. This class encompasses numerous essential distributions such as log-concave, piecewise polynomials, and Gaussian mixtures.

Theorem 2 shows that for any $p \in \mathcal{P}$, with sample size $\Theta(s\log(s/\varepsilon)/\varepsilon^3)$, the CML estimator achieves

$$\Pr(\|p - p_\varphi\|_1 > \varepsilon) \leq \mathcal{O}(\varepsilon).$$

Note that Yatracos' method, described in Section 3, finds the approximate minimizer to the empirical distribution in $\mathcal{A}_k$-distance, and achieves better sample efficiency. However, in general, no efficient algorithm is known for finding such a minimizer, e.g., Theorem 6.4 in Devroye & Lugosi (2012).

By contrast, the CML approach transforms the learning task to a maximum likelihood problem. This establishes the efficiency of ML methods for learning structured distributions, and enables the use of numerous ML optimization algorithms. Some of these algorithms, like the EM algorithm (Bishop, 2006) are heuristic, while others are rigorous.

Learning distributions with bounded crossings was also considered in Acharya et al. (2017c). However, they assume access to an $\mathcal{A}_k$-projection oracle that finds the approximate $\mathcal{A}_k$-distance minimizers to the empirical distribution.

Discrete distributions Section 5 applies CML to learn discrete *i.i.d.* distributions where as above, every two distributions cross values at most s times.

We present two different CML formulations.

The first formulation in Section 5.1 applies the compressor in Section 4 along with a random mapping that transforms the sample from a discrete distribution to a continuous analog while maintaining the structural properties.

The second formulation constructs CML directly from the discrete sample. The approach essentially performs maximum likelihood on the quantized empirical distribution, with a well-tuned quantization level. The resulting ML objective, presented in Section 5.2, resembles that of the PML estimator in Acharya et al. (2017a).

Theorem 3 shows that for any distribution p with support $\{1,\dots,N\}$ satisfying $N \ge s/\varepsilon$, given a sample from p of size $n = \Theta(s\log(N)/\varepsilon^3)$, with probability at least $1-e^{-s}$,

$$\|p - p_\varphi\|_1 \le \mathcal{O}(\varepsilon).$$

Probability multisets Section 6 is geared towards the original PML method, a special CML whose compressor φ maps samples to *profiles* (Orlitsky et al., 2004; 2011; Das, 2012; Acharya et al., 2012; 2017a; Hao & Orlitsky, 2019a; 2020b; Charikar et al., 2019a;b; Han & Shiragur, 2021).

PML yields a natural estimate for the distribution probability multiset. Given support bound N and desired accuracy ε, it is necessary (in the worst case) to obtain $\Theta(N/(\varepsilon^2\log N))$ observations from the distribution to estimate its probability multiset under the sorted ℓ_1-distance (Valiant & Valiant, 2011c; 2013; Han et al., 2018; Hao & Orlitsky, 2019a).

Over the years, a sequence of works established the optimality of PML for multiset learning in sorted ℓ_1 distance with different accuracy levels (Das, 2012; Hao & Orlitsky, 2019a; Han & Shiragur, 2021). A different multiset learning guarantee stated in terms of the earth-mover's distance was considered in Valiant & Valiant (2016).

Theorem 4 shows that PML also enjoys this learning guarantee, hence is sample-optimal for both sorted ℓ_1-distance and *τ-truncated relative earth-mover distance*. For any level $\tau \in [0,1]$, the latter metric is

$$R_\tau(p,q) := \inf_{\gamma\in\Gamma_{p,q}} \mathop{\mathbb{E}}_{(X,Y)\sim\gamma} \left|\log \frac{\max\{p(X),\tau\}}{\max\{q(Y),\tau\}}\right|,$$

where $\Gamma_{p,q}$ represents all possible couplings of p and q. Section 6.2 shows that with probability at least $1-\mathcal{O}(1/n)$, for any $w \in [1, \log n]$, the compressed estimator p_φ achieves

$$R_{\frac{w}{n\log n}}(p_\varphi,\, p) = \mathcal{O}\left(\frac{1}{\sqrt{w}}\right).$$

Distribution functionals Section 7 explores the application of CML to functional estimation. The compressor maps each sample sequence to the multiset of multiplicities that are small relative to n, leading to a unified algorithm that optimally learns several symmetric functionals.

3. Preliminaries

General notation The rest of the paper considers two types of *univariate domains* $\mathcal{X}$, continuous, $\mathbb{R}$, and discrete, $[N] := \{1,\dots,N\}$. We consider distributions p over $\mathcal{X}$. Given a domain partition $\mathcal{I}$, the *quantized distribution* $p_{\mathcal{I}}$ assigns to each part $I \in \mathcal{I}$ probability $p(I)$.

More generally, $\mathcal{I}$ can be any collection of disjoint measurable subsets of $\mathcal{X} = \mathbb{R}$, and f can be any real measurable function over $\mathbb{R}$. Then, $f_{\mathcal{I}}$ becomes the *quantized function* that assigns any $I \in \mathcal{I}$ a value of $\int_{x\in I} f(x)\,dx$ and the remaining set $\mathcal{X}\setminus\cup_{I\in\mathcal{I}} I$ a probability mass of 0. Slightly abusing notation, we use $f_{\mathcal{I}}$ also to represent the *flattened distribution* that assigns to every x in any $I \in \mathcal{I}$ of positive measure $|I| > 0$, the value

$$f_{\mathcal{I}}(x) := \frac{f_{\mathcal{I}}(I)}{|I|}.$$

Finally, write $a\wedge b$ for $\min\{a,b\}$, and $a\vee b$ for $\max\{a,b\}$.

Function norms We utilize a few norms of functions. The *ℓ_1-norm* that evaluates $\|f\|_1 := \int_{-\infty}^{\infty} |f(x)|\,dx$. Another norm that will come in handy is the *$\mathcal{A}_k$-norm*. Specifically, for any positive integer k, let $\mathcal{I}_k$ denote the collection of all unions of k disjoint intervals. Then, for any measurable function $f:\mathbb{R}\to\mathbb{R}$, the $\mathcal{A}_k$-norm of f is

$$\|f\|_{\mathcal{A}_k} := \sup_{I\in\mathcal{I}_k} \|f_I\|_1.$$

Note that the above notations *naturally extend to* $\mathcal{X} = [N]$ if we replace the integrals with the respective finite sums.

VC inequality A well-known convergence bound associated with the $\mathcal{A}_k$-norm is the VC inequality (Devroye & Lugosi, 2012). Often, the inequality appears as an upper bound on the expected $\mathcal{A}_k$-norm of the difference between the empirical and actual distributions.

We leverage the following variant in Acharya et al. (2017c), making the error probability explicit.

Lemma 2 (VC inequality)**.** *For any $\varepsilon, \delta > 0$, and continuous distribution p, draw a sample $X^n \sim p$ and let $\hat{p}$ denote the empirical distribution. For $n = \Theta((k + \log(1/\delta))/\varepsilon^2)$, with probability at least $1 - \delta$,*

$$\|p - \hat{p}\|_{\mathcal{A}_k} \le \varepsilon.$$

4. Structured Continuous Distributions

A distribution collection $\mathcal{C}$ is an *ε-cover* for $\mathcal{P}$ under ℓ_1-distance if every distribution in $\mathcal{P}$ is within ℓ_1-distance ε from some distribution in $\mathcal{C}$. The *crossing number* of two real functions f and g is the number of times they cross each other, or equivalently, the number of sign changes of $f(x) - g(x)$. The crossing number of a distribution collection is the largest crossing number of any pair of distributions in the collection.

Numerous important distribution families have ε-covers whose crossing number grows moderately as ε decreases. For example, t-piecewise, degree-d polynomials have 0-covers with crossing number $\mathcal{O}(t(d+1))$, and t-mixtures of univariate Gaussians have ε-covers with crossing number $\mathcal{O}(t\log(1/\varepsilon))$ growing only logarithmically in ε.

4.1. Compressor

Let $\mathcal{P}$ have an ε-cover with crossing number s. Draw a sample Y^n from an unknown $p \in \mathcal{P}$ and let $\hat{q}$ be the empirical distribution. Partition the real line into $t := s/\varepsilon$ intervals, $\mathcal{I} := (I_i)_{i=1}^t$, such that $\hat{q}_{\mathcal{I}}$ is uniform, where for simplicity, we assume that s/ε is an integer.

Next, draw an independent sample $X^n \sim p$. Let $\hat{p}$ denote the empirical distribution, and let the compressor φ be its flattened version over $\mathcal{I}$,

$$\varphi(X^n) := \hat{p}_{\mathcal{I}}.$$

Note that $\varphi(X^n)$ depends implicitly on Y^n, and its definition consists of both the Y^n partition and the empirical probabilities of each part according to X^n.

4.2. CML Estimator

Given a sample Y^n and the corresponding interval partition $\mathcal{I}$, the respective CML estimator is

$$p_\varphi := \arg\max_{p\in\mathcal{P}} \Pr_{X^n\sim p}(\varphi(X^n) \mid \mathcal{I}).$$

Note that the probability term corresponds to a multinomial distribution. We can rewrite the CML estimator as

$$p_\varphi = \arg\min_{p\in\mathcal{P}} H(\hat{p}_{\mathcal{I}}, p_{\mathcal{I}}).$$

Hence, CML minimizes the cross entropy between the quantized empirical distribution and the actual one.

Theorem 2. *For any ε and $p \in \mathcal{P}$, draw $(X^n, Y^n) \sim p$. If $n = \Theta(t\log(t/\varepsilon)/\varepsilon^2)$, with probability at least $1 - 2\varepsilon$,*

$$\|p - p_\varphi\|_1 \le 34\varepsilon.$$

Note that by traditional VC theory, the empirical $\mathcal{A}_k$-norm minimizer based on Y^n alone achieves sample complexity of $\Theta(t/\varepsilon)$. However, this method is usually not used for learning structured densities. The theorem on the other hand shows that the practically ubiquitous ML technique can also be applied to learn structured continuous distributions with a rigorous guarantee. It thereby demonstrates a different, continuous, application of Theorem 1 beyond the original application to discrete domains. The use of the samples X^n, while not strictly necessary, facilitates the application of the competitive argument.

The rest of the section is devoted to the theorem's proof.

4.3. Typical Events

First, we leverage standard concentration inequalities to establish several claims, each holds with high probability. The subsequent analysis will assume all claims hold.

By construction, each interval $I_i \in \mathcal{I}$ carries an empirical probability mass of $\hat{q}(I_i) = \varepsilon/s$. The respective probability mass under the actual density, $p(I_i)$, turns out to follow a $\mathrm{Beta}(n/t, n(1-1/t))$ distribution. The following inequality (Hao et al., 2020) shows the typical values of $p(I_i)$.

Lemma 3. *For any $\alpha \in [0,1]$ and $1 \le t \le n$,*

$$\Pr\left(\left|p(I_i) - \frac{1}{t}\right| \ge \frac{\alpha}{\sqrt{t}}\right) \le e^{-\frac{n\alpha^2}{2}} + e^{-\frac{n\alpha^2}{2(1+\alpha\sqrt{t})}}.$$

For any $\gamma \in (0,1)$ and $n \ge 12t\log(2t/\gamma)$, choosing $\alpha = 1/(2\sqrt{t})$ in the lemma yields

$$\Pr\left(\exists i \in [t],\ p(I_i) \notin \left(\frac{1}{2t}, \frac{3}{2t}\right)\right) \le \gamma.$$

Henceforth, we *assume* that $n \ge 12t\log(2t/\gamma)$ and for each i, the actual probability mass of I_i falls in $(1/(2t), 3/(2t))$.

Next, consider the second sample X^n and an arbitrary index $i \in [t]$. By independence, $n\hat{p}(I_i) \sim \mathrm{bin}(n, p(I_i))$, where $p(I_i) > 1/(2t)$ as above. From the binomial Chernoff and union bounds, for all $i \in [k]$, with probability at least $1 - \gamma$,

$$|\hat{p}(I_i) - p(I_i)| \le \sqrt{\frac{3p(I_i)}{n}\log\frac{2t}{\gamma}}.$$

Our lower bounds on n and $p(I_i)$ also imply that for all i,

$$\hat{p}(I_i) \le p(I_i) + \sqrt{\frac{3p(I_i)}{n}\log\frac{2t}{\gamma}} \le \frac{1+\sqrt{2}}{\sqrt{2}} \cdot p(I_i) < \frac{3}{t}.$$

Below we *assume* that $\hat{p}$ satisfies these inequalities.

4.4. Guarantees of CML

Given Y^n's empirical distribution $\hat{q}$, hence also $\mathcal{I}$, there are finitely many flattened X^n-distributions $\hat{p}_{\mathcal{I}}$ over $\mathcal{I}$. Given a probability threshold δ, we consider two disjoint cases according to the probability of a particular distribution $\hat{p}_{\mathcal{I}}$.

Likely $\hat{p}_{\mathcal{I}}$: The probability $p(\hat{p}_{\mathcal{I}}) > \delta$. We show that for such $\hat{p}_{\mathcal{I}}$, with high probability, the CML distribution p_φ is close to p. The following argument assumes that the claims made in Section 4.3 hold.

First, since $\hat{p}(I_i) = \hat{p}_{\mathcal{I}}(I_i) < 3/t$ for all i, then for any distribution p satisfying $\|p - \hat{p}_{\mathcal{I}}\|_{\mathcal{A}_s} > 7\varepsilon$,

$$\|p-\hat{p}\|_{\mathcal{A}_s} \ge \|p-\hat{p}_{\mathcal{I}}\|_{\mathcal{A}_s} - \|\hat{p}_{\mathcal{I}}-\hat{p}\|_{\mathcal{A}_s} > 7\varepsilon - 2s\cdot\frac{3}{t} = \varepsilon.$$

By the VC inequality, this event happens with probability at most $\delta/2$ for a sample size of $n = \Theta((s+\log(2/\delta))/\varepsilon^2)$, contradicting our assumption that $p(\hat{p}_{\mathcal{I}}) > \delta$. Therefore,

$$\|p - \hat{p}_{\mathcal{I}}\|_{\mathcal{A}_s} \le 7\varepsilon.$$

On the other hand, a fundamental attribute of the ML method is that $p_\varphi(\hat{p}_{\mathcal{I}}) \ge p(\hat{p}_{\mathcal{I}}) \ge \delta$. Hence,

$$\|p_\varphi - \hat{p}_{\mathcal{I}}\|_{\mathcal{A}_s} \le 7\varepsilon.$$

For any distribution $q \in \mathcal{P}$, we denote by q' the closest distribution in the ε-cover mentioned above under ℓ_1 distance. From the above and the triangle inequality,

$$\begin{aligned}\|p_\varphi - p\|_1 &\le \|p'_\varphi - p'\|_1 + 2\varepsilon \\ &\le 2\left(\|p-\hat{p}_{\mathcal{I}}\|_{\mathcal{A}_s} + \|p_\varphi - \hat{p}_{\mathcal{I}}\|_{\mathcal{A}_s}\right) + 6\varepsilon \\ &\le 34\varepsilon,\end{aligned}$$

where the second inequality follows as

$$\|q_1 - q_2\|_1 = 2\sup_{A\in\mathbb{R}} |q_1(A) - q_2(A)|.$$

Unlikely $\hat{p}_{\mathcal{I}}$: The probability $p(\hat{p}_{\mathcal{I}}) \le \delta$. Note that the number of possible $\hat{p}_{\mathcal{I}}$'s is at most

$$\binom{t+n-1}{t} \le \left(\frac{e(t+n-1)}{t}\right)^t \le e^{t\log\left(e\left(1+\frac{n}{t}\right)\right)}.$$

Denote by $L_{t,n}$ the exponent of the last term. Setting $\delta = e^{-2L_{t,n}}$ and $n = \Theta(t\log(t/\varepsilon)/\varepsilon^2)$, the total probability of $\hat{p}_{\mathcal{I}}$'s falling into this category is at most

$$e^{L_{t,n}} \cdot e^{-2L_{t,n}} = e^{-L_{t,n}} \le \varepsilon^{-t},$$

where the last step follows by choosing a sufficiently large absolute constant in the above asymptotic expression for n.

Summary Choosing $\gamma = \varepsilon/2$ in Section 4.3, if $n = \Theta(t\log(t/\varepsilon)/\varepsilon^2)$, the union bound implies that with probability at least $1-2\varepsilon$,

$$\|p_\varphi - p\|_1 \le 34\varepsilon.$$

5. Structured Discrete Distributions

Moving from $\mathbb{R}$ to the discrete domain $[N]$, we apply CML to learn structured discrete distributions. Again, we assume that $\mathcal{P}$ has an ε-cover $\mathcal{C}$ with a crossing number at most s.

5.1. Reduction from Continuous CML

First, we present a concrete *random mapping* from discrete to continuous domains that enables the use of the continuous CML estimator.

For any distribution p over $[k]$, we define by $\tilde{p}$ its *flattened version*, a continuous distribution that assigns

$$\tilde{p}(x) := p(\lceil x\rceil),\ \forall x \in (0,k],$$

and $\tilde{p}(x) := 0$ for $x \notin (0,k]$. This notation yields a bijective mapping that maintains the number of sign changes of the difference between any two distributions.

Note that we only have sample access to the underlying distribution p. To apply the CML estimator in Section 4, we need to simulate a sample from $\tilde{p}$ from $X^n \sim p$. A random mapping $\mathcal{S}$ for this purpose counts the number of times each symbol $i \in [N]$ observed in X^n, say n_i, and respectively draws n_i independent sample points from the uniform distribution over $(i-1, i]$.

Now, it is straightforward to apply the continuous CML. Specifically, draw samples $(X^n, Y^n) \sim p$, define the compressor φ based on $\mathcal{S}(Y^n)$, and let the CML estimate be

$$p_\varphi := \arg\max_{p\in\mathcal{P}} \Pr_{X^n\sim p}(\varphi(\mathcal{S}(X^n)) \mid \mathcal{I}).$$

5.2. Discrete CML

Next, we present an alternative CML method that applies directly to discrete samples.

Compressor Draw a size-n sample $X^n \sim p$, and denote by $\hat{p}$ the empirical distribution. Then, sort elements in X^n in a non-descending order, say, $X_{(1)}, \ldots, X_{(n)}$, such that

$$X_{(i)} \le X_{(j)},\ \forall i < j.$$

Without loss of generality, assume that $t := s/\varepsilon$ is an integer and n is divisible by t. Sequentially partition the sequence into t sub-sequences, such that the i-th sub-sequence contains observations with indices from $(i-1)n/t+1$ to in/t.

The compressor φ maps every sample X^n to the boundary symbols that of these sub-sequences. Specifically,

$$\varphi(X^n) := \left(X_{(in/t)}\right)_{i=1}^{t},$$

which is *essentially* $\hat{p}_{\mathcal{I}}$ with $\mathcal{I} := (I_i)_{i=1}^{t}$ and

$$I_i := [X_{((i-1)n/t)} : X_{(in/t)}],\ \forall i \in [t].$$

CML estimator Consequently, the CML estimator for the above compressor takes the form

$$p_\varphi := \arg\max_{p\in\mathcal{P}} \Pr_{X^n\sim p}(\varphi(X^n) \mid \mathcal{I}).$$

Equivalently, the CML estimator can be written as

$$p_\varphi := \arg\max_{p\in\mathcal{P}} \sum_{y^n:\varphi(y^n)=\varphi(X^n)} \prod_{i=1}^{n} p(y_i).$$

The estimator often performs well, as described below.

Theorem 3. *For $p \in \mathcal{P}$ and $X^n \sim p$, if $n = \Theta((t \wedge N)\log(t \vee N)/\varepsilon^2)$, with probability at least $1 - e^{-t\wedge N}$,*

$$\|p - p_\varphi\|_1 = \mathcal{O}(\varepsilon).$$

5.3. Hypothesis Selection Algorithm

First, we utilize the assumptions and the standard VC inequality to show the *existence* of a hypothesis selection algorithm that, with high probability, finds an accurate estimate of the underlying distribution.

By the VC inequality, for any $\varepsilon, \delta > 0$, and $n = \Theta((s + \log(2/\delta))/\varepsilon^2)$, with probability at least $1 - \delta/2$,

$$\|\hat{p} - p\|_{\mathcal{A}_s} \le \varepsilon.$$

Denote by $\mathcal{I}_s$ the set of unions of at most s disjoint intervals in $\mathcal{I}$. By our construction, for any set $I \in \mathcal{I}_s$,

$$\|\hat{p}_I\|_1 \le \frac{1}{t} \cdot s = \varepsilon.$$

For any $q \in \mathcal{P}_{s,\varepsilon}$, let q' be a distribution in the ε-cover $\mathcal{C}$ with minimal ℓ_1-distance to q. For any distribution pair $q_1, q_2 \in \mathcal{P}_{s,\varepsilon}$, define the following *variant* of the Scheffé set (Devroye & Lugosi, 2012) as

$$S'_{12} := \{x \in \mathbb{R} : q'_1(x) > q'_2(x)\},$$

and the original Scheffé set as S_{12}.

Given the previous inequality and distribution $\hat{p}_{\mathcal{I}}$, we can approximate $\hat{p}(q_1 \ge q_2) := \hat{p}(S_{12})$ to a 3ε additive error by summing up $\hat{p}_{\mathcal{I}}(I_i)$ over indices i satisfying $I_i \in S'_{12}$. Note that we assumed that $q'_1 - q'_2$ has at most s sign changes.

Next, we show that if the VC inequality holds, there exists a selection algorithm that uses $\hat{p}_{\mathcal{I}}$ to find a density $p^\star \in \mathcal{C}$ satisfying $\|p^\star - p\| \le \mathcal{O}(\varepsilon)$. In addition, this algorithm is a variant of that in Theorem 6.4 of Devroye & Lugosi (2012).

Let $\mathcal{S}'$ denote the collection of all the modified Scheffé sets induced by distributions in $\mathcal{C}$. Let $p^\star$ be the distribution q in $\mathcal{C}$ minimizing $\sup_{S\in\mathcal{S}} |\hat{p}_{\mathcal{I}}(S) - q(S)|$ up to an additive 4ε error, where we enlarged the error bound by ε to guarantee the existence of such a distribution. Then, the triangle inequality yields

$$\|p - p^\star\|_1 \le \|p - p'\|_1 + \|p' - p^\star\|_1.$$

For any pair (q_1, q_2) of distributions, define

$$D_{\mathcal{S}}(q_1,\, q_2) := \sup_{S\in\mathcal{S}} |q_1(S) - q_2(S)|.$$

Consider the last term in the above inequality. By a recursive application of the triangle inequality and properties embedded in the prior constructions,

$$\begin{aligned}
\|p' - p^\star\|_1 &= 2D_{\mathcal{S}}(p', p^\star) \\
&\le 2D_{\mathcal{S}}(p', \hat{p}) + 2D_{\mathcal{S}}(\hat{p}, p^\star) \\
&\le 2D_{\mathcal{S}}(p', \hat{p}_{\mathcal{I}}) + 2D_{\mathcal{S}}(\hat{p}_{\mathcal{I}}, p^\star) + \mathcal{O}(\varepsilon) \\
&\le 4D_{\mathcal{S}}(p', \hat{p}_{\mathcal{I}}) + \mathcal{O}(\varepsilon) \\
&\le 4D_{\mathcal{S}}(p', \hat{p}) + \mathcal{O}(\varepsilon) \\
&\le 4D_{\mathcal{S}}(p', p) + 4D_{\mathcal{S}}(p, \hat{p}) + \mathcal{O}(\varepsilon) \\
&\le 2\|p' - p\|_1 + 4D_{\mathcal{S}}(p, \hat{p}) + \mathcal{O}(\varepsilon).
\end{aligned}$$

Recall that $\mathcal{C}$ is an ε-cover of $\mathcal{P}_{s,\varepsilon}$, we have $\|p' - p\|_1 \le \varepsilon$. In addition, the collection $\mathcal{S}'$ of modified Scheffé sets has a VC dimension of at most $2s + 2$. Therefore by the VC inequality, for a sample size of $n = \Theta((s + \log(2/\delta))/\varepsilon^2)$, with probability at least $1 - \delta/2$,

$$D_{\mathcal{S}}(p, \hat{p}) \le \varepsilon.$$

Consolidating these results shows that with probability at least $1 - \delta$, the selected hypothesis $p^\star$ achieves

$$\|p - p^\star\|_1 = \mathcal{O}(\varepsilon).$$

Note, however, that the above selection algorithm does not provide a method for finding $p^\star$.

5.4. Guarantees of CML

There are only finitely many possible $\hat{p}_{\mathcal{I}}$, or equivalently, $\varphi(X^n)$. Below, we consider two disjoint cases according to the probability of observing a particular *pattern* of $\varphi(X^n)$ under the actual distribution.

Likely pattern: Pattern ϕ whose probability $p(\phi) := \Pr_{X^n\sim p}(\varphi(X^n) = \phi) > \delta$. By the argument in Section 5.3,

there exists a selection algorithm that maps this pattern to a distribution $p^\star$ whose error probability in achieving

$$\|p - p^\star\|_1 = \mathcal{O}(\varepsilon)$$

is at most δ. Since $p(\phi) > \delta$ by assumption, the selected hypothesis must satisfy the error bound above.

It is straightforward to combine these error bounds through Theorem 1 and the triangle inequality,

$$|p - p_\varphi| \le |p_\varphi - p^\star| + |p - p^\star| = \mathcal{O}(\varepsilon).$$

Unlikely pattern: Pattern ϕ whose probability $p(\phi) \le \delta$. Since ϕ can only be a sorted integer sequence with values in $[N]$, the number of possible patterns equals the number of ways to pick t elements from $[N]$ with repetition, which is

$$\binom{t+N-1}{t} \le \left(\frac{e(t+N)}{t}\right)^t \bigwedge \left(\frac{e(N+t)}{N}\right)^k \le \exp\left((t \wedge N)\log\left(e\left(1 + \frac{N}{t} \vee \frac{t}{N}\right)\right)\right).$$

Denote by $L_{t,n}$ the exponent of the last term. Setting $\delta = e^{-2L_{t,n}}$ and $n = \Theta(L_{t,n}/\varepsilon^2)$, the total probability of patterns falling into this category is at most

$$e^{L_{t,n}} \cdot e^{-2L_{t,n}} = e^{-L_{t,n}} \le e^{-t \wedge N},$$

where the last step follows by choosing a sufficiently large absolute constant in the expression for n.

6. Probability Multisets

In this section, we address probability multiset estimation under permutation invariance.

6.1. Compressor and PML

For any discrete distribution p and sample $X^n \sim p$, let compressor φ map the sequence to its *profile* $\varphi(X^n)$, defined as the *multiplicity multiset* of symbols appearing in X^n. The respective CML estimator, introduced in Orlitsky et al. (2004) as the *PML estimator*. Specifically, PML computes

$$p_\varphi := \arg\max_{p \in \mathcal{P}} \sum_{y^n : \varphi(y^n) = \varphi(X^n)} \prod_{i=1}^n p(y_i).$$

For any $\tau \in [0,1]$, define the *τ-truncated relative earth-mover distance* (Valiant & Valiant, 2015) between two discrete distributions p and q as

$$R_\tau(p,q) := \inf_{\gamma \in \Gamma_{p,q}} \mathop{\mathbb{E}}_{(X,Y)\sim\gamma} \left|\log \frac{p(X) \vee \tau}{q(Y) \vee \tau}\right|,$$

where $\Gamma_{p,q}$ represents all the possible couplings of these two distributions. In the following, we show that the PML distribution estimator satisfies

Theorem 4. *For any discrete distribution p, draw a sample $X^n \sim p$. With probability at least $1 - 4\exp(-\Omega(n^{\frac{1}{3}}))$, for any $w \in [1, \log n]$,*

$$R_{\frac{w}{n\log n}}(p_\varphi,\ p) = \mathcal{O}\left(\frac{1}{\sqrt{w}}\right).$$

For any $\tau \in [0,1]$, define the *τ-truncated sorted ℓ_1-distance* between two distributions $p, q \in \Delta_\mathcal{X}$ as

$$\tilde{\ell}_\tau(p,q) := \min_{p' \in \Delta_\mathcal{X} : \{p'\} = \{p\}} \sum_x |p'(x) \vee \tau - q(x) \vee \tau|,$$

where $\{p\}$ denotes the probability multiset of p. By Fact 1 in Valiant & Valiant (2015), for any distributions $p, q \in \Delta_\mathcal{X}$, $\tilde{\ell}_\tau(p,q) \le 2R_\tau(p,q)$, implying

Corollary 2. *Under the same conditions as Theorem 4,*

$$\tilde{\ell}_{\frac{w}{n\log n}}(p_\varphi,\ p) = \mathcal{O}\left(\frac{1}{\sqrt{w}}\right).$$

6.2. Proof of Theorem 4

This section provides a sketch for the proof of Theorem 4. *We relegate most technical details to Appendix A.* Part of the proof is adapted from Theorem 2 in Valiant & Valiant (2015). The original reasoning is not sufficient for our purpose as the error probability derived is too large to invoke the competitiveness of PML. For this reason, we modify the linear program used in the paper, carefully separate the analysis of the estimators for large and small probabilities, and provide a refined analysis with tighter probability bounds by reducing union-bound arguments.

First, we define histograms and the relative earth-moving cost and give operational meaning to R_τ. For a distribution p, the *histogram* of a multiset $\mathcal{A} \subseteq \{p\}$ is a mapping, denoted by $h_\mathcal{A} : (0,1] \to \mathbb{Z}_{\ge 0}$, that maps each number $y \in (0,1]$ to the number of times it appears in $\mathcal{A}$. Note that every y corresponds to a probability mass of $y \cdot h_\mathcal{A}(y)$. More generally, we also allow *generalized histograms* h with non-integral values $h(y) \in \mathbb{R}_{\ge 0}$.

For any $y_1, y_2 \in (0,1]$, generalized histogram h, and non-negative $m < y_1 \cdot h(y_1)$, we can *move* a probability mass from location y_1 to y_2 by reassigning $h(y_1) - m/y_1$ to y_1, and $h(y_2) + m/y_2$ to y_2. Given $\tau \in [0,1]$, we define the *cost* associated with this operation as

$$c_{\tau,m}(y_1, y_2) := m \cdot \left|\log \frac{y_1 \vee \tau}{y_2 \vee \tau}\right|,$$

and term it as *τ-truncated earth-moving cost*. The cost of multiple operations is additive. Note that $R_\tau(p,q)$ is the minimal total τ-truncated earth-moving cost associated with any operation schemes of moving $h_{\{p\}}$ to yield $h_{\{q\}}$.

For simplicity, we suppress X^n in $\varphi_i(X^n)$ and $\mu_s(X^n)$, representing respectively the *number of symbols appearing i times* and the *number of times symbol s appears*.

For any absolute constants B and C satisfying $0.1 > B > C > \frac{B}{2} > 0$, define $x_n := \frac{n^B+n^C}{n}$ and $S := \{\frac{1}{n^2}, \frac{2}{n^2}, \ldots, x_n\}$. Consider the following linear program.

For each $x \in S$, define the associated variable v_x

$$\text{Minimize} \sum_{i=1}^{n^B} \left| \varphi_i - \sum_{x\in S} \text{bin}(n, x, i) \cdot v_x \right|$$

$$\text{s.t.} \sum_{x\in S} x \cdot v_x = \sum_{i\le n^B+2n^C} \frac{i}{n} \cdot \varphi_i$$

and $\forall x \in S, v_x \ge 0$

Figure 1. Linear program (LP)

Existence of a good feasible point

Let p be the underlying distribution and h be its histogram. First, we show that with *high* probability, the linear program LP has a feasible point (v_x) that is *good* in the following sense: 1) the corresponding objective value is relatively *small*; 2) for $\tau \ge n^{-3/2}$, the generalized histogram $h_0 : x \to v_x$ is *close* to $h_n : y \to h(y) \cdot \mathbb{1}_{y\le x_n}$, admitting a low τ-truncated earth-mover cost.

In the appendix, we leverage the Chernoff bound and union bound to show that with probability at least $1 - \exp(-n^{\frac{1}{3}+\kappa})$, the objective value of the feasible point (v_x) is at most $n^B \cdot \mathcal{O}(n^{\frac{2}{3}+\kappa} + 1) = \mathcal{O}(n^{B+\frac{2}{3}+\kappa})$.

For any $\tau \ge n^{-3/2}$, the minimal τ-truncated earth-moving cost of moving the generalized histogram h_0 corresponding to (v_x), and the histogram $h_n : y \to h(y) \cdot \mathbb{1}_{y\le x_n}$, so that they differ from each other only at $x = x_n$, is at most

$$\log\left(\frac{n^{-3/2}+n^{-2}}{n^{-3/2}}\right) + \mathcal{O}\left(\frac{\log n}{n^{\frac{1}{3}-\kappa}}\right) = \mathcal{O}(n^{-\frac{1}{3}+2\kappa}).$$

All solutions are good solutions

Let (v_x) be the solution described above. The appendix then proceeds to show that for any solution (v'_x) to LP whose objective value is $\mathcal{O}(n^{B+\frac{2}{3}+\kappa})$, the generalized histogram h_1 corresponding to (v'_x) is close to h_0.

Specifically, the proof establishes a $\mathcal{O}(1/\sqrt{\log n})$ discrepancy bound whenever $B = 1.5C = 10\kappa = 0.01$.

Consolidate the previous results. For $w \in [1, \log n]$ and $\tau = w/(n \log n)$, with probability at least $1 - \exp(-n^{\frac{1}{3}+\kappa})$, the solution to LP will yield a generalized histogram h_1, such that the minimal τ-truncated earth-moving cost of moving h_1 and h_n so that they differ only at x_n, is $\mathcal{O}(1/\sqrt{w})$.

Competitiveness of PML

For the PML distribution associated with a sample satisfying our prior assumptions, denote by h_n^{PML} the histogram corresponding to its entries that are at most x_n.

By a recent result in (Han & Shiragur, 2021), for any $w \in [1, \log n]$ and $\tau = w/(n \log n)$, the minimal τ-truncated cost of moving h_n^{PML} and h_n so that they differ only at x_n, is $\mathcal{O}(1/\sqrt{w})$, with probability at least $1 - 2\exp(-\Omega(n^{\frac{1}{3}}))$.

Properties of the empirical histogram

Denote by h^{EMP} the empirical histogram. By the Chernoff bound, for any symbol s, the probability that

$$|n \cdot p(s) - \mu_s| \ge \mu_s^{3/4} \text{ and } \mu_s > n^B + 2n^C$$

is at most $2np(s)\exp(-\Omega(n^{2C-B}))$, and similarly,

$$n \cdot p(s) \ge n^B + 4n^C \text{ and } \mu_s \le n^B + 2n^C$$

will happen with probability at most $2\exp(-\Omega(n^{2C-B}))$. Hence, we assume that $|n\cdot p(s) - \mu_s| < \mu_s^{3/4}$ for all symbols s appearing more than $n^B + 2n^C$ times, and that any symbol s with probability $p(s) \ge (n^B + 4n^C)/n$ appears more than $n^B + 2n^C$ times. By the union bound, we will be correct with probability at least $1 - 4n\exp(-\Omega(n^{2C-B}))$.

Let $y_n := (n^B + 4n^C)/n$ for notational convenience. If for each symbol s satisfying $\mu_s \ge n^B + 2n^C$, we move a μ_s/n probability mass of h^{EMP} from μ_s/n to $p(s)$, then at all locations $y \ge y_n$, the total discrepancy between the resulting generalized histogram and the actual one is at most

$$\sum_{j>n^B+2n^C} \varphi_j \frac{j^{\frac{3}{4}}}{n} = \frac{1}{n} \sum_{j>n^B+2n^C} \varphi_j^{\frac{1}{4}} (\varphi_j j)^{\frac{3}{4}} \le n^{-\frac{B}{4}},$$

where the last step follows by Hölder's inequality. Moreover, the associated total earth-moving cost is at most

$$\sum_{j>n^B+2n^C} \varphi_j \frac{j}{n} \log\left|\frac{j}{j \pm j^{\frac{3}{4}}}\right| \le \sum_{j>n^B+2n^C} \varphi_j \frac{j^{\frac{3}{4}}}{n} \le n^{-\frac{B}{4}}.$$

The y_n-truncated earth-mover distance between h^{EMP} and h is thus at most $2n^{-\frac{B}{4}} \log n = \mathcal{O}(n^{-\frac{B}{5}})$, which, together with the above error probability bound, upper bounds the expected value of $R_{y_n}(h^{\text{EMP}}, h)$ by $O(n^{-\frac{B}{5}}) + 4n\exp(-\Omega(n^{2C-B}))\log n = O(n^{-\frac{B}{5}})$.

Moreover, changing any element in the sample sequence changes the value of $R_{y_n}(h^{\text{EMP}}, h)$ by at most $(\log n)/n$. Hence, by McDiarmid's inequality, with probability at least $1 - 2\exp(-2\sqrt{n})$, the value of $R_{y_n}(h^{\text{EMP}}, h)$ is less than $O(n^{-\frac{B}{5}}) + n^{-\frac{1}{4}} \log n = O(n^{-\frac{B}{5}})$.

COMPETITIVENESS OF PML

Consider the PML histogram h^{PML} and its y_n-truncated earth-mover distance to h. Since there are at most $\exp(3\sqrt{n})$ different profiles (Hardy & Ramanujan, 1918), with probability at least $1-2\exp(-\Omega(\sqrt{n}))$,

$$R_{y_n}(h^{\text{PML}}, h) \le \mathcal{O}(n^{-\frac{B}{5}}).$$

PERFORMANCE OF PML

We consolidate the previous results. For any $w \in [1, \log n]$ and $\tau = w/(n\log n)$, we design an earth-moving scheme that moves h^{PML} to h.

First, with probability at least $1-2\exp(-\Omega(n^{\frac{1}{3}}))$, we can move the probability mass of h^{PML}_n so that it differs from h_n only at $x_n = (n^B + n^C)/n$ while incurring a τ-truncated earth-mover cost of at most $\mathcal{O}(1/\sqrt{w})$.

Second, by the empirical-histogram argument and $x_n < y_n$, with probability at least $1-2\exp(-\Omega(n^{\frac{1}{3}}))$, we can further move the probability mass of h^{PML} at locations $y \ge x_n$ to coincide with h above y_n while incurring a loss of $\mathcal{O}(n^{-\frac{B}{5}})$.

After the previous two steps, the modified PML histogram differs from the actual histogram h only at locations $y \in I_n = [x_n, y_n]$. Note that the cost of moving a unit mass within I_n is at most $3n^{C-B}$, implying that with probability at least $1-2\exp(-\Omega(n^{\frac{1}{3}}))$,

$$R_\tau(h^{\text{PML}}, h) \le \mathcal{O}\left(\frac{1}{\sqrt{w}} + n^{-\frac{B}{5}}\right) + 3n^{C-B} = \mathcal{O}\left(\frac{1}{\sqrt{w}}\right).$$

7. Symmetric Functionals

In this section, we apply the idea of CML to estimate multisets of low probabilities and leverage the respective plug-in estimator to approximate several distribution functionals.

The study of functional estimation dates back more than half a century (Carlton, 1969; Good, 1953; Good & Toulmin, 1956) and has steadily grown over the years. While the empirical distribution plug-in estimator performs well in the large-sample regime, modern data science applications often study high-dimensional data, for which more sophisticated methods lead to estimators that possess better guarantees (Jiao et al., 2015; Orlitsky et al., 2016; Valiant & Valiant, 2011a;b; Wu & Yang, 2016; Hao et al., 2018; Wu & Yang, 2019; Hao & Orlitsky, 2019b; 2020a;b).

Recently, a line of research works studied the use of PML in functional estimation (Orlitsky et al., 2004; 2011; Das, 2012; Acharya et al., 2012; 2017a; Hao & Orlitsky, 2019a; 2020b; Charikar et al., 2019a;b; Han & Shiragur, 2021).

For functionals addressed in this section, the plain PML is known to be sample-optimal only for additive errors $\varepsilon \ge 1/n^{1/3}$. Concurrently with our result, Charikar et al. (2019c) proposed a different PML-type functional estimation method with optimal sample complexity down to $\varepsilon \gg 1/\sqrt{n}$ accuracy. We note that some specialized estimators also achieve better accuracy of order $1/\sqrt{n}$.

Recently, an efficient algorithm (Anari et al., 2020) was proposed to compute the PML-type estimators in Charikar et al. (2019c). The techniques also apply to the algorithms presented in this section.

7.1. Compressor

For any distribution p over $\mathcal{X}$, sample $X^n \sim p$, and multiplicity i, recall that $\varphi_i(X^n)$ denotes the number of symbols appearing exactly i times.

Our objective is to extract information about low probabilities of p, regardless of symbol permutations. It is natural to consider, for an integer $t \le n$, the compressor

$$\varphi(X^n) := (\varphi_i(X^n))_{i=1}^t,$$

where t determines the inference horizon.

7.2. CML Estimator

Similar to previous sections, for fixed t and sample $X^n \sim p$, the CML estimator for the above compressor takes the form

$$p_\varphi := \arg\max_{p\in\mathcal{P}} \Pr_{X^n\sim p}(\varphi(X^n)).$$

Equivalently, the CML estimator can be written as

$$p_\varphi := \arg\max_{p\in\mathcal{P}} \sum_{y^n:\varphi(y^n)=\varphi(X^n)} \prod_{i=1}^n p(y_i).$$

For space considerations, we relegate the rest of this section and the paper's appendix to the supplementary material.

Conclusion

The paper proposes a simple, novel, and unified compressed maximum likelihood (CML) approach for several fundamental tasks. The new technique bridges algorithms and results in several research directions over both discrete and continuous domains.

Acknowledgements

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

We also thank an anonymous reviewer for an excellent summary of our contribution, incorporated into Section 2.1, and another anonymous reviewer for suggesting the approximate CML estimator (Definition 1).

References

Acharya, J., Das, H., Jafarpour, A., Orlitsky, A., and Pan, S. Estimating multiple concurrent processes. In *Proceedings 2012 IEEE International Symposium on Information Theory (ISIT)*, pp. 1628–1632. IEEE, 2012.

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

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

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

Anari, N., Charikar, M., Shiragur, K., and Sidford, A. Instance based approximations to profile maximum likelihood. *Advances in neural information processing systems*, 2020.

Barbour, A. D. and Hall, P. On the rate of Poisson convergence. *Mathematical Proceedings of the Cambridge Philosophical Society*, 95(3):473–480, 1984.

Bishop, C. M. *Pattern recognition and machine learning*. springer, 2006.

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

Carlton, A. G. On the bias of information estimates. *Psychological Bulletin*, 71(2):108, 1969.

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

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

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

Charikar, M., Shiragur, K., and Sidford, A. The Bethe approximation for structured matrices: an improved approximation for the profile maximum likelihood. In *NeurIPS 2019 Workshop on Information Theory and Machine Learning*, 2019a.

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

Charikar, M., Shiragur, K., and Sidford, A. A general framework for symmetric property estimation. *Advances in neural information processing systems*, 2019c.

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

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

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

Das, H. *Competitive tests and estimators for properties of distributions*. PhD thesis, UC San Diego, 2012.

Devroye, L. and Lugosi, G. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

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

Friedman, J., Hastie, T., Tibshirani, R., et al. *The elements of statistical learning*. Number 10 in 1. Springer series in statistics New York, 2001.

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

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

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

Haas, P. J., Naughton, J. F., Seshadri, S., and Stokes, L. Sampling-based estimation of the number of distinct values of an attribute. *VLDB*, 95:311–322, 1995.

Han, Y. and Shiragur, K. On the competitive analysis and high accuracy optimality of profile maximum likelihood. In *Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1317–1336. SIAM, 2021.

Han, Y., Jiao, J., and Weissman, T. Local moment matching: A unified methodology for symmetric functional estimation and distribution estimation under wasserstein distance. In *Conference On Learning Theory*, pp. 3189–3221, 2018.

Hao, Y. and Orlitsky, A. The broad optimality of profile maximum likelihood. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10989–11001, 2019a.

Hao, Y. and Orlitsky, A. Unified sample-optimal property estimation in near-linear time. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 11106–11116, 2019b.

Hao, Y. and Orlitsky, A. Data amplification: Instance-optimal property estimation. In *International Conference on Machine Learning (ICML)*, pp. 4049–4059. PMLR, 2020a.

Hao, Y. and Orlitsky, A. Profile entropy: A fundamental measure for the learnability and compressibility of discrete distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b.

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

Hao, Y., Jain, A., Orlitsky, A., and Ravindrakumar, V. SURF: A simple, universal, robust, fast distribution learning algorithm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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

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

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

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

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

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

Orlitsky, A., Santhanam, N. P., Viswanathan, K., and Zhang, J. On modeling profiles instead of values. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 426–435. AUAI Press, 2004.

Orlitsky, A., Santhanam, N. P., Viswanathan, K., and Zhang, J. On estimating the probability multiset. *Online Draft*, 2011. URL http://alon.ucsd.edu/papers/pml1.pdf.

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

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

Shannon, C. E. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

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

Valiant, G. and Valiant, P. Estimating the unseen: An $n/log(n)$-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings 43rd Annual ACM Symposium on Theory of Computing (STOC)*, pp. 685–694. ACM, 2011a.

Valiant, G. and Valiant, P. The power of linear estimators. In *Proceedings 52nd IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 403–412. IEEE, 2011b.

Valiant, G. and Valiant, P. Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new CLTs. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 685–694, 2011c.

Valiant, G. and Valiant, P. Instance optimal learning. *arXiv preprint, arXiv: 1504.05321*, 2015.

Valiant, G. and Valiant, P. Instance optimal learning of discrete distributions. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 142–155, 2016.

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

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

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

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

7. Symmetric Functionals (Continued)

The next corollary of Theorem 1 states that the CML plug-in estimator is competitive to other functional estimators.

Corollary 3. *Let f be a symmetric distribution functional. If for a sample size of n, there exists an estimator $\hat{\varphi} : \Phi \to \mathbb{R}$ such that for any $p \in \mathcal{P}$ and $X^n \sim p$,*

$$\Pr(|f(p) - \hat{\varphi}(\varphi(X^n))| > \varepsilon) < \delta,$$

then

$$\Pr(|f(p) - f(p_\varphi)| > 2\varepsilon) < \delta \cdot en^t.$$

Proof. For any $x^n \in \mathcal{X}^n$, the compressed sequence $\varphi(x^n)$ is a vector of integers from 0 to n. Hence, the compressor is $((n+1)^t, 0)$-typical. In addition, we have $(n+1)^t/n^t \le (1+1/n)^n < e$ as we assumed $t \le n$. The corollary then directly follows from Theorem 1. Note that the term en^t in the upper bound is sub-optimal for large t values but is sufficient for our purposes. □

7.3. Shannon Entropy

We begin with the fundamental entropy estimation problem. The *Shannon entropy* of a distribution p over $\mathcal{X} = [N]$ is

$$H(p) := \sum h(p(x)) := \sum_x -p(x) \log p(x).$$

Shannon entropy is the primary measure of randomness and information (Cover & Thomas, 2012; Shannon, 1948) with wide machine learning (Bresler, 2015; Chow & Liu, 1968; Quinn et al., 2013) and neuroscience applications (Gerstner & Kistler, 2002; Mainen & Sejnowski, 1995).

Given sample access to an unknown p, we draw a pair of independent samples $(X^n, Y^n) \sim p$. Denote by $\hat{p}$ and $\hat{q}$ the respective empirical distributions, and μ_x and ν_x the respective empirical counts of an arbitrary symbol $x \in [N]$.

We separate the empirical counts into a few categories by four thresholds of order-$\log n$,

$$\tau_i := c_i \cdot \log n, \forall i \in [4],$$

where c_i's are positive absolute constants to be properly chosen later. Since p is unknown, we perform a *soft truncation* and partition our target $H(p)$ into a *low-probability part*

$$T(p) := \sum_x h(p(x)) \cdot \Pr(\mu_x \le \tau_1)$$

and a *high-probability part*

$$L(p) := H(p) - T(p).$$

We can estimate $L(p)$ by a variant of the Miller-Mallow estimator (Carlton, 1969):

$$\hat{L} := \sum_x \left(h(\hat{p}(x)) + \frac{1}{2n} \right) \cdot (1 - \mathbb{1}_{\mu_x \le \tau_2,\, \nu_x \le \tau_1}).$$

For $c_2 \gg c_1 \gg 1$, the estimator's bias satisfies

$$\left| \mathbb{E}[\hat{L}] - L(p) \right| = \mathcal{O}\left(\frac{N}{n \log n} \right).$$

The following result shows that a simple combination of CML and the Miller-Mallow variant is sample-optimal for nearly the entire applicable accuracy range.

Theorem 5. *For any desired estimation accuracy ε satisfying $\varepsilon = \Omega(1/n^{0.49})$, the entropy estimator*

$$\hat{H} := T(p_\varphi) + \hat{L}$$

achieves the optimal $\Theta(N/(\varepsilon \log N))$ sample complexity.

As a remark, 0.49 in the exponent can be replaced by an absolute constant smaller than $1/2$, but a larger value implies a larger asymptotic constant in the sample complexity. Refer to Appendix B.2 for proof of this theorem.

7.4. Support Size

Support size functional is an essential distribution attribute, arising in the research of vocabulary size (Efron & Thisted, 1976; McNeil, 1973; Thisted & Efron, 1987), population (Good, 1953; Mao & Lindsay, 2007), and database systems (Haas et al., 1995).

Following the previous discussion, we consider estimating the *normalized support size*

$$S(p) := \sum_x s(p(x)) := \sum_x \frac{\mathbb{1}_{p(x)>0}}{N}.$$

The problem is *ill-defined* without additional assumptions since symbols with an arbitrarily small probability mass can

modify the quantity by some nontrivial constant. A common assumption for support estimation requires the minimal non-zero probability of p to be at least $1/N$. In the following derivations, we will assume that this bound holds.

Analogous to Section 7.3, we draw a sample X^n from p and separate the empirical counts by a threshold of order-$\log n$,

$$\tau := c \log n,$$

for some absolute constant c to be properly chosen later.

For any $x \in [N]$, let μ_x be the number of times x appearing in the sample. We partition $S(p)$ into a *low-probability part*

$$T(p) := \sum_x s(p(x)) \cdot \Pr(\mu_x \le \tau)$$

and a *high-probability part*

$$L(p) := S(p) - T(p).$$

A natural choice for $L(p)$ is the *unbiased* estimator

$$\hat{L} := \sum_x s(\mu_x - \tau).$$

By McDiarmid's inequality, with probability at least $1-2/n$,

$$|\hat{L} - L(p)| \le \frac{\log^2 n}{\sqrt{n}}.$$

We construct a support estimator with the following guarantee based on CML and the unbiased estimator.

Theorem 6. *For any desired estimation accuracy ε satisfying $\varepsilon = \Omega(1/n^{0.49})$, the support estimator*

$$\hat{S} := T(p_\varphi) + \hat{L}$$

achieves the optimal $\Theta\left(\frac{N}{\log N} \log^2 \frac{1}{\varepsilon}\right)$ sample complexity.

Similar to Section 7.3, the 0.49 exponent can be replaced by an absolute constant smaller than $1/2$, but a larger value implies a larger asymptotic constant in the complexity. Refer to Appendix B.4 for proof of this theorem.

7.5. Support Coverage

For a given parameter m, the *normalized support coverage* of a distribution p is

$$C(p) := \sum_x c(p(x)) := \sum_x \frac{1-(1-p(x))^m}{m},$$

the (normalized) expected number of distinct symbols in a size-m sample. Support coverage estimation closely relates to the well-known *unseen species problem* (Acharya et al., 2017b), arising in biological (Chao, 1984; Kroes et al., 1999) and ecological studies (Chao, 1984; Chao & Lee, 1992; Chao & Chiu, 2014; Colwell et al., 2012).

Adapting the notation in Section 7.3, we can partition $C(p)$ into a *low-probability part*

$$T(p) := \sum_x c(p(x)) \cdot \Pr(\mu_x \le \tau)$$

and a *high-probability part*

$$L(p) := C(p) - T(p),$$

where we draw an independent sample pair $(X^n, Y^n) \sim p$, and denote by μ_x and ν_x the respective empirical counts of an arbitrary symbol $x \in [N]$.

We assume that $N \gg m$ without loss of generality since one can always append symbols of zero probability to a distribution. Then, we estimate $L(p)$ by

$$\hat{L} := \frac{1}{m} \sum_x \mathbb{1}_{\mu_x > 0} \cdot \mathbb{1}_{\nu_x > \tau}.$$

By McDiarmid's inequality, with probability at least $1-2/n$,

$$|\hat{L} - L(p)| \le \frac{\log n}{\sqrt{n}}.$$

We construct a support estimator based on CML and estimator $\hat{L}$, possessing the following guarantee.

Theorem 7. *For any desired estimation accuracy ε satisfying $\varepsilon = \Omega(1/n^{0.49})$, the coverage estimator*

$$\hat{C} := T(p_\varphi) + \hat{L}$$

achieves the optimal $\Theta\left(\frac{m}{\log m} \log \frac{1}{\varepsilon}\right)$ sample complexity.

Similar to Section 7.3, the 0.49 exponent can be replaced by an absolute constant smaller than $1/2$, but a larger value implies a larger asymptotic constant in the complexity. Refer to Appendix B.3 for proof of this theorem.

APPENDIX

In this appendix, we provide full versions for proofs of the theorems in this paper. In particular, we establish Theorem 4, 5, 6, and 7 in Appendix A, B.2, B.4, B.3, respectively, showing the broad applicability of our CML methodology. The proofs leverage several recent advances in techniques for learning distributions and their functionals.

A. Probability Multisets

This section provides the proof of Theorem 4.

First, we define histograms and the relative earth-moving cost and give operational meaning to R_τ. For a distribution p, the *histogram* of a multiset $\mathcal{A} \subseteq \{p\}$ is a mapping,

denoted by $h_{\mathcal{A}} : (0,1] \to \mathbb{Z}_{\geq 0}$, that maps each number $y \in (0,1]$ to the number of times it appears in $\mathcal{A}$. Note that every y corresponds to a probability mass of $y \cdot h_{\mathcal{A}}(y)$. More generally, we also allow *generalized histograms* h with non-integral values $h(y) \in \mathbb{R}_{\geq 0}$.

For any $y_1, y_2 \in (0,1]$, generalized histogram h, and non-negative $m < y_1 \cdot h(y_1)$, we can *move* a probability mass from location y_1 to y_2 by reassigning $h(y_1) - m/y_1$ to y_1, and $h(y_2) + m/y_2$ to y_2. Given $\tau \in [0,1]$, we define the *cost* associated with this operation as

$$c_{\tau,m}(y_1, y_2) := m \cdot \left| \log \frac{y_1 \vee \tau}{y_2 \vee \tau} \right|,$$

and term it as *τ-truncated earth-moving cost*. Note that the cost of multiple operations is additive. With this formulation, $R_\tau(p,q)$ represents the minimal total τ-truncated earth-moving cost associated with any operation schemes of moving $h_{\{p\}}$ to yield $h_{\{q\}}$.

One can readily verify that $c_{\tau,m}(y_1,y_2) = c_{\tau,m}(y_2,y_1)$ and $R_\tau(p,q) = R_\tau(q,p)$, for any locations $y_1, y_2 \in (0,1]$ and distributions $p, q \in \Delta_{\mathcal{X}}$, respectively.

For simplicity, we denote the binomial- and Poisson-type probabilities by

$$\mathrm{bin}(n,x,i) := \binom{n}{i} x^i (1-x)^{n-i}$$

and

$$\mathrm{Poi}(\mu, j) := e^{-\mu} \frac{\mu^j}{j!}.$$

For any absolute constants B and C satisfying $0.1 > B > C > \frac{B}{2} > 0$, define $x_n := \frac{n^B + n^C}{n}$ and $S := \{\frac{1}{n^2}, \frac{2}{n^2}, \ldots, x_n\}$. Consider the following linear program.

For each $x \in S$, define the associated variable v_x

$$\text{Minimize} \sum_{i=1}^{n^B} \left| \varphi_i - \sum_{x \in S} \mathrm{bin}(n,x,i) \cdot v_x \right|$$

$$\text{s.t.} \sum_{x \in S} x \cdot v_x = \sum_{i \leq n^B + 2n^C} \frac{i}{n} \cdot \varphi_i$$

and $\forall x \in S, v_x \geq 0$

Figure 2. Linear program (LP)

For any multiplicity j and solution (v_x) of the LP, we simplify the notation and define

$$\mathrm{P}(v,j) := \sum_{x \in S} \mathrm{Poi}(nx, j) \cdot v_x,$$

approximating the expected *total number* φ_j of symbols appearing j times. Analogously, define

$$\mathrm{P}_\star(v,j) := \sum_{x \in S} \mathrm{Poi}(nx, j) \cdot x \cdot v_x,$$

approximating the expected *total probability mass* M_j of symbols appearing j times, in a sample of size n.

The definitions for quantity $\mathrm{b}(v,j)$ and $\mathrm{b}_\star(v,j)$ are *similar except that* we replace the Poisson probabilities $\mathrm{Poi}(nx,j)$ by their $\mathrm{bin}(n,x,j)$ binomial analogs. Note that these quantities have implicit dependence on the sample size n.

A.1. Existence of A Good Feasible Point

Let p be the underlying distribution and h be its histogram. First, we show that with *high* probability, the linear program LP has a feasible point (v_x) that is *good* in the following sense: 1) the corresponding objective value is relatively *small*; 2) for $\tau \geq n^{-3/2}$, the generalized histogram $h_0 : x \to v_x$ is *close* to $h_n : y \to h(y) \cdot \mathbb{1}_{y \leq x_n}$, admitting a low τ-truncated earth-mover cost.

For each $y \leq x_n$ satisfying $h(y) > 0$, find $x = \min\{x' \in S : x' \geq y\}$ and set $v_x = h(y) \cdot \frac{y}{x}$.

Denote $\mathcal{F} := \sum_{i \leq n^B + 2n^C} \frac{i}{n} \cdot \varphi_i$. By construction,

$$T_n(h) := \sum_{y:\ y \leq x_n, h(y) > 0} h(y) \cdot y = \sum_{x \in S} x \cdot v_x.$$

By the binomial Chernoff bound, the expectation of estimator $\mathcal{F}$ satisfies

$$\mathbb{E}[\mathcal{F}] = \sum_{i \leq n^B + 2n^C} \frac{i}{n} \mathbb{E}[\varphi_i] \geq T_n(h) - \exp(-\Omega(n^{2C-B})).$$

Since changing one observation changes the estimator's value by at most n^{-1}, we bound its tail probability using McDiamid's inequality,

$$\Pr(|\mathcal{F} - \mathbb{E}[\mathcal{F}]| > n^{-\frac{1}{3}+\kappa}) \leq 2\exp(-2n^{\frac{1}{3}+2\kappa}),$$

where $\kappa \in (0, 0.1)$ is a constant to be determined later.

Henceforth we assume $|\mathcal{F} - \mathbb{E}[\mathcal{F}]| \leq n^{-\frac{1}{3}+\kappa}$, holding with probability at least $1 - 2\exp(-2n^{\frac{1}{3}+2\kappa})$. To ensure that (v_x) is a feasible point of the linear program LP, we may need to modify its entries.

For $y \in (0,1]$, let $f_i(y) := \frac{\mathrm{bin}(n,y,i)}{y}$. For $i \geq 1$, we can verify that $|f_i(y)| \leq n$ and $|f_i'(y)| \leq n^2$.

Without any modifications, for $i \leq n^B$, the difference between $\mathbb{E}[\varphi_i] = \sum_{y:h(y)>0} \mathrm{bin}(n,y,i) \cdot h(y)$ and $\mathrm{b}(v,i) = \sum_{x \in S} \mathrm{bin}(n,x,i) \cdot v_x$ is at most $n^{-2} \cdot \sup_{y \in [0,1]} |f_i'(y)| +$

$n\exp(-\Omega(n^{2C-B})) = \mathcal{O}(1)$. Furthermore, by the McDiarmid's inequality,

$$\Pr(|\varphi_i - \mathbb{E}[\varphi_i]| \geq n^{\frac{2}{3}+\kappa}) \leq 2\exp(-2n^{\frac{1}{3}+2\kappa}).$$

Define $m = \mathcal{F}(X^n) - \sum_x x \cdot v_x$ and consider two cases. If $m > 0$, we choose $x = x_n$ and increase v_x by m/x. For any i satisfying $1 \leq i \leq n^B$, this modifies the value of $\mathrm{b}(v,i) = \sum_{x\in S} \mathrm{bin}(n,x,i)\cdot v_x$ by at most $\mathrm{bin}(n,x_n,n^B)\cdot x_n^{-1} \leq \exp(-\Omega(n^{2C-B}))$.

By the assumption that $|\mathcal{F} - \mathbb{E}[\mathcal{F}]| \leq n^{-\frac{1}{3}+\kappa}$,

$$\begin{aligned}\mathcal{F} &\geq \sum_x x\cdot v_x - n^{-\frac{1}{3}+\kappa} - \exp(-\Omega(n^{2C-B}))\\ &\geq \sum_x x\cdot v_x - \mathcal{O}(n^{-\frac{1}{3}+\kappa}).\end{aligned}$$

If $m < 0$, we can remove a total probability mass of at most $\mathcal{O}(n^{-\frac{1}{3}+\kappa})$ by decreasing the entries of (v_x), in an arbitrary manner. Since $|f_i(y)| \leq n$, this operation modifies the value of $\mathrm{b}(v,i)$ by at most $\mathcal{O}(n^{\frac{2}{3}+\kappa})$.

Then, by the union bound, with probability at least $1 - \exp(-n^{\frac{1}{3}+\kappa})$, the objective value of the feasible point (v_x) is at most $n^B\cdot\mathcal{O}(n^{\frac{2}{3}+\kappa}+1) = \mathcal{O}(n^{B+\frac{2}{3}+\kappa})$. Note that $\kappa > 0$ is useful for our subsequent analysis.

Finally, for any level $\tau \geq n^{-3/2}$, the minimal τ-truncated earth-moving cost of moving the generalized histogram h_0 corresponding to (v_x) so that it differs from histogram $h_n : y \to h(y)\cdot \mathbb{1}_{y\leq x_n}$ only at $x = x_n$, is at most

$$\log\left(\frac{n^{-3/2}+n^{-2}}{n^{-3/2}}\right) + \mathcal{O}\left(\frac{\log n}{n^{\frac{1}{3}-\kappa}}\right) = \mathcal{O}(n^{-\frac{1}{3}+2\kappa}).$$

A.2. All Solutions Are Good Solutions

Let (v_x) be the solution described above. In this section, we will show that for any solution (v'_x) to the LP whose objective value is $\mathcal{O}(n^{B+\frac{2}{3}+\kappa})$, the generalized histogram h_1 corresponding to (v'_x) is close to h_0.

Consider the earth-moving scheme described in Valiant & Valiant (2015) that moves all the probability mass to a sequence $\{c_i\}$ of center points satisfying $c_i = \Omega(1/(n\log n))$. We apply this scheme to h_0 and h_1 with the following modification: For any probability mass that should be moved to a center c_i with $c_i > x_n$ under the original earth-moving scheme, we move it to x_n. Since $x_n = \max S$, this modification only reduces the cost of the scheme. By Proposition 5 in Valiant & Valiant (2015), for any $w \in [1,\log n]$ and $\tau = \frac{w}{n\log n}$, the corresponding τ-truncated earth-moving cost is at most $\mathcal{O}(1/\sqrt{w})$.

We first consider h_0. After applying the modified earth-moving scheme, the probability mass at each center $c_i < x_n$ is $\sum_{j\geq 0}\alpha_{i,j}\sum_{x\in S}\mathrm{Poi}(nx,j)xv_x$ for some set of coefficients $\{\alpha_{i,j}\}$ satisfying: $\sum_{j\geq 0}|\alpha_{i,j}| \leq 2n^{0.3}$ for all i; $\alpha_{i,j} = 0$ for $i \leq 0.2\log n \leq j/2$; and $\alpha_{i,j} = \mathbb{1}_{i-1=j}$ for $i > 0.2\log n$.

As for h_1, the probability mass at each center $c_i < x_n$ is $\sum_{j\geq 0}\alpha_{i,j}\mathrm{P}_\star(v',j)$. which differs from that of h_0 by

$$\sum_{j\geq 0}\alpha_{i,j}\,|\mathrm{P}_\star(v',j) - \mathrm{P}_\star(v,j)| \leq \sum_{j\geq 1}\frac{j\alpha_{i,j-1}}{n}\,|\mathrm{P}_\star(v'-v,j)|\,.$$

By our assumption on the corresponding objective values of LP, for any positive integer $i \leq n^B$,

$$|\varphi_i - \mathrm{b}(v,i)| \vee |\varphi_i - \mathrm{b}(v',i)| = \mathcal{O}(n^{B+\frac{2}{3}+\kappa}),$$

which, together with $|\mathrm{Poi}(nx,j) - \mathrm{bin}(n,x,j)| \leq 2x, \forall x \in [0,1]$ (Barbour & Hall, 1984), implies that

$$\begin{aligned}&\sum_{j\geq 1}\alpha_{i,j-1}\frac{j}{n}\,|\mathrm{P}(v'-v,j)|\\ &\leq \sum_{j\geq 1}\alpha_{i,j-1}\frac{j}{n}\left(\mathrm{b}(v'-v,j) + \sum_{x\in S}2x|v'_x - v_x|\right)\\ &\leq 2n^{0.3}\cdot\frac{n^B}{n}\cdot\left(\mathcal{O}(n^{B+\frac{2}{3}+\kappa})+4\right)\\ &= \mathcal{O}(n^{2B+\kappa-1/30}),\end{aligned}$$

where n is sufficiently large so that $n^B > 0.4\log n$.

Therefore, for $\tau \geq 1/(n\log n)$, the minimal τ-truncated earth-moving cost of moving h_0 and h_1 so that they differ only at x_n, is at most

$$\begin{aligned}&n^B\cdot\mathcal{O}(n^{2B+\kappa-1/30})\cdot 2\log n + \log\left(\frac{n^B+n^C}{n^B}\right)\\ &= \mathcal{O}(n^{3B+\kappa-1/30}\log n + n^{C-B})\\ &= \mathcal{O}\left(\frac{1}{\sqrt{\log n}}\right),\end{aligned}$$

where the last step holds for $B = 1.5C = 10\kappa = 0.01$. We consolidate the previous results. For $w \in [1,\log n]$ and $\tau = w/(n\log n)$, with probability at least $1-\exp(-n^{\frac{1}{3}+\kappa})$, the solution to LP will yield a generalized histogram h_1, such that the minimal τ-truncated earth-moving cost of moving h_1 and h_n to make them differ only at x_n, is $\mathcal{O}(1/\sqrt{w})$.

A.3. Competitiveness of PML over Low Probabilities

For the PML distribution associated with a sample satisfying our prior assumptions, denote by h_n^{PML} the histogram corresponding to its entries that are at most x_n.

Leveraging the first theorem in Han & Shiragur (2021), for any $w \in [1,\log n]$ and $\tau = w/(n\log n)$, the minimal τ-truncated earth-moving cost of moving h_n^{PML} and h_n so that

they differ only at x_n, is $\mathcal{O}(1/\sqrt{w})$, with probability at least $1-2\exp(-\Omega(n^{\frac{1}{3}}))$.

On the technical side, our major contribution is the high-confidence guarantee crossing the $1-\exp(-n^{1/3})$ threshold, which is nontrivial to establish. Once we obtain a multiset estimator with such strong concentration guarantees, the theorem in Han & Shiragur (2021) directly shows that PML achieves a comparable performance. Note that prior results in Hao & Orlitsky (2019a); Han & Shiragur (2021) showed that the PML is optimal for estimating the probability multiset under the sorted ℓ_1-distance. Hence, our theorem makes PML the first multiset sample-optimal estimator under both the sorted ℓ_1-distance and the *τ-truncated relative earth-mover distance*.

A.4. Properties of the Empirical Histogram

Denote by h^{EMP} the empirical histogram. By the Chernoff bound, for any symbol s, the probability that

$$|n\cdot p(s)-\mu_s|\geq \mu_s^{3/4} \text{ and } \mu_s > n^B+2n^C$$

is at most $2np(s)\exp(-\Omega(n^{2C-B}))$, and similarly,

$$n\cdot p(s)\geq n^B+4n^C \text{ and } \mu_s \leq n^B+2n^C$$

will happen with probability at most $2\exp(-\Omega(n^{2C-B}))$. Hence, we assume that $|n\cdot p(s)-\mu_s|<\mu_s^{3/4}$ for all symbols s appearing more than n^B+2n^C times, and that any symbol s with probability $p(s)\geq (n^B+4n^C)/n$ appears more than n^B+2n^C times. By the union bound, we will be correct with probability at least $1-4n\exp(-\Omega(n^{2C-B}))$.

Let $y_n := (n^B+4n^C)/n$ for notational convenience. If for each symbol s satisfying $\mu_s\geq n^B+2n^C$, we move a μ_s/n probability mass of h^{EMP} from μ_s/n to $p(s)$, then at all locations $y\geq y_n$, the total discrepancy between the resulting generalized histogram and the actual one is at most

$$\sum_{j>n^B+2n^C}\varphi_j\frac{j^{\frac{3}{4}}}{n}=\frac{1}{n}\sum_{j>n^B+2n^C}\varphi_j^{\frac{1}{4}}(\varphi_j j)^{\frac{3}{4}}\leq n^{-\frac{B}{4}},$$

where the last step follows from Hölder's inequality:

$$\Big(\sum_{j>n^B+2n^C}\varphi_j\Big)^{\frac{1}{4}}\Big(\sum_{j>n^B+2n^C}\varphi_j j\Big)^{\frac{3}{4}}\leq\left(\frac{n}{n^B+2n^C}\right)^{\frac{1}{4}}n^{\frac{3}{4}}.$$

Moreover, the associated total earth-moving cost is bounded from above by

$$\sum_{j>n^B+2n^C}\varphi_j\frac{j}{n}\log\left|\frac{j}{j\pm j^{\frac{3}{4}}}\right|\leq\sum_{j>n^B+2n^C}\varphi_j\frac{j^{\frac{3}{4}}}{n}\leq n^{-\frac{B}{4}}.$$

Hence, the y_n-truncated earth-mover distance between h^{EMP} and h is at most $2n^{-\frac{B}{4}}\log n=\mathcal{O}(n^{-\frac{B}{5}})$, which, combined with the above error probability bound, yields

$$\begin{aligned}\mathbb{E}R_{y_n}(h^{\text{EMP}},h)&\leq O(n^{-\frac{B}{5}})+4n\exp(-\Omega(n^{2C-B}))\log n\\&=O(n^{-\frac{B}{5}}).\end{aligned}$$

In addition, changing any element in the sample sequence changes the value of $R_{y_n}(h^{\text{EMP}},h)$ by at most $(\log n)/n$. Hence, by McDiarmid's inequality, with probability at least $1-2\exp(-2\sqrt{n})$, the value of $R_{y_n}(h^{\text{EMP}},h)$ is less than $O(n^{-\frac{B}{5}})+n^{-\frac{1}{4}}\log n=O(n^{-\frac{B}{5}})$.

A.5. Competitiveness of PML over High Probabilities

Consider the PML histogram h^{PML} and its y_n-truncated earth-mover distance to h. Since there are at most $\exp(3\sqrt{n})$ different profiles (Hardy & Ramanujan, 1918), with probability at least $1-2\exp(-\Omega(\sqrt{n}))$,

$$R_{y_n}(h^{\text{PML}},h)\leq\mathcal{O}(n^{-\frac{B}{5}}).$$

A.6. Performance of PML

We consolidate the previous results. For any $w\in[1,\log n]$ and $\tau=w/(n\log n)$, we design an earth-moving scheme that moves h^{PML} to h.

First, with probability at least $1-2\exp(-\Omega(n^{\frac{1}{3}}))$, we can move the probability mass of h^{PML}_n so that it differs from h_n only at $x_n=(n^B+n^C)/n$ while incurring a τ-truncated earth-mover cost of at most $\mathcal{O}(1/\sqrt{w})$.

Second, by the empirical-histogram argument and $x_n<y_n$, with probability at least $1-2\exp(-\Omega(n^{\frac{1}{3}}))$, we can further move the probability mass of h^{PML} at locations $y\geq x_n$ to coincide with h above y_n while incurring a loss of $\mathcal{O}(n^{-\frac{B}{5}})$.

After the previous two steps, the modified PML histogram differs from the actual histogram h only at locations $y\in I_n=[x_n,y_n]$. Note that the cost of moving a unit mass within I_n is at most $3n^{C-B}$, implying that with probability at least $1-2\exp(-\Omega(n^{\frac{1}{3}}))$,

$$R_\tau(h^{\text{PML}},h)\leq\mathcal{O}\left(\frac{1}{\sqrt{w}}+n^{-\frac{B}{5}}\right)+3n^{C-B}=\mathcal{O}\left(\frac{1}{\sqrt{w}}\right).$$

B. Symmetric Functionals

This section presents the proofs of Theorem 5, 7, and 6. Note that we prove Theorem 6 before 7 since the former proof provides valuable tools for the latter.

B.1. Concentration of Linear Estimators

Define the *sensitivity* of a real function $f:\mathcal{X}^n\to\mathbb{R}$ as the maximum absolute difference $s_n(f)$ in the function value if we change any single coordinate in the input. The following result is a corollary of McDiarmid's inequality.

Lemma 4. *For any distribution p over $\mathcal{X}$, sample $X^n \sim p$, function $f : \mathcal{X}^n \to \mathbb{R}$, and threshold $\tau \geq 0$,*

$$\Pr\left(|f(X^n) - \mathbb{E}[f(X^n)]| > \tau\right) \leq 2\exp\left(-\frac{2\tau^2}{ns_n^2(f)}\right).$$

Next, we consider the concentration property of estimators linear in the φ_i's, the number of symbols appearing i times in sample X^n. Specifically, define

$$\hat{f} := \sum_{i\geq 1} c_i \cdot \varphi_i.$$

The lemma below bounds the sensitivity of $\hat{f}$ in terms of the maximum absolute difference between consecutive $c_i's$.

Lemma 5. *For any n and estimator $\hat{f} := \sum_{i\geq 1} c_i \cdot \varphi_i$, we have $s_n(\hat{f}) \leq 2\max_{i\geq 1}|c_i - c_{i-1}|$, where $c_0 := 0$.*

Proof. Let x^n and y^n be two arbitrary sequences over $\mathcal{X}$ that differ by one element. Let i be the index where $x_i \neq y_i$.

By definition, the multiplicities satisfy the following equalities: $\mu_{x_i}(x^n) = \mu_{x_i}(y^n) + 1$, $\mu_{y_i}(y^n) = \mu_{y_i}(x^n) + 1$, and $\mu_x(x^n) = \mu_x(y^n)$ for $x \in \mathcal{X}$ satisfying $x \neq x_i, y_i$. For simplicity, write $a := \mu_{x_i}(x^n)$, $b := \mu_{y_i}(y^n)$, and for any $i \geq 1$, let $\hat{f}_i := c_{i-1} \cdot \varphi_{i-1} + c_i \cdot \varphi_i$.

The first equality implies that $\varphi_a(x^n) = \varphi_a(y^n) + 1$ and $\varphi_{a-1}(x^n) = \varphi_{a-1}(y^n) - 1$. And hence, $\hat{f}_a(x^n) - \hat{f}_a(y^n) = c_a - c_{a-1}$. By analogy, the second multiplicity equality implies $\hat{f}_b(x^n) - \hat{f}_b(y^n) = -c_b + c_{b-1}$. The third equality combines these two results to yield

$$\hat{f}(x^n) - \hat{f}(y^n) = c_a - c_{a-1} + (-c_b + c_{b-1}).$$

The lemma then follows by the triangle inequality. □

By the above two lemmas, we have the following result for the concentration of linear estimators.

Corollary 4. *For any $\tau \geq 0$, distribution p over $\mathcal{X}$, sample $X^n \sim p$, and estimator $\hat{f} := \sum_{i\geq 1} c_i \cdot \varphi_i(X^n)$,*

$$\Pr(|\hat{f} - \mathbb{E}[\hat{f}]| \geq \tau) \leq 2\min_{i\geq 1} \exp\left(-\frac{2\tau^2}{n(c_i - c_{i-1})^2}\right).$$

B.2. Shannon Entropy

We begin with the fundamental entropy estimation problem. The *Shannon entropy* of a distribution p over $\mathcal{X} = [N]$ is

$$H(p) := \sum h(p(x)) := \sum_x -p(x)\log p(x).$$

Shannon entropy is the primary measure of information with wide applications to machine learning and neuroscience.

Given sample access to an unknown p, we draw a pair of independent samples $(X^n, Y^n) \sim p$. Denote by $\hat{p}$ and $\hat{q}$ the respective empirical distributions, and μ_x and ν_x the respective empirical counts of an arbitrary symbol $x \in [N]$.

We separate the empirical counts into a few categories by four thresholds of order-$\log n$,

$$\tau_i := c_i \cdot \log n,\ \forall i \in [4],$$

where c_i's are positive absolute constants to be determined later. Since p is unknown, we perform a *soft truncation* and partition our target $H(p)$ into a *low-probability part*

$$T(p) := \sum_x h(p(x)) \cdot \Pr\left(\mu_x \leq \tau_1\right)$$

and a *high-probability part*

$$L(p) := H(p) - T(p).$$

We estimate $L(p)$ by a simple variant of the Miller-Mallow estimator (Carlton, 1969):

$$\hat{L} := \sum_x \left(h\left(\hat{p}(x)\right) + \frac{1}{2n}\right) \cdot \left(1 - \mathbb{1}_{\mu_x \leq \tau_2,\, \nu_x \leq \tau_1}\right).$$

For $c_2 \gg c_1 \gg 1$, the estimator's bias satisfies

$$\left|\mathbb{E}[\hat{L}] - L(p)\right| = \mathcal{O}\left(\frac{N}{n\log n}\right).$$

The following result shows that a simple combination of CML and the Miller-Mallow variant is sample-optimal for nearly the entire applicable accuracy range.

Theorem 8. *For any desired estimation accuracy ε satisfying $\varepsilon = \Omega(1/n^{0.49})$, the entropy estimator*

$$\hat{H} := T(p_\varphi) + \hat{L}$$

achieves the optimal $\Theta(N/(\varepsilon \log N))$ sample complexity.

Proof. As the previous sections demonstrated, to establish the above theorem, we need to find another estimator for $T(p)$ that highly concentrates around its mean value.

We will make use of an estimator similar to that in Wu & Yang (2016). Let $g(z) := \sum_{i=0}^{\tau_4} a_i \cdot z^i$ denote the min-max polynomial approximation of $h(z)$ over $I_n := [0, \tau_3/n]$.

For any natural numbers A, B, denote by $A^{\underline{B}}$ the order-B *falling factorial* of A. Consider the following estimator.

$$\hat{T} := \sum_x \left(\sum_{i=0}^{\tau_4} a_i \cdot \frac{\mu_x^{\underline{i}}}{n^{\underline{i}}}\right) \mathbb{1}_{\mu_x \leq \tau_2} \cdot \mathbb{1}_{\nu_x \leq \tau_1}.$$

Choose parameters $c_2 \gg c_3 \gg c_1$ and $1 \gg c_4$. Following the derivations in Wu & Yang (2016), and utilizing the Chernoff bound and $\max_{z\in I_n}|g(z) - h(z)| = \mathcal{O}\left(1/(n\log n)\right)$,

we bound the bias of $\hat{T}$ by $\mathcal{O}(N/(n\log n))$. In addition, since $|a_i| = \mathcal{O}(2^{3\tau_4}(n/\log n)^{i-1})$, for any absolute constant $\lambda \in (0, 1/2)$, we can pick a sufficiently small c_2 so that the sensitivity of $\hat{T}$ is at most $\mathcal{O}(n^\lambda/n)$.

One technical subtlety here is that the (X^n, Y^n) splitting breaks the compressor formulation in Section 7.1. An easy fix is to view (X^n, Y^n) as a single size-$2n$ sample and consider all the possible equal splittings. For each of them, our estimator $\hat{T}$ yields an estimate. Summing up all such estimates and taking the average as the final estimate will maintain both the bias of the estimator and the sensitivity bound we derived above.

Due to the two indicator functions in the definition of $\hat{T}$, we can view $\hat{T}$ as an estimator over the co-domain of a compressor in Section 7.1 with parameter $t = (\tau_1 + \tau_2)$. Lemma 4 and sensitivity bound $\mathcal{O}(n^\lambda/n)$ on $\hat{T}$ yield that

$$\Pr(|\hat{T} - \mathbb{E}[\hat{T}]| \geq \tau) \leq 2\exp(-\Omega(\tau^2 n^{1-2\lambda})),\ \forall \tau \geq 0.$$

The triangle inequality combines this with the bias bound to yield, with probability at least $1 - 2\exp(-\Omega(\tau^2 n^{1-2\lambda}))$,

$$|\hat{T} - T(p)| - \tau = \mathcal{O}\left(\frac{N}{n\log n}\right),\ \forall \tau \geq 0.$$

For $\tau = \Omega((\log^{1.5} n)/n^{1/2-\lambda})$, the estimator achieves a near-optimal error guarantee, with probability at least $1 - 2\exp(-\log^3 n)$. Finally, Corollary 3 requires only the error probability to be smaller than $n^t = n^{\tau_1+\tau_2} = n^{\Theta(\log n)}$, holding trivially for large enough n.

The work of Carlton (1969); Wu & Yang (2016) has already established the optimality of the large-probability estimator $\hat{L}$, and our theorem follows from here. □

B.3. Support Coverage

For a given parameter m, the *normalized support coverage* of a distribution p is

$$C(p) := \sum_x c(p(x)) := \sum_x \frac{1 - (1 - p(x))^m}{m},$$

the (normalized) expected number of distinct symbols in a size-m sample. Support coverage estimation closely relates to the well-known unseen species problem arising in biological and ecological studies.

Adopting the notation in Section 7.3, we can partition $C(p)$ into a *low-probability part*

$$T(p) := \sum_x c(p(x)) \cdot \Pr(\mu_x \leq \tau)$$

and a *high-probability part*

$$L(p) := C(p) - T(p),$$

where we draw an independent sample pair $(X^n, Y^n) \sim p$, and denote by μ_x and ν_x the respective empirical counts of an arbitrary symbol $x \in [N]$.

We assume that $N \gg m$ without loss of generality since one can always append symbols of zero probability to a distribution. Then, we estimate $L(p)$ by

$$\hat{L} := \frac{1}{m}\sum_x \mathbb{1}_{\mu_x > 0} \cdot \mathbb{1}_{\nu_x > \tau}.$$

By McDiarmid's inequality, with probability at least $1 - 2/n$,

$$|\hat{L} - L(p)| \leq \frac{\log n}{\sqrt{n}}.$$

We construct a support estimator based on CML and estimator $\hat{L}$, possessing the following guarantee.

Theorem 9. *For any desired estimation accuracy ε satisfying $\varepsilon = \Omega(1/n^{0.49})$, the coverage estimator*

$$\hat{C}_m := T(p_\varphi) + \hat{L}$$

achieves the optimal $\Theta\left(\frac{m}{\log m}\log\frac{1}{\varepsilon}\right)$ sample complexity.

Proof. By Acharya et al. (2017b), for any positive absolute constant α, error parameter $\varepsilon \geq 6n^\alpha/n$, and parameters m, n satisfying $2n \leq m \leq \alpha\frac{n\log(n/2^{1/\alpha})}{\log(3/\varepsilon)}$, there is a linear estimator $\hat{C}_0 := \sum_{i\geq 1} c_i \cdot \varphi_i$ such that

$$\left|C(p) - \mathbb{E}[\hat{C}_0]\right| \leq \frac{3n^\alpha}{m} + \frac{\varepsilon}{3}\cdot\frac{m \wedge N}{m}$$

and $\max_i |c_i| \leq n^{\alpha-1}$. For $c_2 = 10c_1$, we leverage sample Y^n and estimate $T(p)$ by

$$\hat{C} := \sum_x \sum_{i=1}^{\tau_2} c_{\mu_x = i} \cdot \mathbb{1}_{\nu_x \leq \tau_1},$$

where $c_{\mu_x=1} := \mathbb{1}_{\mu_x = i} \cdot c_i$. The bias of this estimator admits

$$\begin{aligned}
\left|\mathbb{E}[\hat{C}] - T(p)\right| &\leq \sum_x |\mathbb{E}[c_{\mu_x}] - c_m(p(x))| \cdot \mathbb{E}[\mathbb{1}_{\mu_x \leq \tau_1}] \\
&\quad + \sum_x \left|\mathbb{E}\left[\sum_{i>\tau_2} c_{\mu_x = i}\right] \cdot \mathbb{E}[\mathbb{1}_{\mu_x \leq \tau_1}]\right| \\
&\leq \frac{3n^\alpha}{n} + \frac{\varepsilon}{3} + \frac{n^\alpha}{n} \cdot \sum_x np(x) \times \\
&\qquad \mathbb{E}[\mathbb{1}_{\mu_x(X^{n-1}) \geq \tau_2} \cdot \mathbb{1}_{\nu_x \leq \tau_1}] \\
&\leq \varepsilon,
\end{aligned}$$

where the last step follows by assuming that c_1 is sufficiently large and the Chernoff bound for binomial random variables.

Note that changing one point in X^n or Y^n changes the value of $\hat{C}$ by at most $4n^\alpha/n$. Viewing $Z^{2n} := (X^n, Y^n)$

as a single sample, we can apply $\hat{C}$ to all the equal-size partitions of Z^{2n} and re-denote by $\hat{C}$ the average of all the induced estimates. The resulting estimator corresponds to a compressor in Section 7.1 with $t := \tau_1 + \tau_2$, and has the same bias and sensitivity bound as the original.

For any parameter $\tau \geq 0$, Lemma 4 and the n-sensitivity bound $\mathcal{O}(n^\alpha/n)$ yield that

$$\Pr\left(|\hat{C} - \mathbb{E}[\hat{C}]| \geq \tau\right) \leq 2\exp(-\Omega(\tau^2 n^{1-2\alpha})).$$

Applying Corollary 3, we establish a similar guarantee for the CML plug-in estimator, i.e., with probability at least $1 - 6n^t \exp(-\Omega(\tau^2 n^{1-2\alpha}))$,

$$|T(p) - T(p_\varphi)| \leq 2\tau + 2\varepsilon.$$

The error probability vanishes as fast as $2\exp(-\log^2 n)$ for $\tau = \Omega((\log n)/n^{1/2-\alpha})$.

We consider the large-probability estimator $\hat{L}$. We will make the dependence on m explicit by writing $c_m(z) := (1-z)^m$.

The bias of estimator $\hat{L}$ satisfies

$$\begin{aligned}
\left|\mathbb{E}[\hat{L}] - L(p)\right| &= \left|\frac{1}{m}\sum_x (c_n - c_m)(p(x)) \cdot \mathbb{E}\left[\mathbb{1}_{\mu_x > \tau_1}\right]\right| \\
&\leq \sum_x p(x) c_n(p(x)) \cdot \mathbb{E}\left[\mathbb{1}_{\mu_x \geq \tau_1}\right] \\
&\leq \sum_{x: p(x) < \tau_1/n} p(x) \cdot \mathbb{E}\left[\mathbb{1}_{\mu_x(X^{n-1}) \geq \tau_1}\right] \\
&+ \sum_{x: p(x) \geq \tau_1/n} p(x)\left(1 - \frac{\tau_1}{n}\right)^n \\
&\leq 2\exp(-\Omega(c_1 \log n)),
\end{aligned}$$

where the last step follows by the Chernoff bound. Hence, the bias is $\mathcal{O}(1/n)$ for sufficiently large c_1. Furthermore, the sensitivity of $\hat{L}$ is exactly $1/m < 1/n$.

By McDiarmid's inequality, with probability at least $1 - 2\exp(-\log^2 n)$,

$$|\hat{L} - L(p)| = \Theta\left(\frac{\log n}{\sqrt{n}}\right).$$

Consolidating these previous results yields the theorem. □

B.4. Support Size

Support size is an essential distribution attribute arising in vocabulary size, population, and database research.

Following the previous discussion, we consider estimating the *normalized support size*

$$S(p) := \sum_x s(p(x)) := \sum_x \frac{\mathbb{1}_{p(x)>0}}{N}.$$

The problem is *ill-defined* without additional assumptions since symbols with arbitrarily small probability mass can modify the quantity by a nontrivial constant. A common assumption for support estimation requires the minimal non-zero probability of p to be at least $1/N$. In the following derivations, we will assume that this bound holds.

Analogous to Section 7.3, we draw a sample $X^n \sim p$ and separate the empirical counts by a threshold of order-$\log n$,

$$\tau := c \log n,$$

for some absolute constant c to be determined later.

For any $x \in [N]$, let μ_x be the number of times x appearing in the sample. We partition $S(p)$ into a *low-probability part*

$$T(p) := \sum_x s(p(x)) \cdot \Pr\left(\mu_x \leq \tau\right)$$

and a *high-probability part*

$$L(p) := S(p) - T(p).$$

A natural choice for $L(p)$ is the unbiased estimator

$$\hat{L} := \sum_x s\left(\mu_x - \tau\right).$$

By McDiarmid's inequality, with probability at least $1 - 2/n$,

$$|\hat{L} - L(p)| \leq \frac{\log^2 n}{\sqrt{n}}.$$

We construct a support estimator with the following guarantee based on CML and the unbiased estimator.

Theorem 10. *For any desired estimation accuracy ε satisfying $\varepsilon = \Omega(1/n^{0.49})$, the support estimator*

$$\hat{S} := T(p_\varphi) + \hat{L}$$

achieves the optimal $\Theta\left(\frac{N}{\log N} \log^2 \frac{1}{\varepsilon}\right)$ sample complexity.

Proof. For clarity, let T_S and T_C be the small-probability parts of support size and coverage, respectively. We proceed by relating T_S to T_C. Note that we replaced $2n$ with n in T_C. For any error parameter ε, choose $m = N \log(1/\varepsilon)$,

$$\begin{aligned}
&|T_S(p) - T_C(p) \cdot \log(1/\varepsilon)| \\
&= \left|\frac{1}{N}\sum_x (\mathbb{1}_{p(x)>0} - m \cdot c(p(x))) \Pr\left(\mu_x \leq \tau_1\right)\right| \\
&\leq \frac{1}{N}\sum_x (1 - p(x))^m \\
&\leq \varepsilon.
\end{aligned}$$

Hence by results in the last section, for $\varepsilon \geq 12n^{\alpha}/n$, and N, n such that $n \leq N\log(1/\varepsilon) \leq \alpha\frac{n\log(n/2^{1+1/\alpha})}{2\log(3/\varepsilon)}$, the bias of $\hat{C}\cdot\log(1/\varepsilon)$ in estimating $T_S(p)$ satisfies

$$\begin{aligned}
&|\mathbb{E}[\hat{C}]\cdot\log(1/\varepsilon) - T_S(p)| \\
&\leq |\mathbb{E}[\hat{C}] - T_C(p)|\log(1/\varepsilon) + |T_C(p)\log(1/\varepsilon) - T_S(p)| \\
&\leq \left(\frac{3n^{\alpha}}{m} + \frac{\varepsilon}{3}\cdot\frac{m\wedge N}{m}\right)\cdot\log(1/\varepsilon) + \varepsilon \\
&\leq \frac{n^{\alpha}}{n}\log n + \frac{4\varepsilon}{3}.
\end{aligned}$$

Moreover, changing one observation in sample X^n modifies $\hat{C}\cdot\log(1/\varepsilon)$ by at most $(8n^{\alpha}\log n)/n$. Hence by Lemma 4, for any real parameter $\tau_0 \geq 0$ and with probability at most $2\exp(-\Omega(\tau_0^2\cdot n^{1-2\alpha}/\log^2 n))$,

$$\left|\hat{C}\cdot\log(1/\varepsilon) - \mathbb{E}[\hat{C}\cdot\log(1/\varepsilon)]\right| \geq \tau_0.$$

Applying Corollary 3 and letting $t = \tau_1 + \tau_2$ yield, with probability at least $1 - 2\exp(-\Omega(\tau^2\cdot n^{1-2\alpha}/\log^2 n))\cdot en^t$,

$$|T_C(p) - T_C(p_{\varphi})| \leq 2\tau + \frac{8\varepsilon}{3} + \frac{2n^{\alpha}}{n}\log n.$$

For any $\tau = \Omega((\log n)^2/n^{1/2-\alpha})$, the right-hand side vanishes as fast as $2\exp(-\log^2 n)$.

Consolidating the previous results yields the theorem. The estimator's optimality follows by

$$N\log\frac{1}{\varepsilon} \leq \alpha\frac{n\log(n/2^{1+1/\alpha})}{2\log(3/\varepsilon)},$$

matching with the tight lower bound $n = \Omega\left(\frac{N}{\log N}\log^2\frac{1}{\varepsilon}\right)$ in a paper by Wu & Yang (2019). □