Dimension-Free Empirical Entropy Estimation

Anonymous Author(s)
Affiliation
Address
email

Abstract

We seek an entropy estimator for discrete distributions with fully empirical accuracy bounds. As stated, this goal is infeasible without some prior assumptions on the distribution. We discover that a certain information moment assumption renders the problem feasible. We argue that the moment assumption is natural and, in some sense, *minimalistic* — weaker than finite support or tail decay conditions. Under the moment assumption, we provide the first finite-sample entropy estimates for infinite alphabets, nearly recovering the known minimax rates. Moreover, we demonstrate that our empirical bounds are significantly sharper than the state-of-the-art bounds, for various natural distributions and non-trivial sample regimes. Along the way, we give a dimension-free analogue of the Cover-Thomas result on entropy continuity (with respect to total variation distance) for finite alphabets, which may be of independent interest.

1 Introduction

Estimating the entropy of a discrete distribution based on a finite iid sample is a classic problem with theoretical and practical ramifications. Considerable progress has been made in the case of a finite alphabet, and the countably infinite case has also attracted a fair amount of attention in recent years. A less-addressed issue is one of *empirical* accuracy estimates: data-dependent bounds adaptive to the particular distribution being sampled.

Our point of departure is the simpler (to analyze) problem of estimating a discrete distribution $\boldsymbol{\mu}$ in total variation norm $\|\cdot\|_{\mathsf{TV}} = \frac{1}{2}\|\cdot\|_1$, where the most recent advance was made by Cohen et al. [2020]; see therein for a literature review. If $\boldsymbol{\mu}$ is a distribution on $\mathbb{N}$ and $\hat{\boldsymbol{\mu}}_n$ is its empirical realization based on a sample of size n, then Theorem 2.1 of Cohen et al. states that with probability at least $1-\delta$,

$$\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}_n\|_1 \quad \leq \quad \frac{2}{\sqrt{n}} \sum_{j \in \mathbb{N}} \sqrt{\hat{\boldsymbol{\mu}}_n(j)} + 6\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{1}$$

This bound has the advantage of being valid for all distributions on $\mathbb{N}$, without any prior assumptions, and being fully empirical: it yields a risk estimate that is computable based on the observed sample, not depending on any unknown quantities. (Additionally, Cohen et al. argue that (1) is near-optimal in a well-defined sense.) The question we set out to explore in this paper is: What analogues of (1) are possible for discrete entropy estimation?

When $\boldsymbol{\mu}$ has support size $d < \infty$, an answer to our question is readily provided by combining (1) with Cover and Thomas [2006, Theorem 17.3.3], which asserts that, for $\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_1 \leq 1/2$, we have

$$|\mathrm{H}(\boldsymbol{\mu}) - \mathrm{H}(\boldsymbol{\nu})| \quad \leq \quad \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_1 \log \frac{d}{\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_1}, \tag{2}$$

where $\mathrm{H}(\cdot)$ is the entropy functional defined in (3). Indeed, taking $\boldsymbol{\mu}$ as in (1) and $\boldsymbol{\nu}$ to be $\hat{\boldsymbol{\mu}}_n$ yields a fully empirical estimate on $|\mathrm{H}(\boldsymbol{\mu}) - \mathrm{H}(\hat{\boldsymbol{\mu}}_n)|$. For fixed $d < \infty$, no technique relying on the

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

plug-in estimator can yield minimax rates [Wu and Yang, 2016]. The plug-in is, however, minimax optimal for $d = \infty$, [Paninski, 2003] and is among the few methods for which explicitly computable finite-sample risk bounds are known.

The thrust of this paper is to replace the restrictive finite-support assumption with considerably more general moment conditions. It is well-known that when estimating the mean of some random variable X, the first-moment assumption $\mathbb{E}|X| \leq M$ is not sufficient to yield any finite-sample information.[1] Strengthening the assumption to $\mathbb{E}|X|^\alpha \leq M$, for any $\alpha > 1$, immediately yields finite-sample empirical estimates on $\left|\mathbb{E}X - \frac{1}{n}\sum_{i=1}^n X_i\right|$ via the von Bahr and Esseen [1965] inequality.[2] In this sense, a bound on the $(1+\varepsilon)$th moment is a *minimal* requirement for empirical mean estimation. However, it is not immediately obvious how to apply this insight to the entropy estimation problem: the corresponding random variable is $X = -\log \boldsymbol{\mu}(I)$, where $I \sim \boldsymbol{\mu}$, but rather than being given iid samples of X, we are only given draws of I.

Our contribution. In Theorem 1, we provide a dimension-free analogue of (2), which, combined with (1), allows for empirical accuracy bounds on the plug-in entropy estimator under a minimalistic moment assumption. Moreover, for this rich class of distributions, the plug-in estimator turns out to be asymptotically optimal, as we show in Theorem 4. Our moment assumption is natural and essentially the weakest one that makes *any* empirical bounds feasible, as we argue in Theorem 3. As we demonstrate in Section 6, the rates provided by our empirical bound compare favorably against the state of the art.

2 Definitions and notation

Our logarithms will always be base e by default. For discrete distributions, there is no loss of generality in taking the domain to be the natural numbers $\mathbb{N} = \{1, 2, 3, \ldots\}$. For $k \in \mathbb{N}$, we write $[k] := \{i \in \mathbb{N} : i \leq k\}$. The set of all probability distributions on $\mathbb{N}$ will be denoted by $\Delta_\mathbb{N}$. For $d \in \mathbb{N}$, we write $\Delta_d \subset \Delta_\mathbb{N}$ to denote those $\boldsymbol{\mu}$ whose support is contained in $[d]$.

We define the operator $(\cdot)^\downarrow$, which maps any $\boldsymbol{\mu} \in \Delta_\mathbb{N}$ to its non-increasing rearrangement $\boldsymbol{\mu}^\downarrow$. The set of all non-increasing distributions will be denoted by $\Delta_\mathbb{N}^\downarrow := \left\{\boldsymbol{\mu}^\downarrow : \boldsymbol{\mu} \in \Delta_\mathbb{N}\right\}$.

We write $\mathbb{R}_+ := [0, \infty)$. For any $\xi : \mathbb{N} \to \mathbb{R}_+$ and $\alpha \geq 0$, define

$$\mathrm{H}^{(\alpha)}(\xi) := \sum_{j \in \mathbb{N} : \xi(j) > 0} \xi(j) \left|\log \xi(j)\right|^\alpha . \tag{3}$$

For $\xi \in \mathbb{R}^\mathbb{N}$, denote by $|\xi| \in \mathbb{R}_+^\mathbb{N}$ the elementwise application of $|\cdot|$ to ξ. When $\xi \in \Delta_\mathbb{N}$ and $\alpha = 1$, (3) recovers the standard definition of entropy, which we denote by $\mathrm{H}(\xi) := \mathrm{H}^{(1)}(\xi)$. For general $\alpha > 0$, this quantity may be referred to as the αth *moment of information*. For $h \geq 0$, define

$$\Delta_\mathbb{N}^{(\alpha)}[h] = \left\{\boldsymbol{\mu} \in \Delta_\mathbb{N} : \mathrm{H}^{(\alpha)}(\boldsymbol{\mu}) \leq h\right\}$$

and also $\Delta_\mathbb{N}^{(\alpha)} := \bigcup_{h \geq 0} \Delta_\mathbb{N}^{(\alpha)}[h]$ and $\Delta_\mathbb{N}^{\downarrow(\alpha)}[h] := \Delta_\mathbb{N}^\downarrow \cap \Delta_\mathbb{N}^{(\alpha)}[h]$.

For $n \in \mathbb{N}$ and $\boldsymbol{\mu} \in \Delta_\mathbb{N}$, we write $\boldsymbol{X} = (X_1, \ldots, X_n) \sim \boldsymbol{\mu}^n$ to mean that the components of the vector $\boldsymbol{X}$ are drawn iid from from $\boldsymbol{\mu}$. The empirical measure $\hat{\boldsymbol{\mu}}_n \in \Delta_\mathbb{N}$ induced by the sample $\boldsymbol{X}$ is defined by $\hat{\boldsymbol{\mu}}_n(j) = \frac{1}{n}\sum_{i \in [n]} \mathbf{1}[X_i = j]$. For any $\xi \in \mathbb{R}^\mathbb{N}$ and $0 < p < \infty$, the ℓ_p (pseudo)norm is defined by $\|\xi\|_p^p = \sum_{j \in \mathbb{N}} |\xi(j)|^p$ and $\|\xi\|_\infty = \sup_{j \in \mathbb{N}} |\xi(j)|$.

For $\alpha, h > 0$, and $n \in \mathbb{N}$, define the L_1 *minimax risk* for the αth moment by

$$\mathcal{R}_n^{(\alpha)}(h) \quad := \quad \inf_{\hat{H}} \sup_{\boldsymbol{\mu} \in \Delta_\mathbb{N}^{(\alpha)}[h]} \mathbb{E}|\hat{H}(X_1, \ldots, X_n) - \mathrm{H}(\boldsymbol{\mu})|, \tag{4}$$

where the infimum is over all mappings $\hat{H} \colon \mathbb{N}^n \to \mathbb{R}_+$.

[1] Even distinguishing, for $X \geq 0$, between $\mathbb{E}X = 0$ and $\mathbb{E}X = M$ based on a finite sample is impossible with any degree of confidence. Of course, $\frac{1}{n}\sum_{i=1}^n X_i \to \mathbb{E}X$ almost surely, by the strong law of large numbers.

[2] Put $Y = X - \mathbb{E}X$; then $\mathbb{E}|Y| \leq 2M$. For $1 < \alpha < 2$, a sharper version of the Bahr-Esseen inequality [Pinelis, 2015] states that $\mathbb{E}\left[\left|\sum_{i=1}^n Y_i\right|^\alpha\right] \leq 2n(2M)^\alpha$, which implies tail bounds via Markov's inequality. Better rates are available via the median-of-means estimator, see Lugosi and Mendelson [2019].

3 Main results

Our first result is a dimension-free analogue of (2):

Theorem 1. *For all $h > 0$ and $\alpha > 1$, $\mathrm{H} : \Delta_{\mathbb{N}}^{(\alpha)}[h] \to \mathbb{R}_+$ is uniformly continuous under ℓ_1. In particular, for all $\boldsymbol{\mu}, \boldsymbol{\nu} \in \Delta_{\mathbb{N}}^{(\alpha)}$ satisfying $\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_\infty < 1/2$, we have*

$$\begin{aligned} |\mathrm{H}(\boldsymbol{\mu}) - \mathrm{H}(\boldsymbol{\nu})| &\leq \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_1^{1-1/\alpha} \left(2\alpha^\alpha + \mathrm{H}^{(\alpha)}(\boldsymbol{\mu}) + \mathrm{H}^{(\alpha)}(\boldsymbol{\nu})\right)^{1/\alpha} \\ &\leq \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_1^{1-1/\alpha} \left(2\alpha + \mathrm{H}^{(\alpha)}(\boldsymbol{\mu})^{1/\alpha} + \mathrm{H}^{(\alpha)}(\boldsymbol{\nu})^{1/\alpha}\right). \end{aligned}$$

The requirement in Theorem 1 that $\alpha > 1$ cannot be dispensed with, as the function $\mathrm{H} : \Delta_{\mathbb{N}}^{(\alpha)}[h] \to \mathbb{R}_+$ is not continuous under ℓ_1 for $\alpha = 1$ (see Remark following Lemma 5), and, a fortiori, is not uniformly continuous. Thus, there can be no function $F : \mathbb{R}_+^2 \to \mathbb{R}_+$ satisfying

$$|\mathrm{H}(\boldsymbol{\mu}) - \mathrm{H}(\boldsymbol{\nu})| \leq F(\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_1, h), \qquad h > 0, \boldsymbol{\mu}, \boldsymbol{\nu} \in \Delta_{\mathbb{N}}^{(1)}[h]$$

with the additional property that for any two sequences $\boldsymbol{\mu}_n, \boldsymbol{\nu}_n \in \Delta_{\mathbb{N}}$ satisfying $\varepsilon_n := \|\boldsymbol{\mu}_n - \boldsymbol{\nu}_n\|_1 \to 0$, it holds that $F(\varepsilon_n, h) \to 0$. Perhaps surprisingly,[3] it turns out that $\mathrm{H} : \Delta_{\mathbb{N}}^{(\alpha)}[h] \to \mathbb{R}_+$ *is* uniformly continuous under ℓ_p for all $\alpha > 1$, $p \in [1, \infty]$:

Theorem 2. *There is a function $F : \mathbb{R}_+^4 \to \mathbb{R}_+$ such that*

$$|\mathrm{H}(\boldsymbol{\mu}) - \mathrm{H}(\boldsymbol{\nu})| \leq F(\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_p, h, \alpha, p), \qquad h > 0, \alpha > 1, p \in [1, \infty], \boldsymbol{\mu}, \boldsymbol{\nu} \in \Delta_{\mathbb{N}}^{(\alpha)}[h]$$

with the additional property that whenever $\varepsilon_n := \|\boldsymbol{\mu}_n - \boldsymbol{\nu}_n\|_p \to 0$, we have $F(\varepsilon_n, h, \alpha, p) \to 0$.

Remark. Although Theorem 2 establishes uniform continuity, it gives no hint as to the functional dependence of the modulus of continuity F on α, p, h, and $\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_p$. We leave this as a fascinating open problem — even though the practical applications are likely to be limited: it follows from Wyner and Foster [2003] that for $p = \alpha = 2$ and fixed h, $F(\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2, h, 2, 2)$ cannot decay at a faster rate than $1/\log(1/\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2)$.

Combining Theorem 1 with (1) yields an empirical (under moment assumptions) bound for the plug-in entropy estimator:

Corollary 1. *For all $\alpha > 1$, $h > 0$, $\delta \in (0, 1)$, $n \geq 2 \log \frac{4}{\delta}$, and $\boldsymbol{\mu} \in \Delta_{\mathbb{N}}^{(\alpha)}[h]$, we have that*

$$|\mathrm{H}(\boldsymbol{\mu}) - \mathrm{H}(\hat{\boldsymbol{\mu}}_n)| \leq \left(2\alpha^\alpha + h + \mathrm{H}^{(\alpha)}(\hat{\boldsymbol{\mu}}_n)\right)^{1/\alpha} \left(\frac{2\|\hat{\boldsymbol{\mu}}_n\|_{1/2}^{1/2}}{\sqrt{n}} + 6\sqrt{\frac{\log(4/\delta)}{2n}}\right)^{1-1/\alpha}$$

holds with probability at least $1 - \delta$.

In Section 6, we compare the rates implied by Corollary 1 to the state of the art on various distributions.

Next, we examine the optimality of the plug-in estimate by analyzing the minimax risk, defined in (4). It was known [Silva, 2018, Appendix A] that assuming $\mathrm{H}(\boldsymbol{\mu}) < \infty$ does not suffice to yield a minimax rate for the L_2 risk:

$$\inf_{\hat{H} : \mathbb{N}^n \to \mathbb{R}_+} \sup_{\boldsymbol{\mu} \in \Delta_{\mathbb{N}}^{(1)}} \mathbb{E}\left(\hat{H}(X_1, \ldots, X_n) - \mathrm{H}(\boldsymbol{\mu})\right)^2 = \infty.$$

This technique yields an analogous result for the L_1 risk as well. We strengthen these results in two ways: (i) by lower-bounding the L_1 risk (rather than L_2, which is never smaller), and (ii) by restricting $\boldsymbol{\mu}$ to $\Delta_{\mathbb{N}}^{(1)}[h]$ and obtaining a finitary, quantitative lower bound:

Theorem 3. *For $\alpha = 1$, there is a universal constant $C > 0$ such that for all $h > 1$ and $n \in \mathbb{N}$, we have $\mathcal{R}_n^{(1)}(h) \geq Ch$.*

[3] Since ℓ_1 dominates all of the ℓ_p norms, continuity of a function under ℓ_p trivially implies continuity under ℓ_1, but the reverse implication is generally not true.

Remark. The above result complements — but is not directly comparable to — Antos and Kontoyiannis [2001a, Theorem 4]. Ours gives a quantitative dependence on h but constructs an adversarial distribution for each sample size n; theirs is asymptotic only but a single adversarial distribution suffices for all n.

Remark. Our technique immediately yields a lower bound of Ch^2 on the L_2 minimax risk.

In contradistinction to the $\alpha = 1$ case, where no minimax rate exists, we show that the plug-in estimator is minimax for all $\alpha > 1$:

Theorem 4. *The following bounds hold for the L_1 minimax risk:*

(a) Upper bound: for all $h > 0, \alpha > 1$,

$$\mathcal{R}_n^{(\alpha)}(h) \quad \leq \quad \frac{1 + \log n}{\sqrt{n}} + \frac{2^{\alpha-1} h}{\log^{\alpha-1} n}, \qquad n \in \mathbb{N};$$

further, this bound is achieved by the plug-in estimate $\mathrm{H}(\hat{\boldsymbol{\mu}}_n)$.

(b) Lower bound: for each $\alpha > 0$, $n \in \mathbb{N}$ there is an $h > 0$ such that

$$\mathcal{R}_n^{(\alpha)}(h) \quad \geq \quad \frac{h}{4 \cdot 3^{\alpha} \log^{\alpha-1} n}.$$

Open problem. Close the gap in the dependence on α in the upper and lower bounds.

4 Related work

Continuity and convergence. Zhang [2007] gave a sharpened version of (2) and Ho and Yeung [2010] presented analogous bounds; Audenaert [2007] proved a non-commutative generalization. Sason [2013, Theorem 5] upper-bounds $|\mathrm{H}(\boldsymbol{\mu}) - \mathrm{H}(\boldsymbol{\nu})|$ in terms of quantities related to $\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_1$, where (at most) one of them is allowed to have infinite support. Even though $\mathrm{H}(\cdot)$ is not continuous on $\Delta_{\mathbb{N}}$, the plug-in estimate $\mathrm{H}(\hat{\boldsymbol{\mu}}_n)$ converges to $\mathrm{H}(\boldsymbol{\mu})$ almost surely and in L_2 [Antos and Kontoyiannis, 2001a]. Silva [2018] studied a variety of restrictions on distributions over infinite alphabets to derive strong consistency results and rates of convergence.

Entropy estimation. Recent surveys of entropy estimation results may be found in Jiao et al. [2015], Verdú [2019]. The finite-alphabet case is particularly well-understood. For fixed alphabet size $d < \infty$, the plug-in estimate is asymptotically minimax optimal [Paninski, 2003]. Paninski [2004] non-constructively established the existence of a sublinear (in d) entropy estimator. The optimal dependence on d (at fixed accuracy) was was settled by Valiant and Valiant [2011a, 2017] as being $\Theta(d/\log d)$.

The $\Theta(d/\log d)$ dependence on the alphabet size is also relevant in the so-called *high dimensional* asymptotic regime, where d grows with n. Here, the plug-in estimate is no longer optimal, and more sophisticated techniques are called for [Valiant and Valiant, 2011a,b, 2017]. The works of Wu and Yang [2016], Jiao et al. [2015], Han et al. [2015], Jiao et al. [2017] characterized the minimax rates for the high-dimensional regime: a small additive error of ε requires $\Theta(d/\varepsilon \log d)$ samples.

Finally, there is the infinite-alphabet case. Although here the plug-in estimate is again universally strongly consistent, control of the convergence rate requires some assumption on the sampling distribution — and Antos and Kontoyiannis [2001a] compellingly argue that moment assumptions are natural and minimalistic. Absent any prior assumptions, the L_1 (and hence L_2) convergence rate of *any* estimator can be made arbitrarily slow (Theorem 4 ibid.). The present paper proves a variant of this result (see Theorem 3 and the Remark following it). Antos and Kontoyiannis [2001a] further show that even under moment assumptions, there is no polynomial rate of convergence for the plug-in estimate: there is no $\beta > 0$ such that its risk decays as $O(n^{-\beta})$. Wyner and Foster [2003] showed that the plug-in estimate achieves a rate of $O(\frac{1}{\log n})$ for bounded second moment, and this is minimax optimal. Brautbar and Samorodnitsky [2007] exhibited a function of the higher moments that can be used in place of alphabet size to give a multiplicative approximation to the entropy.

5 Proofs

5.1 Proof of Theorem 1

We begin with a subadditivity result for the αth moment of information (which we state for $\alpha > 0$, even though only the range $\alpha > 1$ will be needed).

Lemma 1. *For $\alpha > 0$ and $\boldsymbol{\mu}, \boldsymbol{\nu} \in \Delta_{\mathbb{N}}^{(\alpha)}$, we have*

$$\mathrm{H}^{(\alpha)}(|\boldsymbol{\mu} - \boldsymbol{\nu}|) \leq 2\alpha^{\alpha} + \mathrm{H}^{(\alpha)}(\boldsymbol{\mu}) + \mathrm{H}^{(\alpha)}(\boldsymbol{\nu}).$$

Proof. Define $\mathrm{h}^{(\alpha)} \colon [0,1] \to \mathbb{R}_+$ by $z \to z \ln^{\alpha}(1/z)$, where $\mathrm{h}^{(\alpha)}(0) = 0$. The function $\mathrm{h}^{(\alpha)}$ is increasing on $[0, \mathrm{e}^{-\alpha}]$ and decreasing on $[\mathrm{e}^{-\alpha}, 1]$. The maximum is therefore achieved at $z = \mathrm{e}^{-\alpha}$, and

$$\max_{z \in [0,1]} \mathrm{h}^{(\alpha)}(z) = \mathrm{h}^{(\alpha)}(\mathrm{e}^{-\alpha}) = \mathrm{e}^{-\alpha}\alpha^{\alpha}. \tag{5}$$

Now decompose $\mathrm{H}^{(\alpha)}$:

$$\mathrm{H}^{(\alpha)}(|\boldsymbol{\mu} - \boldsymbol{\nu}|) = \sum_{i : \boldsymbol{\mu}(i) \vee \boldsymbol{\nu}(i) > \mathrm{e}^{-\alpha}} \mathrm{h}^{(\alpha)}(|\boldsymbol{\mu}(i) - \boldsymbol{\nu}(i)|) + \sum_{i : \boldsymbol{\mu}(i) \vee \boldsymbol{\nu}(i) \leq \mathrm{e}^{-\alpha}} \mathrm{h}^{(\alpha)}(|\boldsymbol{\mu}(i) - \boldsymbol{\nu}(i)|).$$

For the first term, since $\boldsymbol{\mu} \in \Delta_{\mathbb{N}}$, it must be that $|\{i \in \mathbb{N} \colon \boldsymbol{\mu}(i) > \mathrm{e}^{-\alpha}\}| \leq \mathrm{e}^{\alpha}$, and similarly for $\boldsymbol{\nu}$. Thus,

$$\begin{aligned}
\sum_{i : \boldsymbol{\mu}(i) \vee \boldsymbol{\nu}(i) > \mathrm{e}^{-\alpha}} \mathrm{h}^{(\alpha)}(|\boldsymbol{\mu}(i) - \boldsymbol{\nu}(i)|) &\leq \left(\left|\left\{i \colon \boldsymbol{\mu}(i) > \mathrm{e}^{-\alpha}\right\}\right| + \left|\left\{i \colon \boldsymbol{\nu}(i) > \mathrm{e}^{-\alpha}\right\}\right| \right) \max_{z \in [0,1]} \mathrm{h}^{(\alpha)}(z) \\
&\leq 2\mathrm{e}^{\alpha}\mathrm{e}^{-\alpha}\alpha^{\alpha} = 2\alpha^{\alpha}.
\end{aligned}$$

For the second term, notice that when $\boldsymbol{\mu}(i) \vee \boldsymbol{\nu}(i) \leq \mathrm{e}^{-\alpha}$, the monotonicity of $\mathrm{h}^{(\alpha)}$ implies

$$\mathrm{h}^{(\alpha)}(|\boldsymbol{\mu}(i) - \boldsymbol{\nu}(i)|) \leq \mathrm{h}^{(\alpha)}(\boldsymbol{\mu}(i) \vee \boldsymbol{\nu}(i)),$$

and hence

$$\begin{aligned}
\sum_{i \in \mathbb{N} : \boldsymbol{\mu}(i) \vee \boldsymbol{\nu}(i) \leq \mathrm{e}^{-\alpha}} \mathrm{h}^{(\alpha)}(|\boldsymbol{\mu}(i) - \boldsymbol{\nu}(i)|) &\leq \sum_{i : \in \boldsymbol{\mu}(i) \vee \boldsymbol{\nu}(i) \leq \mathrm{e}^{-\alpha}} \mathrm{h}^{(\alpha)}(\boldsymbol{\mu}(i) \vee \boldsymbol{\nu}(i)) \\
&\leq \sum_{i : \in \boldsymbol{\mu}(i) \vee \boldsymbol{\nu}(i) \leq \mathrm{e}^{-\alpha}} \mathrm{h}^{(\alpha)}(\boldsymbol{\mu}(i)) + \mathrm{h}^{(\alpha)}(\boldsymbol{\nu}(i)) \\
&\leq \mathrm{H}^{(\alpha)}(\boldsymbol{\mu}) + \mathrm{H}^{(\alpha)}(\boldsymbol{\nu}).
\end{aligned}$$

□

Proof of Theorem 1. The concavity argument in the proof of Cover and Thomas [2006, Theorem 17.3.3], immediately implies $|\mathrm{H}(\boldsymbol{\mu}) - \mathrm{H}(\boldsymbol{\nu})| \leq \mathrm{H}(|\boldsymbol{\mu} - \boldsymbol{\nu}|)$. Then, via an application of Hölder's inequality,

$$\begin{aligned}
\mathrm{H}(|\boldsymbol{\mu} - \boldsymbol{\nu}|) &= \sum_{i \in \mathbb{N}} |\boldsymbol{\mu}(i) - \boldsymbol{\nu}(i)| \log \frac{1}{|\boldsymbol{\mu}(i) - \boldsymbol{\nu}(i)|} = \sum_{i \in \mathbb{N}} |\boldsymbol{\mu}(i) - \boldsymbol{\nu}(i)|^{1-1/\alpha} \cdot |\boldsymbol{\mu}(i) - \boldsymbol{\nu}(i)|^{1/\alpha} \log \frac{1}{|\boldsymbol{\mu}(i) - \boldsymbol{\nu}(i)|} \\
&\leq \left(\sum_{i \in \mathbb{N}} \left(|\boldsymbol{\mu}(i) - \boldsymbol{\nu}(i)|^{1-1/\alpha} \right)^{1/(1-1/\alpha)} \right)^{1-1/\alpha} \left(\sum_{i \in \mathbb{N}} \left(|\boldsymbol{\mu}(i) - \boldsymbol{\nu}(i)|^{1/\alpha} \log \frac{1}{|\boldsymbol{\mu}(i) - \boldsymbol{\nu}(i)|} \right)^{\alpha} \right)^{1/\alpha} \\
&= \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_1^{1-1/\alpha} \, \mathrm{H}^{(\alpha)}(|\boldsymbol{\mu} - \boldsymbol{\nu}|)^{1/\alpha}.
\end{aligned}$$

The claim follows by invoking Lemma 1 and the subadditivity of $t \mapsto t^{1/\alpha}$ for $t \geq 0$ and $\alpha > 1$. □

5.2 Proof of Theorem 3

For $h > 1$ and $n \in \mathbb{N}$, put $a_n = (1-1/(2n))\ln(1-1/(2n))$ and define the support size $S = S(h,n)$ by $S = \lfloor (1/2n)\exp(2n(h+a_n))\rfloor$. Consider the distributions $\boldsymbol{\mu}_0 = (1,0,0,\dots)$ and $\boldsymbol{\mu}_n$ defined by $\boldsymbol{\mu}_n(1) = 1-1/(2n)$, and $\boldsymbol{\mu}_n(i) = \frac{1}{2nS}$ for $2 \le i \le 1+S(h,n)$. We compute the Kullback-Leibler divergence and entropy:

$$D_{\mathrm{KL}}\left(\boldsymbol{\mu}_0 || \boldsymbol{\mu}_n\right) = \log\frac{1}{1-1/(2n)} \le \frac{1}{1-1/(2n)} - 1 \le \frac{1}{n}; \qquad \mathrm{H}(\boldsymbol{\mu}_0) \;=\; 0 \le h. \tag{6}$$

For $x \ge 2$, always $\lfloor x \rfloor \ge x/2$. Additionally, from $2na_n \ge -1$, and $\frac{1}{2n}\exp(2nh-1) > 2$, we obtain that $S > (1/4n)\exp(2n(h+a_n))$, hence we also have that $h \ge \mathrm{H}(\boldsymbol{\mu}_n) > h - \frac{1}{2n}\ln 2$. Since $\frac{1}{2x}\ln 2 \le 1/2$ on $(0,\infty)$ and $h > 1$, it follows that $\mathrm{H}(\boldsymbol{\mu}_n) \ge \frac{h}{2}$, whence $|\mathrm{H}(\boldsymbol{\mu}_0) - \mathrm{H}(\boldsymbol{\mu}_n)| \ge h/2$. To bound the L_1 minimax risk (defined in (4)), we invoke Markov's inequality:

$$\mathbb{E}|\hat{H}(X_1,\dots,X_n) - \mathrm{H}(\boldsymbol{\mu})| \;\ge\; \frac{h}{4}\,\mathbb{P}\left(|\hat{H}(X_1,\dots,X_n) - \mathrm{H}(\boldsymbol{\mu})| > \frac{h}{4}\right).$$

It follows via Le Cam's two point method [Tsybakov, 2008, Section 2.4.2] that $\mathcal{R}_n^{(1)}(h) \ge \frac{h}{4}\mathrm{e}^{-nD_{\mathrm{KL}}(\boldsymbol{\mu}_0||\boldsymbol{\mu})} \ge \frac{h}{4\mathrm{e}}$, where the second inequality stems from (6). □

5.3 Proof of Theorem 4

We begin with an auxiliary lemma, of possible independent interest.

Lemma 2. *For all $\boldsymbol{\mu} \in \Delta_{\mathbb{N}}$ and $n \in \mathbb{N}$, we have*

$$\mathrm{H}(\boldsymbol{\mu}) \;\ge\; \mathbb{E}\mathrm{H}(\hat{\boldsymbol{\mu}}_n) \;\ge\; \mathrm{H}(\boldsymbol{\mu}) - \sup_{0<\varepsilon<1}\left[\sum_{i\in\mathbb{N}:\boldsymbol{\mu}(i)<\varepsilon}\boldsymbol{\mu}(i)\log\frac{1}{\boldsymbol{\mu}(i)} + \log\left(1+\frac{1}{\varepsilon n}\right)\right].$$

Proof. The first inequality follows from Jensen's, since $\mathrm{H}(\cdot)$ is concave and $\mathbb{E}\hat{\boldsymbol{\mu}}_n = \boldsymbol{\mu}$. To prove the second inequality, choose $\varepsilon > 0$, put $J := \{i \in \mathbb{N} : \boldsymbol{\mu}(i) < \varepsilon\}$, and compute

$$\begin{aligned}
\mathbb{E}\mathrm{H}(\hat{\boldsymbol{\mu}}_n) \;&=\; \mathbb{E}\left[\sum_{i\in\mathbb{N}\setminus J}\hat{\boldsymbol{\mu}}_n(i)\log\frac{1}{\hat{\boldsymbol{\mu}}_n(i)} + \sum_{i\in J}\hat{\boldsymbol{\mu}}_n(i)\log\frac{1}{\hat{\boldsymbol{\mu}}_n(i)}\right] \\
&\ge\; \mathbb{E}\left[\sum_{i\in\mathbb{N}\setminus J}\hat{\boldsymbol{\mu}}_n(i)\log\frac{1}{\hat{\boldsymbol{\mu}}_n(i)} + \left(\sum_{i\in J}\hat{\boldsymbol{\mu}}_n(i)\right)\log\frac{1}{\sum_{i\in J}\hat{\boldsymbol{\mu}}_n(i)}\right] =: \mathbb{E}\mathrm{H}(\tilde{\boldsymbol{\mu}}_n),
\end{aligned}$$

where $\tilde{\boldsymbol{\mu}}_n$ is the "collapsed" version of $\hat{\boldsymbol{\mu}}_n$, where all of the masses in J have been replaced by a single mass equal to their sum, and the inequality follows from the data processing inequality (any transformation of the distribution can only decrease the entropy). We observe that $\tilde{\boldsymbol{\mu}}_n$ has support size at most $1 + 1/\varepsilon$ and invoke [Paninski, 2003, Proposition 1]:

$$\mathbb{E}\mathrm{H}(\tilde{\boldsymbol{\mu}}_n) \;\ge\; \mathrm{H}(\tilde{\boldsymbol{\mu}}) - \log\left(1+\frac{1}{\varepsilon n}\right), \tag{7}$$

where $\tilde{\boldsymbol{\mu}}$ is the "collapsed" version of $\boldsymbol{\mu}$. Now

$$\begin{aligned}
\mathrm{H}(\tilde{\boldsymbol{\mu}}) \;&=\; \mathrm{H}(\boldsymbol{\mu}) + \left(\sum_{\boldsymbol{\mu}(i)<\varepsilon}\boldsymbol{\mu}(i)\right)\log\frac{1}{\sum_{\boldsymbol{\mu}(i)<\varepsilon}\boldsymbol{\mu}(i)} - \sum_{i:\boldsymbol{\mu}(i)<\varepsilon}\boldsymbol{\mu}(i)\log\frac{1}{\boldsymbol{\mu}(i)} \\
&\ge\; \mathrm{H}(\boldsymbol{\mu}) - \sum_{i:\boldsymbol{\mu}(i)<\varepsilon}\boldsymbol{\mu}(i)\log\frac{1}{\boldsymbol{\mu}(i)},
\end{aligned}$$

which concludes the proof. □

The first part of the theorem will follow from the following proposition.

Proposition 1. *For $\alpha \geq 1$, $h > 0$, $n \in \mathbb{N}$ and $\boldsymbol{\mu} \in \Delta_{\mathbb{N}}^{(\alpha)}[h]$, we have*

$$\mathbb{E}|\mathrm{H}(\boldsymbol{\mu}) - \mathrm{H}(\hat{\boldsymbol{\mu}}_n)| \quad \leq \quad \frac{\log n}{\sqrt{n}} + \inf_{0<\varepsilon\leq 1}\left[\left(\log\frac{1}{\varepsilon}\right)^{1-\alpha} h + \log\left(1+\frac{1}{\varepsilon n}\right)\right].$$

Proof. Since $|\mathrm{H}(\boldsymbol{\mu}) - \mathbb{E}\mathrm{H}(\hat{\boldsymbol{\mu}}_n)| = \mathrm{H}(\boldsymbol{\mu}) - \mathbb{E}\mathrm{H}(\hat{\boldsymbol{\mu}}_n)$ (Lemma 2), it follows from the triangle and Jensen inequalities that

$$\begin{aligned}
\mathbb{E}|\mathrm{H}(\boldsymbol{\mu}) - \mathrm{H}(\hat{\boldsymbol{\mu}}_n)| &\leq \mathbb{E}|\mathrm{H}(\hat{\boldsymbol{\mu}}_n) - \mathbb{E}\mathrm{H}(\hat{\boldsymbol{\mu}}_n)| + \mathrm{H}(\boldsymbol{\mu}) - \mathbb{E}\mathrm{H}(\hat{\boldsymbol{\mu}}_n) \\
&\leq \sqrt{\mathbb{V}\mathbf{ar}\left[\mathrm{H}(\hat{\boldsymbol{\mu}}_n)\right]} + \mathrm{H}(\boldsymbol{\mu}) - \mathbb{E}\mathrm{H}(\hat{\boldsymbol{\mu}}_n) \\
&\leq \frac{\log n}{\sqrt{n}} + \mathrm{H}(\boldsymbol{\mu}) - \mathbb{E}\mathrm{H}(\hat{\boldsymbol{\mu}}_n), \qquad (8)
\end{aligned}$$

where the variance bound is from Antos and Kontoyiannis [2001b, Proposition 1(iv)]. For any $\varepsilon > 0$, Lemma 2 implies

$$\begin{aligned}
\mathbb{E}\mathrm{H}(\hat{\boldsymbol{\mu}}_n) &\geq \mathrm{H}(\boldsymbol{\mu}) - \sum_{i\in\mathbb{N}:\boldsymbol{\mu}(i)<\varepsilon} \boldsymbol{\mu}(i)\log\frac{1}{\boldsymbol{\mu}(i)} - \log\left(1+\frac{1}{\varepsilon n}\right) \\
&\geq \mathrm{H}(\boldsymbol{\mu}) - \left(\log\frac{1}{\varepsilon}\right)^{1-\alpha} \sum_{i\in\mathbb{N}:\boldsymbol{\mu}(i)<\varepsilon} \boldsymbol{\mu}(i)\left(\log\frac{1}{\boldsymbol{\mu}(i)}\right)^{\alpha} - \log\left(1+\frac{1}{\varepsilon n}\right) \\
&\geq \mathrm{H}(\boldsymbol{\mu}) - \left(\log\frac{1}{\varepsilon}\right)^{1-\alpha} \mathrm{H}^{(\alpha)}(\boldsymbol{\mu}) - \log\left(1+\frac{1}{\varepsilon n}\right), \qquad (9)
\end{aligned}$$

where the second and third inequalities follow from the obvious relations

$$\sum_{i:\boldsymbol{\mu}(i)<\varepsilon} \boldsymbol{\mu}(i)\log\frac{1}{\boldsymbol{\mu}(i)} \leq \left(\log\frac{1}{\varepsilon}\right)^{1-\alpha} \sum_{i:\boldsymbol{\mu}(i)<\varepsilon} \boldsymbol{\mu}(i)\left(\log\frac{1}{\boldsymbol{\mu}(i)}\right)^{\alpha} \leq \left(\log\frac{1}{\varepsilon}\right)^{1-\alpha} \mathrm{H}^{(\alpha)}(\boldsymbol{\mu}).$$

The claim follows by combining (8) with (9). □

Proof of Theorem 4(a). Invoke Proposition 1 with $\varepsilon = \frac{1}{\sqrt{n}}$ and use $\log(1+x) \leq x$. □

Proof of Theorem 4(b). Let $\alpha > 0$, $n \in \mathbb{N}$ and define two families of distributions:

$$\mathcal{U}_1 := \left\{\boldsymbol{\mu}_1 = \mathrm{Uniform}([n^3])\right\}, \quad \mathcal{U}_2 := \left\{\boldsymbol{\mu}_2 = \mathrm{Uniform}(A) : A \subset [n^3], |A| = n^2\right\}.$$

Let $h := 3^\alpha \log^\alpha n$ and note that $\mathcal{U}_1 \cup \mathcal{U}_2 \subseteq \Delta_{\mathbb{N}}^{(\alpha)}[h]$. Let E be the event that $\boldsymbol{X} = (X_1, \ldots, X_n)$ has no repeating elements, i.e $|\{X_1, X_2, \ldots, X_n\}| = n$. Let $\boldsymbol{\mu}_1 \in \mathcal{U}_1, \boldsymbol{\mu}_2 \in \mathcal{U}_2$ and consider the values $\mathbb{P}_{\boldsymbol{X}\sim\boldsymbol{\mu}_1^n}(E)$ and $\mathbb{P}_{\boldsymbol{X}\sim\boldsymbol{\mu}_2^n}(E)$. For $m \in \mathbb{N}$, define $\mathcal{X}(m)$ to be the smallest k such that when uniformly throwing m balls into k buckets, the probability of collision is at least $1/2$. Since $\mathcal{X}(m)$ is known[4] to be at least $\sqrt{m}$ (and hence $\mathcal{X}(n^2) > n$) we have a lower bound of $\frac{1}{2}$ on both $\mathbb{P}_{\boldsymbol{X}\sim\boldsymbol{\mu}_1^n}(E)$ and $\mathbb{P}_{\boldsymbol{X}\sim\boldsymbol{\mu}_2^n}(E)$. Define $\boldsymbol{\mu}_1^n|E$ as the distribution on $\mathbb{N}^n$ induced by conditioning the product $\boldsymbol{\mu}_1^n$ on the event E, and define $\boldsymbol{\mu}_2^n|E$ analogously. Our key observation is that conditional on E, (i) both are effectively distributions on ordered n-tuples from $[n^3]$, and (ii) $\boldsymbol{\mu}_1^n$ is uniform on $([n^3])_n$ whereas $\boldsymbol{\mu}_2^n = \mathrm{Uniform}(A)$ is uniform on $(A)_n$, where

[4] Better bounds exist [Brink, 2012].

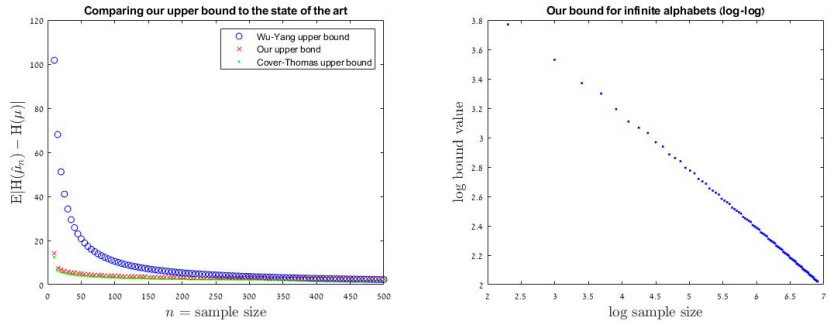

Figure 1: Left: A comparison of the three bounds for $d = 10$, $D = 1000$, $p = 0.95$. Our bound considerably outperforms Wu and Yang [2016] on small samples, and performs nearly as well as the finite-dimensional Cover-Thomas bound. Right: for our value of $q = 2$, the log-log plot shows roughly the correct slope of $-1/2$.

$(J)_k := \left\{(x_1, \ldots, x_k) \in J^k : |\{x_1, \ldots, x_k\}| = k\right\}, \qquad J \subset \mathbb{N}, k \in \mathbb{N}$. Then

$$
\begin{aligned}
\mathcal{R}_n^{(\alpha)}(h) &\geq \inf_{\hat{H}} \sup_{\boldsymbol{\mu} \in \mathcal{U}_1 \cup \mathcal{U}_2} \mathop{\mathbb{E}}_{\boldsymbol{X} \sim \boldsymbol{\mu}^n} \left[|\hat{H}(\boldsymbol{X}) - \mathrm{H}(\boldsymbol{\mu})| \right] \\
&\overset{\text{(a)}}{\geq} \inf_{\hat{H}} \sup_{\boldsymbol{\mu} \in \mathcal{U}_1 \cup \mathcal{U}_2} \mathop{\mathbb{E}}_{\boldsymbol{X} \sim \boldsymbol{\mu}^n | E} \left[|\hat{H}(\boldsymbol{X}) - \mathrm{H}(\boldsymbol{\mu})| \right] \mathop{\mathbb{P}}_{\boldsymbol{X} \sim \boldsymbol{\mu}^n} (E) \\
&\geq \inf_{\hat{H}} \frac{1}{2} \sup_{\boldsymbol{\mu} \in \mathcal{U}_1 \cup \mathcal{U}_2} \mathop{\mathbb{E}}_{\boldsymbol{X} \sim \boldsymbol{\mu}^n | E} \left[|\hat{H}(\boldsymbol{X}) - \mathrm{H}(\boldsymbol{\mu})| \right] \\
&\overset{\text{(b)}}{\geq} \inf_{\hat{H}} \frac{1}{4} \left(\mathop{\mathbb{E}}_{\boldsymbol{X} \sim \boldsymbol{\mu}_1^n | E} \left[|\hat{H}(\boldsymbol{X}) - \mathrm{H}(\boldsymbol{\mu}_1)| \right] + \sup_{\boldsymbol{\mu}_2 \in \mathcal{U}_2} \mathop{\mathbb{E}}_{\boldsymbol{X} \sim \boldsymbol{\mu}_2^n | E} \left[|\hat{H}(\boldsymbol{X}) - \mathrm{H}(\boldsymbol{\mu}_2)| \right] \right) \\
&\overset{\text{(c)}}{\geq} \inf_{\hat{H}} \frac{1}{4} \left(\mathop{\mathbb{E}}_{\boldsymbol{X} \sim \boldsymbol{\mu}_1^n | E} \left[|\hat{H}(\boldsymbol{X}) - \mathrm{H}(\boldsymbol{\mu}_1)| \right] + \mathop{\mathbb{E}}_{\boldsymbol{\mu}_2 \sim \mathrm{Uniform}(\mathcal{U}_2)} \left[\mathop{\mathbb{E}}_{\boldsymbol{X} \sim \boldsymbol{\mu}_2^n | E} \left[|\hat{H}(\boldsymbol{X}) - \mathrm{H}(\boldsymbol{\mu}_2)| \right] \right] \right) \\
&\overset{\text{(d)}}{=} \inf_{\hat{H}} \frac{1}{4} \left(\mathop{\mathbb{E}}_{\boldsymbol{X} \sim \boldsymbol{\mu}_1^n | E} \left[|\hat{H}(\boldsymbol{X}) - \mathrm{H}(\boldsymbol{\mu}_1)| \right] + \mathop{\mathbb{E}}_{\boldsymbol{X} \sim \boldsymbol{\mu}_1^n | E} \left[|\hat{H}(\boldsymbol{X}) - \mathrm{H}(\boldsymbol{\mu}_2)| \right] \right) \\
&= \inf_{\hat{H}} \frac{1}{4} \left(\mathop{\mathbb{E}}_{\boldsymbol{X} \sim \boldsymbol{\mu}_1^n | E} \left[|\hat{H}(\boldsymbol{X}) - \mathrm{H}(\boldsymbol{\mu}_1)| + |\hat{H}(\boldsymbol{X}) - \mathrm{H}(\boldsymbol{\mu}_2)| \right] \right) \\
&\overset{\text{(e)}}{\geq} \frac{1}{4} |\mathrm{H}(\boldsymbol{\mu}_1) - \mathrm{H}(\boldsymbol{\mu}_2)| = \frac{1}{4} \log n = \frac{1}{4} \frac{h}{3^\alpha \log^{\alpha-1} n},
\end{aligned}
$$

where (a) is from the law of total expectation (the complement of E is discarded), (b) and (c) are bounding a supremum by an average, (e) is from the triangle inequality, and (d) is by symmetry, the operators $\mathbb{E}_{\boldsymbol{\mu}_2 \sim \mathrm{Uniform}(\mathcal{U}_2)} \left[\mathbb{E}_{\boldsymbol{X} \sim \boldsymbol{\mu}_2^n | E} [\cdot] \right]$ and $\mathbb{E}_{\boldsymbol{X} \sim \boldsymbol{\mu}_1^n | E} [\cdot]$ are equivalent. (There is a minor abuse of notation in transitions after (c), since we write $\boldsymbol{\mu}_2$ without specifying a *particular* member of $\mathcal{U}_2$. However, $\boldsymbol{\mu}_2$ only occurs therein as $\mathrm{H}(\boldsymbol{\mu}_2)$, and this value is identical for all $\boldsymbol{\mu}_2 \in \mathcal{U}_2$.) □

6 Rates

Our bounds have the crucial characteristic of being empirical. When we *observe* favorable distributions (even without a priori knowledge of the fact), we will benefit from tighter bounds. There is some cost which this naturally entails as well, and in the worst case our bounds will be sub-optimal. In this section, we illustrate these trade-offs for various natural classes of distributions.

For the class of all finite alphabet distributions, our bound is sub-optimal. The MLE is competitive with the optimal estimator up to logarithmic factors in d, but our bounds on the MLE are loose nearly quadratically in d/n, in the worst case. The convergence of the empirical distribution on a

finite alphabet in ℓ_1 occurs at rate $\Theta(\sqrt{d/n})$, whereas the MLE entropy estimator converges at rate $O\left(\sqrt{\left(\frac{d}{n}\right)^2 + \frac{\log^2 d}{\sqrt{n}}}\right)$, as follows from Wu and Yang [2016, Proposition 1]. So any approach that upper bounds the entropy risk via ℓ_1 (as our Theorem 1 or Section 4 of Ho and Yeung [2010]) will be worst-case suboptimal for this class of distributions.

Nevertheless, for certain classes of distributions our bounds can significantly outperform the state of the art, for small and moderate-sized samples. To calculate the expected rate of our approach, we apply Hölder's inequality, as in the proof of Theorem 1:

$$\mathbb{E}|\mathrm{H}(\hat{\boldsymbol{\mu}}_n) - \mathrm{H}(\boldsymbol{\mu})| \quad \leq \quad \left(\mathbb{E}\left[2\alpha^\alpha + \mathrm{H}^{(\alpha)}(\boldsymbol{\mu}) + \mathrm{H}^{(\alpha)}(\hat{\boldsymbol{\mu}}_n)\right]\right)^{1/\alpha} \left(\mathbb{E}\|\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}\|_1\right)^{1-1/\alpha}.$$

Now, as in the proof of Lemma 1 (recall that $\mathrm{h}^{(\alpha)}(z) := z\ln^\alpha(1/z)$) ,

$$\begin{aligned}
\mathbb{E}\mathrm{H}^{(\alpha)}(\hat{\boldsymbol{\mu}}_n) &= \sum_{i\in[d]} \mathbb{E}\mathrm{h}^{(\alpha)}(\hat{\boldsymbol{\mu}}_n(i)) \\
&\leq e^{\alpha-1} \max_{z\in[0,e^{1-\alpha}]} \mathrm{h}^{(\alpha)}(z) + \sum_{\substack{i\in[d] \\ \hat{\boldsymbol{\mu}}_n(i)<e^{1-\alpha}}} \mathbb{E}\mathrm{h}^{(\alpha)}(\hat{\boldsymbol{\mu}}_n(i)) \\
&\overset{(i)}{\leq} e^{\alpha-1} \max_{z\in[0,e^{1-\alpha}]} \mathrm{h}^{(\alpha)}(z) + \mathrm{H}^{(\alpha)}(\boldsymbol{\mu}) \overset{(ii)}{\leq} \frac{\alpha^\alpha}{e} + \mathrm{H}^{(\alpha)}(\boldsymbol{\mu}),
\end{aligned}$$

where (i) follows from Jensen's inequality and (ii) from (5).

By Berend and Kontorovich [2013, Lemma 6], we have $\mathbb{E}\|\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}\|_1 \leq \Lambda_n(\boldsymbol{\mu})$, where $\Lambda_n(\boldsymbol{\mu}) := 2\sum_{\boldsymbol{\mu}(j)<1/n} \boldsymbol{\mu}(j) + \frac{1}{\sqrt{n}} \sum_{\boldsymbol{\mu}(j)\geq 1/n} \sqrt{\boldsymbol{\mu}(j)}$. This quantity is always finite and $\Lambda_n(\boldsymbol{\mu}) \underset{n\to\infty}{\longrightarrow} 0$ for all $\boldsymbol{\mu} \in \Delta_{\mathbb{N}}$ (ibid). Thus, we obtain the bound

$$\mathbb{E}|\mathrm{H}(\hat{\boldsymbol{\mu}}_n) - \mathrm{H}(\boldsymbol{\mu})| \quad \leq \quad \left(\frac{\alpha^\alpha}{e} + 2\alpha^\alpha + 2\mathrm{H}^{(\alpha)}(\boldsymbol{\mu})\right)^{1/\alpha} \Lambda_n(\boldsymbol{\mu})^{1-1/\alpha}. \tag{10}$$

Finite support. For distributions with a large support but concentrated mass, the bound in (10) compares favorably to the state of the art, especially for smaller sample sizes. To illustrate this, consider a mixture of two distributions with support sizes d and D: $\boldsymbol{\mu}'$ is uniform over $[d]$, $\boldsymbol{\mu}''$ is uniform over $d + [D]$, and $\boldsymbol{\mu} := p\boldsymbol{\mu}' + (1-p)\boldsymbol{\mu}''$, for some $p \in [0,1]$.

The state-of-the-art upper bound for the plug-in estimator can be inferred from Wu and Yang [2016, Appendix D], and has the form

$$\mathbb{E}|\mathrm{H}(\hat{\boldsymbol{\mu}}_n) - \mathrm{H}(\boldsymbol{\mu})| \leq \mathrm{WY}(d,D,p,n) := \frac{d+D}{n} + \min\left(C\frac{\log(d+D)}{\sqrt{n}}, \frac{\log n}{\sqrt{n}}\right)$$

for some $C > 1$; notice that it is insensitive to p. For a fair comparison to (10), our estimator's only a priori knowledge of $\boldsymbol{\mu}$ is that its support is of size at most $d + D$. By Proposition 2, we have $\max_{\boldsymbol{\mu}\in\Delta_K} \mathrm{H}^{(\alpha)}(\boldsymbol{\mu}) \leq \max\{\alpha, \log K\}^\alpha + (\alpha/\mathrm{e})^\alpha$. This allows us to optimize over α for each n:

$$\mathrm{OUR}(d,D,p,n) := \inf_{\alpha>1} \left(\frac{\alpha^\alpha}{e} + 2\alpha^\alpha + 2\max\{\alpha, \log(d+D)\}^\alpha + 2(\alpha/\mathrm{e})^\alpha\right)^{1/\alpha} \Lambda_n(\boldsymbol{\mu})^{1-1/\alpha}.$$

Since $\boldsymbol{\mu}$ has finite support, the Cover-Thomas inequality (2) also applies to yield an adaptive estimate when combined with (1). As $t\log(1/t)$ is convex, the latter has the form

$$\mathbb{E}|\mathrm{H}(\hat{\boldsymbol{\mu}}_n) - \mathrm{H}(\boldsymbol{\mu})| \leq \mathbb{E}\left[\|\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}\|_1 \log\frac{d+D}{\|\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}\|_1}\right] \leq \Lambda_n(\boldsymbol{\mu}) \log\frac{d+D}{\Lambda_n(\boldsymbol{\mu})} =: \mathrm{CT}(d,D,p,n).$$

Infinite support. In some cases our bound is nearly tight (at least for the plug-in estimate), such as for the family of zeta distributions $\boldsymbol{\mu}_q(i) \sim 1/i^q$ with parameter $q > 1$. For this family, Antos and Kontoyiannis [2001a, Theorem 7] establish a lower bound of order $n^{\frac{1-q}{q}}$ on $\mathbb{E}\left|\mathrm{H}(\hat{\boldsymbol{\mu}}_n) - \mathrm{H}(\boldsymbol{\mu}_q)\right|$. It is straightforward to verify[5] that $\boldsymbol{\mu}_q \in \Delta_{\mathbb{N}}^{(\alpha)}$ for all $q, \alpha > 1$. Thus, we can optimize our bound in (10) over all $\alpha > 1$; the results are presented in Figure 1.

[5] One can, for example, apply Cauchy's condensation test, followed up by the ratio test.

References

A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, 2001a.

A. Antos and I. Kontoyiannis. Estimating the entropy of discrete distributions. In *IEEE International Symposium on Information Theory*, pages 45–45, 2001b.

K. M. R. Audenaert. A sharp continuity estimate for the von Neumann entropy. *J. Phys. A*, 40 (28):8127–8136, 2007. ISSN 1751-8113. doi: 10.1088/1751-8113/40/28/S18. URL https://doi.org/10.1088/1751-8113/40/28/S18.

D. Berend and A. Kontorovich. A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters*, 83(4):1254–1259, 2013. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2013.01.023. URL https://www.sciencedirect.com/science/article/pii/S0167715213000242.

D. Berend, A. Kontorovich, and G. Zagdanski. The expected missing mass under an entropy constraint. *Entropy*, 19(7):315, 2017. doi: 10.3390/e19070315. URL https://doi.org/10.3390/e19070315.

M. Brautbar and A. Samorodnitsky. Approximating entropy from sublinear samples. In N. Bansal, K. Pruhs, and C. Stein, editors, *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2007, New Orleans, Louisiana, USA, January 7-9, 2007*, pages 366–375. SIAM, 2007. URL http://dl.acm.org/citation.cfm?id=1283383.1283422.

D. Brink. A (probably) exact solution to the birthday problem. *The Ramanujan Journal*, 28, 06 2012. doi: 10.1007/s11139-011-9343-9.

D. Cohen, A. Kontorovich, and G. Wolfer. Learning discrete distributions with infinite support. In *Neural Information Processing Systems (NIPS)*, 2020.

T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2006. ISBN 978-0-471-24195-9; 0-471-24195-4.

Y. Han, J. Jiao, and T. Weissman. Adaptive estimation of shannon entropy. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 1372–1376. IEEE, 2015.

S.-W. Ho and R. W. Yeung. The interplay between entropy and variational distance. *IEEE Transactions on Information Theory*, 56(12):5906–5929, 2010.

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

J. Jiao, K. Venkat, Y. Han, and T. Weissman. Maximum likelihood estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 63(10):6774–6798, 2017. doi: 10.1109/TIT.2017.2733537.

N. Kusolitsch. Why the theorem of Scheffé should be rather called a theorem of Riesz. *Period. Math. Hungar.*, 61(1-2):225–229, 2010. ISSN 0031-5303. doi: 10.1007/s10998-010-3225-6. URL https://doi.org/10.1007/s10998-010-3225-6.

E. H. Lieb and M. Loss. *Analysis*, volume 14 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2001. ISBN 0-8218-2783-9. doi: 10.1090/gsm/014. URL https://doi.org/10.1090/gsm/014.

P.-S. Loh. Convexity. 2013. URL https://www.math.cmu.edu/~ploh/docs/math/mop2008/convexity-soln.pdf.

G. Lugosi and S. Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Found. Comput. Math.*, 19(5):1145–1190, 2019. doi: 10.1007/s10208-019-09427-x. URL https://doi.org/10.1007/s10208-019-09427-x.

P. Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283, 1990. ISSN 0091-1798. URL http://links.jstor.org/sici?sici=0091-1798(199007)18:3<1269:TTCITD>2.0.CO;2-Q&origin=MSN.

L. Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003. doi: 10.1162/089976603321780272.

L. Paninski. Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50(9):2200–2203, 2004. doi: 10.1109/TIT.2004.833360.

I. Pinelis. Best possible bounds of the von Bahr–Esseen type. *Annals of Functional Analysis*, 6(4):1 – 29, 2015. doi: 10.15352/afa/06-4-1. URL https://doi.org/10.15352/afa/06-4-1.

I. Sason. Entropy bounds for discrete random variables via maximal coupling. *IEEE Trans. Inf. Theory*, 59(11):7118–7131, 2013. doi: 10.1109/TIT.2013.2274515. URL https://doi.org/10.1109/TIT.2013.2274515.

H. Scheffé. A useful convergence theorem for probability distributions. *Ann. Math. Statistics*, 18: 434–438, 1947. ISSN 0003-4851. doi: 10.1214/aoms/1177730390. URL https://doi.org/10.1214/aoms/1177730390.

J. F. Silva. Shannon entropy estimation in ∞-alphabets from convergence results: studying plug-in estimators. *Entropy*, 20(6):397, 2018.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

G. Valiant and P. Valiant. Estimating the unseen: An n/log(n)-sample estimator for entropy and support size, shown optimal via new clts. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, STOC '11, page 685–694, New York, NY, USA, 2011a. Association for Computing Machinery. ISBN 9781450306911. doi: 10.1145/1993636.1993727. URL https://doi.org/10.1145/1993636.1993727.

G. Valiant and P. Valiant. The power of linear estimators. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pages 403–412, 2011b. doi: 10.1109/FOCS.2011.81.

G. Valiant and P. Valiant. Estimating the unseen: Improved estimators for entropy and other properties. *J. ACM*, 64(6), Oct. 2017. ISSN 0004-5411. doi: 10.1145/3125643. URL https://doi.org/10.1145/3125643.

S. Verdú. Empirical estimation of information measures: A literature guide. *Entropy*, 21(8):720, 2019.

B. von Bahr and C.-G. Esseen. Inequalities for the rth Absolute Moment of a Sum of Random Variables, $1 \leqq r \leqq 2$. *The Annals of Mathematical Statistics*, 36(1):299 – 303, 1965. doi: 10.1214/aoms/1177700291. URL https://doi.org/10.1214/aoms/1177700291.

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

A. J. Wyner and D. Foster. On the lower limits of entropy estimation. *IEEE Transactions on Information Theory, submitted for publication*, 2003.

Z. Zhang. Estimating mutual information via kolmogorov distance. *IEEE Transactions on Information Theory*, 53(9):3280–3282, 2007. doi: 10.1109/TIT.2007.903122.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 3.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Section 5.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]