Dimension-Free Empirical Entropy Estimation

Doron Cohen, Aryeh Kontorovich, Aaron Koolyk, and Geoffrey Wolfer

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. Additionally, we resolve all of the open problems posed by Jürgensen and Matthews, 2010.

Index Terms—Information, entropy, empirical estimation.

I. 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. (See Section IV for a some background, motivation, and related work.) 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 \( \mu \) in total variation norm \( \| \cdot \|_1 = \frac{1}{2} \| \cdot \|_1 \), where the most recent advance was made by [10]; see therein for a literature review. If \( \mu \) is a distribution on \( \mathbb{N} \) and \( \mu_n \) is its empirical realization based on a sample of size \( n \), then Theorem 2.1 of [10] states that with probability at least \( 1 - \delta \),

\[
\| \mu - \mu_n \|_1 \leq \frac{2}{\sqrt{n}} \sum_{j \in \mathbb{N}} \sqrt{\mu_n(j)} + 6 \sqrt{\frac{\log(2/\delta)}{2n}}.
\]

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, [10] 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 \( \mu \) has support size \( d < \infty \), an answer to our question is readily provided by combining (1) with Theorem 17.3.3 of [12], which asserts that, for \( \| \mu - \nu \|_1 \leq 1/2 \), we have

\[
\| H(\mu) - H(\nu) \| \leq \| \mu - \nu \|_1 \log \frac{d}{\| \mu - \nu \|_1},
\]

where \( H(\cdot) \) is the entropy functional defined in (3). Indeed, taking \( \mu \) as in (1) and \( \nu \) to be \( \mu_n \), yields a fully empirical estimate on \( \| H(\mu) - H(\mu_n) \| \). For fixed \( d < \infty \), no technique relying on the plug-in estimator can yield minimax rates [40]. The plug-in is, however, asymptotically optimal for fixed \( d < \infty \) [29] as well as strongly universally consistent even for \( d = \infty \) [3], 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 \( \mathbb{E}X = \frac{1}{n} \sum_{i=1}^{n} X_i \) via the [39] 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 \( X = -\log \mu(I) \), where \( I \sim \mu \), but rather than being given iid samples of \( X \), we are only given draws of \( I \). In Corollary 1, we provide an empirical entropy estimate under a \((1 + \varepsilon)\)th moment assumption (for any \( \varepsilon > 0 \)) on \( X = -\log \mu(I) \).

\(^1\)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, \( \mathbb{P}\sum_{i=1}^{n} X_i \to \mathbb{E}X \) 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 Bahs-Esseen inequality [30] 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 [26].
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 5. Our moment assumption is natural and considerably less restrictive than the finite-alphabet and tail conditions studied in previous works (see Sections VII and A). Moreover, as we argue in Theorem 4, without such a moment assumption, an empirical bound is not feasible. As we demonstrate in Section VI, the rates provided by our empirical bound compare favorably against the state of the art.

II. 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 $\Delta_N$. For $d \in \mathbb{N}$, we write $\Delta_d \subset \Delta_N$ to denote those $\mu$ whose support is contained in $[d]$.

We define the operator $(\cdot)^d$, which maps any $\mu \in \Delta_N$ to its non-increasing rearrangement $\mu^d$. The set of all non-increasing distributions will be denoted by $\Delta_N^2 := \{\mu^d : \mu \in \Delta_N\}$.

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

$$H(\alpha)(\xi) := \sum_{j \in \mathbb{N}, \xi(j) > 0} \xi(j) \log \xi(j)^\alpha.$$  

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

$$\Delta_N^{(\alpha)}[h] = \{\mu \in \Delta_N : H(\alpha)(\mu) \leq h\}$$

and also $\Delta_N^{(\alpha)} := \bigcup_{h \geq 0} \Delta_N^{(\alpha)}[h]$ and $\Delta_N^{(1)}[h] := \Delta_N^{(1)} \cap \Delta_N^{(\alpha)}[h]$.

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

$$R_n^{(\alpha)}(h) := \inf_{\hat{H}} \sup_{\mu \in \Delta_N^{(\alpha)}[h]} \mathbb{E}[\hat{H}(X_1, \ldots, X_n) - H(\mu)],$$

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

III. MAIN RESULTS

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

**Theorem 1:** For all $\alpha > 1$, $H : \Delta_N^{(\alpha)} \to \mathbb{R}_+$ is uniformly continuous under $\ell_1$. In particular, for all $\mu, \nu \in \Delta_N^{(\alpha)}$ satisfying $\|\mu - \nu\|_\infty < e^{-\alpha}$, we have

$$|H(\mu) - H(\nu)| < 2e^{\alpha} \frac{\log \alpha}{\alpha} \frac{1}{\|\mu - \nu\|_\infty}$$

**Theorem 2:** For every $0 < \varepsilon < 1/2$ and $\alpha \geq 1$, there are $\mu, \nu \in \Delta_N$ such that $\varepsilon := \|\mu - \nu\|_1 \to 0$, it holds that $F(\varepsilon_n, h) \to 0$. Moreover, the upper bound in Theorem 1 is tight, up to a constant factor:

$$H(\mu) - H(\nu) \leq F(\|\mu - \nu\|_1, h), \quad h > 0, \mu, \nu \in \Delta_N^{(1)}[h]$$

with the additional property that for any two sequences $\mu_n, \nu_n \in \Delta_N$ satisfying $\varepsilon_n := \|\mu_n - \nu_n\|_1 \to 0$, it holds that $F(\varepsilon_n, h) \to 0$. Moreover, the upper bound in Theorem 1 is tight, up to a constant factor:

$$H(\mu) - H(\nu) \leq \frac{c\varepsilon^{1-1/\alpha}}{\alpha} \left(2e^{1/\alpha} \log \frac{1}{\varepsilon} + H^{(1)}(\mu)^{1/\alpha} + H^{(1)}(\nu)^{1/\alpha}\right),$$

where $c \geq \frac{1}{\alpha} + \frac{1}{2}$ is a universal constant.

Perhaps surprisingly, it turns out that $H : \Delta_N^{(\alpha)}[h] \to \mathbb{R}_+$ is uniformly continuous not only under $\ell_1$, but actually under all $\ell_p$ norms:

**Theorem 3:** There is a function $F : \mathbb{R}_+^4 \to \mathbb{R}_+$ such that

$$|H(\mu) - H(\nu)| \leq F(\|\mu - \nu\|_p, h), \quad h > 0, \alpha > 1, p \in [1, \infty], \mu, \nu \in \Delta_N^{(\alpha)}[h]$$

with the additional property that whenever $\varepsilon_n := \|\mu_n - \nu_n\|_p \to 0$, we have $F(\varepsilon_n, h, \alpha, p) \to 0$.

**Remark:** Although Theorem 3 establishes uniform continuity, it gives no hint as to the functional dependence of the modulus of continuity $F$ on $\alpha, p, h, \|\mu - \nu\|_p$. We leave this as a fascinating open problem — even though the practical applications are likely to be limited: it follows from [41] and Theorem 5 that for $p = \alpha = 2$ and fixed $h$, $F(\|\mu - \nu\|_2, h, 2, 2)$ cannot decay at a faster rate than $1/\log(1/\|\mu - \nu\|_2)$. 

\footnote{Since $\ell_1$ dominates all of the $\ell_p$ norms, continuity of a function under $\ell_p$ trivially implies continuity under $\ell_1$, but the reverse implication is generally not true.}

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Combining Theorem 1 with (1) yields an empirical (under moment assumptions) bound for the plug-in entropy estimator:

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

\[
|H(\mu) - H(\hat{\mu}_n)| \\
\leq \left( 2 \alpha \log \log \frac{4}{\delta} + 6 \sqrt{\frac{\log(4/\delta)}{2n}} \right)^{1-1/\alpha}
\]

holds with probability at least \( 1 - \delta. \) For all \( \alpha > 1, h > 0, \delta \in (0, 1), n \geq \frac{1}{\alpha} \log \frac{4}{\delta}, \) and \( \mu \in \Delta_{\alpha}^{(1)}[h], \) we have that

\[
|H(\mu) - H(\hat{\mu}_n)| \\
\leq \left( 2 \left( \frac{\alpha}{2} \log \frac{4}{\delta} \right)^{1/\alpha} + h + H(\hat{\mu}_n) \right)^{1/\alpha} \\
+ H(\hat{\mu}_n)^{1/\alpha} \\
\cdot \left( 2 \sqrt{\frac{\log(4/\delta)}{2n}} \right)^{1-1/\alpha}
\]

holds with probability at least \( 1 - \delta. \)

**Remark:** Since the estimates in Corollary 1 involve the random quantity \( ||\hat{\mu}_n||_{1/2}, \) it is natural to inquire as to the behavior of the latter. It follows from [10, Theorem 2.3] that \( n^{-1/2} ||\hat{\mu}_n||_{1/2} \rightarrow 0 \) in almost surely. The rate of convergence must necessarily depend on \( \mu \) itself (cf. Remark 9 of [6]).

In Section VI, 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 [33, Appendix A] that assuming \( H(\mu) < \infty \) does not suffice to yield a minimax rate for the \( L_2 \) risk:

\[
\inf_{\tilde{H}: \tilde{H} \geq H} \sup_{\mu \in \Delta_{\alpha}^{(1)}} \mathbb{E} \left( \tilde{H}(X_1, \ldots, X_n) - H(\mu) \right)^2 = \infty.
\]

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 \( \mu \) to \( \Delta_{\alpha}^{(1)}[h] \) and obtaining a finitary, quantitative lower bound:

**Theorem 4:** For \( \alpha = 1, \) there is a universal constant \( C > 0 \) such that for all \( h > 1 \) and \( n \in \mathbb{N}, n \geq 2, \) we have \( R_{\alpha}^{(1)}(h) \geq Ch. \)

**Remark:** The above result complements — but is not directly comparable to — [3, 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 5:** The following bounds hold for the \( L_1 \) minimax risk:

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

\[
R_{\alpha}^{(1)}(h) \leq \frac{1 + \log n}{\sqrt{n}} + \frac{2^{\alpha-1}h}{\log^{\alpha-1} n}, \quad n \in \mathbb{N};
\]

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

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

\[
R_{\alpha}^{(1)}(h) \geq \frac{h}{4 \cdot 3^\alpha \log^{\alpha-1} n}.
\]

**Open Problem:** Close the gap in the dependence on \( \alpha \) in the upper and lower bounds.

**Open Problem:** Another gap between the upper and lower bounds is the quantified on \( h: \) in the upper bound, it is “for all”, while in the lower bound, it is “exists”. Closing this gap is also of interest.

Finally, in Section VII-C, we resolve most of the conjectures posed by [22].

**IV. RELATED WORK**

**A. Continuity, Convergence, Moments of Information**

Reference [42] gave a sharpened version of (2) and [19] presented analogous bounds; [5] proved a non-computative generalization. [31, Theorem 5] upper-bounds \( H(\mu) - H(\nu) \) in terms of quantities related to \( ||\mu - \nu||_1, \) where (at most) one of them is allowed to have infinite support. Even though \( H(\cdot) \) is not continuous on \( \Delta_{\alpha}, \) the plug-in estimate \( H(\hat{\mu}_n) \) converges to \( H(\mu) \) almost surely and in \( L_2 \) [3]. Reference [33] studied a variety of restrictions on distributions over infinite alphabets to derive strong consistency results and rates of convergence. Moments of information were apparently first defined in [15].

**B. Entropy Estimation**

Recent surveys of entropy estimation results may be found in [21] and [38]. The finite-alphabet case is particularly well-understood. For fixed alphabet size \( d < \infty, \) the plug-in estimate is asymptotically minimax optimal [29]. [28] non-constructively established the existence of a sublinear (in \( d \)) entropy estimator. The optimal dependence on \( d \) (at fixed accuracy) was settled by [35] and [37] 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 [35], [36], and [37]. The works of [16], [20], [21], and [40] characterized the minimax rates for the high-dimensional regime: a small additive error of \( \epsilon \) requires \( \Theta(d/\epsilon \log d) \) samples. Building off of these polynomial-approximation based
constructions, [1] design an additional optimal estimator, this one based on a profile maximum likelihood approach that can also estimate a variety of other important statistics. References [13] and [14] generalize the optimal estimators to estimate any additive functional, recovering in particular the optimal rates for entropy. Reference [2] modify these optimal estimators with the added goal of low space complexity.

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 [3] 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 4 and the Remark following it). Reference [3] 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 $O(n^{-\beta})$. Reference [41] showed that the plug-in estimate achieves a rate of $O(\frac{1}{\log n})$ for bounded second moment, and this is minimax optimal. Reference [8] exhibited a function of the higher moments that can be used in place of alphabet size to give a multiplicative approximation to the entropy.

The empirical nature of Corollary 1 can be seen as a distribution-dependent improvement over otherwise worst-case minimax guarantees. It can be compared, in this light, to the “instance-optimality” program of [17] and [18] and the adaptive guarantees of [16].

V. PROOFS

A. Proof of Theorem 1

We begin with a subadditivity result for the oth 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 $\mu, \nu \in \Delta_\mathbb{N}^{(\alpha)}$, we have

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

(11)

If, additionally, $|\mu - \nu|_\infty \leq \epsilon$, then

$$H^{(\alpha)}(|\mu - \nu|) \leq 2\epsilon \alpha \alpha + \|\mu - \nu\|_\infty \log \frac{1}{\|\mu - \nu\|_\infty} + H^{(\alpha)}(\mu) + H^{(\alpha)}(\nu).$$

(12)

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

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

(13)

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

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

(14)

To prove the lemma, we bound the two terms of (14) separately. The second term can be bound in two ways, yielding (11) and (12), respectively. To bound the first term of (14), notice that $\mu \in \Delta_\mathbb{N}$ implies that $\{i \in \mathbb{N}: \mu(i) > e^{-\alpha}\} \leq e^\alpha$, and similarly for $\nu$. Thus,

$$\sum_{i: \mu(i) \nu(i) > e^{-\alpha}} h^{(\alpha)}(|\mu(i) - \nu(i)|) \leq 2e^\alpha e^{-\alpha} \alpha = 2\alpha.$$

For the second term of (14), notice that when $\mu(i) \nu(i) \leq e^{-\alpha}$, the monotonicity of $h^{(\alpha)}$ implies

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

and hence

$$\sum_{i: \mu(i) \nu(i) \leq e^{-\alpha}} h^{(\alpha)}(|\mu(i) - \nu(i)|) \leq 2e^\alpha e^{-\alpha} \alpha = 2\alpha.$$

Given the additional condition $\|\mu - \nu\|_\infty \leq \epsilon$, to prove (12), put $\epsilon = \|\mu - \nu\|_\infty$ and modify (15) as follows:

$$\sum_{i: \mu(i) \nu(i) > e^{-\alpha}} h^{(\alpha)}(|\mu(i) - \nu(i)|) \leq \sum_{i: \mu(i) \nu(i) > e^{-\alpha}} h^{(\alpha)}(\mu(i) \nu(i)) \leq 2e^\alpha h^{(\alpha)}(\epsilon).$$

The latter holds since $|\mu(i) - \nu(i)| \leq \|\mu - \nu\|_\infty$, and so $h^{(\alpha)}(|\mu(i) - \nu(i)|) \leq h^{(\alpha)}(\epsilon)$, by $h^{(\alpha)}$’s monotonicity on $[0, e^{-\alpha}]$.

Proof of Theorem 1: The concavity argument in the proof of [12, Theorem 17.3.3], immediately implies

$$|H(\mu) - H(\nu)| \leq H(|\mu - \nu|).$$

Then, via an application of Hölder’s inequality,

$$H(|\mu - \nu|) = \sum_{i \in \mathbb{N}} |\mu(i) - \nu(i)| \log \frac{1}{|\mu(i) - \nu(i)|} \cdot |\mu(i) - \nu(i)|^{1/\alpha} \cdot \log \frac{1}{|\mu(i) - \nu(i)|^{1/\alpha}}$$

$$\leq \left( \sum_{i \in \mathbb{N}} |\mu(i) - \nu(i)|^{1/\alpha} \log \frac{1}{|\mu(i) - \nu(i)|} \right)^{1/1-\alpha} \cdot \left( \sum_{i \in \mathbb{N}} |\mu(i) - \nu(i)|^{1/\alpha} \log \frac{1}{|\mu(i) - \nu(i)|^{1/\alpha}} \right)^{1/\alpha}$$

$$= \|\mu - \nu\|_1^{1-\alpha} \alpha \log \frac{1}{|\mu(i) - \nu(i)|^{1/\alpha}}.$$
The claim follows by invoking Lemma 1 and the subadditivity of $t \mapsto t^{1/\alpha}$ for $t \geq 0$ and $\alpha > 1$. □

B. Proof of Theorem 2

Let $0 < \varepsilon \leq 1/2$ be given and choose $\mu, \nu \in \Delta_N$ as follows: $\mu = (1, 0, 0, \ldots)$ and $\nu = (1 - \varepsilon, \varepsilon, 0, 0, \ldots)$. Then left-hand side of (7) is $L(\varepsilon) = H(\nu)$:

$$L(\varepsilon) = (1 - \varepsilon) \log \frac{1}{1 - \varepsilon} + \varepsilon \log \frac{1}{\varepsilon} =: L_1(\varepsilon) + L_2(\varepsilon),$$

and right-hand side of (7), without the constant $c$, is

$$R(\varepsilon) = \left(1 - \varepsilon\right) \log \frac{1}{1 - \varepsilon} + \varepsilon \log \frac{1}{\varepsilon}\left(1 + \frac{1}{1 - \varepsilon}\right)^{1/\alpha} \leq 2^{1/\alpha} \max\left\{\left(1 - \varepsilon\right)^{1/\alpha} \log \frac{1}{1 - \varepsilon}, \frac{1}{\varepsilon}\right\} \leq 2e^{-1/\alpha} \left(1 - \varepsilon\right)^{1/\alpha} \log \frac{1}{1 - \varepsilon} + (2e + 2) \varepsilon \log \frac{1}{\varepsilon} =: R_1(\varepsilon) + R_2(\varepsilon).$$

Now $L_1(\varepsilon) \geq R_1(\varepsilon)/2$ for $\varepsilon \in (0, \frac{1}{2})$ and $L_2(\varepsilon) = \frac{R_2(\varepsilon)}{2\varepsilon + 2}$, and therefore $L_1(\varepsilon) + L_2(\varepsilon) \geq R_1(\varepsilon)/2 + \frac{R_2(\varepsilon)}{2\varepsilon + 2} \geq \frac{1}{2\varepsilon + 2} R(\varepsilon)$. □

C. Proof of Theorem 3

The following fact [24, Theorem 3.5 and Eq. (5) on p. 83] will be useful:

$$\|\mu - \nu\| \leq \|\mu\| \leq \|\mu - \nu\|_p, \quad p \in [1, \infty], \mu, \nu \in \Delta_N. \quad (16)$$

A result of [32] (more accurately credited to Riesz, 1928 [23]) implies that a sequence $\{\xi_n\} \subset \ell_1(\mathbb{N})$ converging pointwise to some $\xi \in \ell_1(\mathbb{N})$ also converges in $\ell_1$ iff $\|\xi_n\| \to \|\xi\|$. This immediately implies

Lemma 2: If $\{\mu_n\} \subset \Delta_N$ converges pointwise to some $\mu \in \Delta_N$, then it also converges in $\ell_1$.

[7, Lemma 1] showed that $\Delta_N^{[1]}[h]$ is compact under $\ell_p$.

We begin by extending this result to general $\alpha, p$.

Lemma 3: For all $\alpha \geq 1, p \in [1, \infty]$, and $h > 0$, the set $\Delta_N^{[1]}[h]$ is compact under $\ell_p$.

Remark: This is quite false if either the non-increasing or the bounded-entropy condition is omitted. For a counterexample to the former, consider the sequence $\mu_n \in \Delta_N$ defined by $\mu_n(i) = 1[i = n]$. For a counterexample to the latter, consider the sequence $\mu_n \in \Delta_N$, where $\mu_n$ is uniform on $[n]$. Proof: We closely follow the proof strategy of [7, Lemma 1]. In a metric space, compactness and sequential compactness are equivalent. Let $\mu_n \in \Delta_N$ be a sequence in $\Delta_N^{[1]}[\alpha][h]$. Since $[0, 1]$ is compact, every $\{\mu_n(i) : n \in \mathbb{N}\}$ has a convergent subsequence, and hence $\mu_n \in \Delta_N$ has a pointwise convergent subsequence. There is thus no loss of generality in assuming that $\mu_n \to \mu$ pointwise. Obviously, $\mu$ is non-negative and non-increasing. It remains to show that

(a) $\sum_{i \in \mathbb{N}} \mu(i) = 1$,
(b) $H(\alpha)(\mu) \leq h$,
(c) $\|\mu_n - \mu\|_p \to 0$.

To show (a), assume, for a contradiction, that $\sum_{i \in \mathbb{N}} \mu(i) > 1$. Then there must be an $i_0 \in \mathbb{N}$ such that $\sum_{i=i_0}^{\infty} \mu(i) > 1$. But the latter must then hold for all $\mu_n$ with $n$ sufficiently large, which contradicts $\mu_n \in \Delta_N$. Now assume $\mu := 1 - \sum_{i \in \mathbb{N}} \mu(i) > 0$. For any $i_0 \in \mathbb{N}$, we have $\sum_{i=i_0}^{\infty} \mu(i) < 1 - \varepsilon/2$ for all sufficiently large $n$. Now every $\nu \in \Delta_N$ satisfies $\nu(i) \leq \frac{1}{n}(\nu(1) + \nu(2) + \ldots + \nu(i)) \approx \frac{1}{i}$. Hence,

$$\sum_{i=i_0+1}^{\infty} \mu_n(i) |\log \mu_n(i)|^\alpha \geq \sum_{i=i_0+1}^{\infty} \mu_n(i)(|\log i_0|)^\alpha \geq \frac{1}{i} (\log io)^\alpha.$$ 

Choosing $i_0$ sufficiently large makes the latter expression exceed $h$, violating the assumption $\mu_n \in \Delta_N^{[1]}[\alpha][h]$. Thus (a) holds.

To show (b), assume, for a contradiction that $H(\alpha)(\mu) > h$ and, in particular, $\sum_{i=i_0}^{\infty} \mu(i) |\log \mu(i)|^\alpha > h$ for some $i_0 \in \mathbb{N}$. But the latter must hold for all $\mu_n$ with $n$ sufficiently large, a contradiction.

Finally, to show (c), we invoke Lemma 2: if $\{\mu_n\} \subset \Delta_N$ converges pointwise to some $\mu \in \Delta_N$, then it also converges in $\ell_1$. Since $\ell_1$ dominates every $\ell_p, p > 1$, this proves (c). □

Next, we examine the continuity of $H(\cdot)$ on $\Delta_N^{[1]}[\alpha][h]$ under $\ell_p$.

Lemma 4: Fix $h > 0, \alpha > 1$, and $p \in [1, \infty]$. If $\{\mu_n\} \subset \Delta_N^{[1]}[\alpha][h]$ converges in $\ell_p$, then its limit is some $\mu \in \Delta_N^{[1]}[\alpha][h]$ and furthermore, $H(\mu_n) \to H(\mu)$. In other words, $H(\cdot)$ is continuous on $\Delta_N^{[1]}[\alpha][h]$ under $\ell_p$.

Remark: We note that $H(\cdot)$ is not continuous on $\Delta_N^{[1]}[\alpha][h]$ under $\ell_p, p \in [1, \infty], as evidenced by the sequence $\mu_n = (1 - \varepsilon_n, \varepsilon_n/n, \ldots, \varepsilon_n/0, \ldots)$, with support size $n + 1$. We can choose $\varepsilon_n$ so that $H(\mu_n) = h$, but of course the limiting $\mu$ has $H(\mu) = 0$ (see Example 1 in [7]). Proof: It follows from Lemma 3 that the limiting $\mu$ belongs to $\Delta_N^{[1]}[\alpha][h]$. Further, Lemma 2 implies that $\mu_n \to \mu$ in $\ell_1$. Invoking the continuity result in Theorem 1 proves the claim. □

Proof of Theorem 3: It follows from Lemma 4 that $H(\cdot)$ is continuous on $\Delta_N^{[1]}[\alpha][h]$ under $\ell_p$. Since, by Lemma 3, $\Delta_N^{[1]}[\alpha][h]$ is compact under $\ell_p$, it follows that $H(\cdot)$ is uniformly continuous on $\Delta_N^{[1]}[\alpha][h]$; there is a function $F$ such that

$$|H(\mu) - H(\nu)| \leq F(||\mu - \nu||_p, h, \alpha, p), \quad \mu, \nu \in \Delta_N^{[1]}[\alpha][h]$$

and $\varepsilon_n := ||\mu_n - \nu_n||_p \to 0 \implies F(\varepsilon_n, h, \alpha, p) \to 0$. Now, for all $\mu, \nu \in \Delta_N^{[1]}[\alpha][h]$ we have

$$|H(\mu) - H(\nu)| = |H(\mu^1) - H(\mu^1)|$$

In the sequel, we make use of the following facts.

The result is stated for functions in $L_2(\mathbb{R}^n)$ and their symmetric-decreasing rearrangements $f^*$, but the specialization to discrete distributions is straightforward. We convert $f$ to a function $g: \mathbb{R}^n \to [0, 1]$ via $g(x) = \mu(|x|)$ and $\nu$ to $g(x)$ analogously. A direct calculation then shows that $||\mu - \nu||_p = ||f - g||_p$ and $||\mu^* - \nu^*||_p = ||f^* - g^*||_p$, to which the result from [24] applies to yield (16).
It follows from (16) that \( \| \mu_n - \nu_n \|_p \to 0 \implies \| \mu_n - \nu_h \|_p \to 0 \), which concludes the proof. \( \square \)

D. Proof of Corollary 1

Proof: Fix \( 0 < \varepsilon < e^{-\alpha} \). Consider two potential “bad” events: \( B_1 \), where \( \| \mu_n - \mu \|_\infty > \varepsilon \), and \( B_2 \), where \( \| \mu - \nu_n \|_1 > 2\| \mu_n - \mu \|_\infty + 6\sqrt{\log(4/\delta)}/n \). Our assumption on the sample size \( n \), together with the Dvoretzky-Kiefer-Wolfowitz inequality [27], implies that \( \mathbb{P}(B_1) \leq \delta/2 \) and (1) implies that \( \mathbb{P}(B_2) \leq \delta/2 \). Thus, with probability at least \( 1 - \delta \), neither of \( B_1 \) or \( B_2 \) occurs, and on the event where \( \| \mu - \nu_n \|_\infty < \varepsilon < e^{-\alpha} \), we may invoke Theorem 1, from which the claims immediately follow. \( \square \)

E. Proof of Theorem 4

For \( h > 1 \) and \( n \in \mathbb{N}, n \geq 2, p_i = (1-1/(2n)) \log(1-1/(2n)) \) and define the support size \( S = S(h, n) \) by \( S = \lfloor (1/2n) \exp(2n(h/2 + a_n)) \rfloor \). Consider the distributions \( \mu_0 = (1, 0, 0, \ldots) \) and \( \mu_n \) defined by \( \mu_n(1) = 1 - 1/(2n) \), and

\[
\mu_n(i) = \frac{1}{2nS}, \quad 2 \leq i \leq 1 + S(h, n).
\]

We compute the Kullback-Leibler divergence and entropy:

\[
D_{\text{KL}}(\mu_0||\mu_n) = \log \frac{1}{1-1/(2n)} \leq \frac{1}{1-1/(2n)} - 1 \leq \frac{1}{n} \quad (17)
\]

\[
H(\mu_0) = 0 \leq h.
\]

For \( x \geq 2 \), always \( |x| \geq x/2 \). Additionally, from \( 2na_n \geq -1 \), and \( x \geq 1/2 \), we obtain that \( S > (1/2n) \exp(2n(h/2 + a_n)) \), hence we also have that \( h \geq H(\mu_n) > h - \frac{1}{2n} \log 2 \). Since \( \frac{1}{2n} \log 2 \leq 1/2 \) on \( [1, \infty) \) and \( h > 1 \), it follows that \( H(\mu_n) \geq \frac{h}{2} \), whence \( H(\mu_0) - H(\mu_n) \geq h/2 \). To bound the \( L_1 \) minimax risk (defined in (4)), we invoke Markov’s inequality:

\[
\mathbb{E}[|\bar{H}(X_1, \ldots, X_n) - H(\mu)|] \geq \frac{h}{4} \mathbb{P} \left( |\bar{H}(X_1, \ldots, X_n) - H(\mu)| > \frac{h}{4} \right).
\]

It follows via Le Cam’s two point method [34, Section 2.4.2] that

\[
R_n(1)(h) \geq \frac{h}{8} \exp(-nD_{\text{KL}}(\mu_0||\mu)) \geq \frac{h}{8e},
\]

where the second inequality stems from (17). \( \square \)

F. Proof of Theorem 5

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

\[
\text{Lemma 5: For all } \mu \in \Delta_N \text{ and } n \in \mathbb{N}, \text{ we have}
\]

\[
H(\mu) \geq \text{EH}(\hat{\mu}_n)
\]

\[
\geq \mathbb{E} \left[ \sum_{i \in \mathbb{N}^\prime} \frac{1}{\mu(i)} \log \frac{1}{\mu(i)} \right] + \log \left( 1 + \frac{1}{\varepsilon n} \right).
\]

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

\[
\text{EH}(\hat{\mu}_n) = \mathbb{E} \left[ \sum_{i \in \mathbb{N}^\prime} \hat{\mu}_n(i) \log \frac{1}{\mu(i)} \right] + \mathbb{E} \left[ \sum_{i \in J} \hat{\mu}_n(i) \log \frac{1}{\sum_{i \in J} \mu(i)} \right]
\]

where \( \hat{\mu}_n \) is the “collapsed” version of \( \mu_n \), where all of the masses in \( J \) have been replaced by a single mass equal to their sum, and the inequality holds because conditioning reduces entropy [12, Eq.(2.157)]. We observe that \( \hat{\mu}_n \) has support size at most \( 1 + 1/\varepsilon \) and invoke [29, Proposition 1]:

\[
\mathbb{E} \text{EH}(\hat{\mu}_n) \geq H(\hat{\mu}) - \log \left( 1 + \frac{1}{\varepsilon n} \right),
\]

where \( \hat{\mu} \) is the “collapsed” version of \( \mu \). Now

\[
H(\hat{\mu}) = \sum_{i \in J} \mu(i) \log \frac{h}{\sum_{i \in \mathbb{N}^\prime} \mu(i)} \geq \sum_{i \in J} \mu(i) \log \frac{1}{\mu(i)}
\]

\[
\geq H(\mu) - \sum_{i \in J} \mu(i) \frac{1}{\mu(i)} \log \frac{1}{\mu(i)},
\]

which concludes the proof. \( \square \)

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

\[
\text{Proposition 1: For all } \alpha \geq 1, h > 0, n \in \mathbb{N} \text{ and } \mu \in \Delta_N^{(\alpha)}[h] \text{, we have}
\]

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

Proof: Since by Lemma 5, \( |H(\mu) - \text{EH}(\hat{\mu}_n)| = H(\mu) - \text{EH}(\hat{\mu}_n) \), it follows from the triangle and Jensen inequalities
that
\[ \text{EH}(\hat{\mu}_n) \geq H(\mu) - \frac{1}{\varepsilon} \sum_{i : \mu(i) < \varepsilon} \frac{1}{\mu(i)} \leq \left( \log \frac{1}{\varepsilon} \right)^{1-\alpha} \sum_{i : \mu(i) < \varepsilon} \left( \log \frac{1}{\mu(i)} \right)^\alpha \leq \left( \log \frac{1}{\varepsilon} \right)^{1-\alpha} H(\alpha)(\mu), \] (20)
where the second and third inequalities follow from the obvious relations
\[ \sum_{i : \mu(i) < \varepsilon} \mu(i) \log \frac{1}{\mu(i)} \leq \left( \log \frac{1}{\varepsilon} \right)^{1-\alpha} \sum_{i : \mu(i) < \varepsilon} \left( \log \frac{1}{\mu(i)} \right)^\alpha \leq \left( \log \frac{1}{\varepsilon} \right)^{1-\alpha} H(\alpha)(\mu). \]

The claim follows by combining (19) with (20). \( \square \)

Proof: [Proof of Theorem 3(a)]
Use the fact that \( R_\alpha^{(n)}(h) \leq \text{EH}(\mu) - \text{EH}(\hat{\mu}_n) \), invoke Proposition 1 with \( \varepsilon = \frac{1}{\sqrt{n}} \) and use log(1 + x) \( \leq x \). \( \square \)

We now prove the second half of the theorem.

Proof: [Proof of Theorem 5(b)] Let \( \alpha > 0, n \in \mathbb{N} \) and define two families of distributions:
\[ \mathcal{U}_1 := \{ \mu_1 = \text{Uniform}([n^3]) \}, \]
\[ \mathcal{U}_2 := \{ \mu_2 = \text{Uniform}(A) : A \subset [n^3], |A| = n^2 \}. \]

Let \( h := 3^\alpha \log^\alpha n \) and note that \( \mathcal{U}_1 \cup \mathcal{U}_2 \subseteq \Delta_n^{(\alpha)}[h] \). Let \( E \) be the event that \( X = (X_1, \ldots, X_n) \) has no repeating elements, i.e. \( \{(X_1, X_2, \ldots, X_n) : n \} \). Let \( \mu_1 \in \mathcal{U}_1, \mu_2 \in \mathcal{U}_2 \) and consider the values \( \mathbb{P}_{X \sim \mu_1^n}(E) \) and \( \mathbb{P}_{X \sim \mu_2^n}(E) \). For \( m \in \mathbb{N} \), define \( K(m) \) to be the smallest \( k \) such that uniformly throwing \( m \) balls into \( k \) buckets, the probability of collision is at least 1/2. Since \( K(m) \) is known\(^5\) to be at least \( \sqrt{n} \) (and hence \( K(n^2) > n \)) we have a lower bound of \( \frac{1}{n} \) on both \( \mathbb{P}_{X \sim \mu_1^n}(E) \) and \( \mathbb{P}_{X \sim \mu_2^n}(E) \). Define \( \mu_1^n \) as the distribution on \( \mathbb{N}^n \) induced by conditioning the product \( \mu_1^n \) on the event \( E \), and define \( \mu_2^n \) analogously. Our key observation is that conditional on \( E \), \( \mu_1^n \) is uniform on \( ([n^3])_n \) whereas \( \mu_2^n = \text{Uniform}(A)^n \) is uniform on \( (A)_n \),

\(^5\)Better bounds exist [9].
we will benefit from tighter bounds. This entails some cost, 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 (plug-in estimator) 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 $\ell_1$ occurs at rate $O(\sqrt{d/n})$, whereas the MLE entropy estimator converges at rate $O\left(\sqrt{\frac{d}{n^2} + \frac{\log^2 d}{n}}\right)$, as follows from [40, Proposition 1]. So any approach that upper bounds the entropy risk via $\Theta(\log n)$, as in the proof of Lemma 1 (recall that $\Lambda_n(\mu)$ is increasing on $[0, e^{-\alpha}]$, 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.

Nevertheless, for certain classes of distributions our bounds (Theorem 1 and Corollary 1) 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:

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

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

$$\mathbb{E}h^{(\alpha)}(\hat{\mu}_n) = \sum_{i \in [d]} h^{(\alpha)}(\hat{\mu}_n(i)) = \sum_{i \in [d]} h^{(\alpha)}(\hat{\mu}_n(i)) 1[\hat{\mu}_n(i) < e^{-\alpha}] + \sum_{i \in [d], \mu(i) < e^{-\alpha}} h^{(\alpha)}(\hat{\mu}_n(i)) 1[\hat{\mu}_n(i) < e^{-\alpha}] + 2\alpha^\alpha \sum_{z \in [e^{-\alpha}, 1]} h^{(\alpha)}(z),$$

where $(i)$ follows from Jensen’s inequality, $(ii)$ is because $h^{(\alpha)}(z)$ is increasing on $z \in [0, e^{-\alpha}]$, and $(iii)$ is from (13).

By [6, Lemma 6], we have $\mathbb{E}[\hat{\mu}_n - \mu]_1 \leq \Lambda_n(\mu)$, where

$$\Lambda_n(\mu) := 2 \sum_{\mu(j) < 1/n} \mu(j) + \frac{1}{\sqrt{n}} \sum_{\mu(j) \geq 1/n} \sqrt{\mu(j)}.$$ 

This quantity is always finite and $\Lambda_n(\mu) \to 0$ for all $\mu \in \Delta_n'$ (ibid). Thus, we can optimize the bound

$$\mathbb{E}[H(\hat{\mu}_n) - H(\mu)] \leq \left(4\alpha^\alpha + 2H(\alpha)(\mu)\right)^{1/\alpha} \Lambda_n(\mu)^{1-1/\alpha}. \tag{21}$$

### A. Finite Support

For distributions with a large support but concentrated mass, the bound in (21) 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 $d'$; $\mu'$ is uniform over $[d]$, $\mu''$ is uniform over $[d + D]$, and $\mu := p\mu' + (1 - p)\mu''$, for some $p \in [0, 1]$. The state-of-the-art upper bound for the plug-in estimator can be inferred from [40, Appendix D], and has the form

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

for some $C > 1$; notice that it is insensitive to $p$. For a fair comparison to (21), our estimator’s only a priori knowledge of $\mu$ is that its support is of size at most $d + D$. By Proposition 2, we have $\max_{\mu \in \Delta_n} H^{(\alpha)}(\mu) \leq \max\{\alpha, \log K\}^{\alpha} + (\alpha/\epsilon)^\alpha$. This allows us to optimize over $\alpha$ for each $n$:

$$\text{OUR}(d, D, p, n) := \inf_{\alpha > 1} \left(4\alpha^\alpha + 2\max\{\alpha, \log(d + D)\}^{\alpha} + 2(\alpha/\epsilon)^\alpha\right)^{1/\alpha} \Lambda_n(\mu)^{1-1/\alpha}.$$

Since $\mu$ 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 concave, the latter has the form

$$\mathbb{E}[H(\hat{\mu}_n) - H(\mu)] \leq \mathbb{E}\left[\|\hat{\mu}_n - \mu\|_1 \log \frac{d + D}{\|\hat{\mu}_n - \mu\|_1}\right] \leq \Lambda_n(\mu) \log \frac{d + D}{\Lambda_n(\mu)},$$

The comparisons are plotted in Figure 1 (Left).

### B. 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 $\mu_q(i) \sim 1/i^q$ with parameter $q > 1$. For this family, [3, Theorem 7] establish a lower bound of order $n^{-q}$ on $\mathbb{E}[H(\hat{\mu}_n) - H(\mu_q)]$. It is straightforward to verify that $\mu_q \in \Delta_n'$ for all $q, \alpha > 1$. Thus, we can optimize our bound in (21) over all $\alpha > 1$; the results are presented in Figure 1 (Right).

---

*One can, for example, apply Cauchy’s condensation test, followed up by the ratio test.*
VII. MOMENTS OF INFORMATION

We motivate our bounded-moment assumption as being considerably less restrictive than the finite-alphabet assumptions and tail conditions studied in previous works [33], [40]; see Section A for a detailed comparison. Obtaining moment-based results is essentially a desideratum laid out by [3], see Section A for a detailed comparison. Obtaining moment-based results is essentially a desideratum laid out by [3], in which it is hypothesized that — in parallel to the asymptotic distribution for the finite alphabet case — moment conditions are the correct notion to achieve finite-sample estimates in the infinite alphabet case. Our Theorem 5(a) shows that, under these assumptions, there is an inverse logarithmic convergence rate (similar to, though distinct from, the results of [41]) and, furthermore, using empirical quantities, this rate can be very much accelerated, as demonstrated in Corollary 1.

In this section, we study some of the mathematical properties of moments of information.

A. Maximizing the α-Moment Over a Fixed Support Size

**Proposition 2:** For $K \geq 2$ and $\alpha \geq 1$,
\[
\max \{ \log K, (\alpha/e) \}^\alpha \leq \max_{\mu \in \Delta_K} \mathbb{H}^{(\alpha)}(\mu) \leq \max \{ \log K, \alpha \}^\alpha + (\alpha/e)^\alpha.
\]

We will need the following useful (and likely known) result.

**Lemma 6 (Folklore):** Suppose that $0 < \alpha < 1$ and $f : [0, 1] \to \mathbb{R}$ is strictly concave on $[0, a]$ and strictly convex on $[a, 1]$. Define the function $F : \Delta_K \to \mathbb{R}$ by
\[
F(\mu) = \sum_{i=1}^K f(\mu(i)).
\]

Then any maximizer $\mu^*$ of $F$ is either the uniform distribution or else has exactly 1 “heavy” mass $v \in [a, 1]$ and $K - 1$ identical “light” masses $(1 - v)/(K - 1)$.

**Proof:** A standard “smoothing” argument [25] shows that if two masses $u \leq v$ occur in the interval $(a, 1)$, there is an $\varepsilon > 0$ such that $f(u - \varepsilon) + f(v + \varepsilon) > f(u) + f(v)$. In other words, such masses can be pushed apart (keeping their sum fixed) to increase the value of $F$, until one of them reaches the boundary of $[a, 1]$. Furthermore, since $0 < a < u < v$ and $u + v \leq 1$, repeated iteration of the “pushing apart” operation will hit the left endpoint (i.e., $a$) rather than the right one (i.e., 1). Having exhausted the “pushing apart” process, we are left with one “heavy” mass $v \in [a, 1]$ and $K - 1$ “lighter” ones in $[0, a]$. But concavity implies that $F$ will be maximized by pulling the lighter masses in (as opposed to pushing them apart), which amounts to replacing each of them by the average of the $K - 1$ values.

**Proof:** [Proof of Proposition 2] Choosing $\mu$ to be the uniform distribution yields $H^{(\alpha)}(\mu) = \log^\alpha K$, and choosing $\mu$ such that $v := \mu(1) = e^{-\alpha}$ yields $H^{(\alpha)}(\mu) \geq v \log(1/v)^\alpha = (\alpha/e)^\alpha$. Thus, the lower bound is proven and it only remains to prove the upper bound.

Let $\mu^*$ be a maximizer for given $\alpha, K$. Recall the function $h^{(\alpha)}(z) = z \log^\alpha(1/z)$ and note that it is strictly concave on $[0, e^{-(\alpha - 1)}]$ and strictly convex on $[e^{-(\alpha - 1)}, 1]$. Then Lemma 6 shows that $\mu^*$ will either be uniform or else attains at most one value $v \in [e^{-(\alpha - 1)}, 1]$ in the convex interval, with the remaining values equal to $1 - v/K$, $v \in [0, e^{-(\alpha - 1)}]$ in the concave interval. Only the latter case is non-trivial:
\[
H^{(\alpha)}(\mu^*) = v \left( \log \frac{1}{v} \right)^\alpha + (1 - v) \left( \log \frac{K - 1}{1 - v} \right)^\alpha
\]
for some $v$ satisfying
\[
0 < \frac{1 - v}{K - 1} \leq e^{-(\alpha - 1)} \leq v < 1.
\]

Now $v \left( \log \frac{1}{v} \right)^\alpha$ is maximized over $[0, 1]$ by $v = e^{-\alpha}$, which yields the value $(\alpha/e)^\alpha$.

To bound the second term, $g(v) := (1 - v) \left( \log \frac{K - 1}{1 - v} \right)^\alpha$, we consider two cases: (i) $K - 1 < e\alpha$ and (ii) $K - 1 \geq e\alpha$.

In case (i), $g$ is maximized by $v^* = 1 - (K - 1)/e\alpha$ and $g(v^*) = (1 - v^*) \left( \log \frac{K - 1}{1 - v^*} \right)^\alpha \leq \left( \log \frac{K - 1}{1 - v} \right)^\alpha = \alpha^\alpha$.

In case (ii), $g$ is monotonically decreasing in $v$. The constraint $1/v \leq v$ from (22) implies $v \geq 1/K$, so in this case,
\[
g(v) \leq \left( \log \frac{K - 1}{1 - (1/v)} \right)^\alpha = \log^\alpha K.
\]

This proves the upper bound. □
B. Moments of Information vs. Moments of Distributions
Since for all $r \geq 1$ and $\mu \in \Delta_N$, we trivially have $\|\mu\|_r \leq 1$, it is only for $r < 1$ that $\|\mu\|_r$ conveys nontrivial tail information. However, as a measure of tail decay, the latter is rather crude: $\|\mu\|_r < \infty$ for any $r < 1$ implies $H(\mu) < \infty$ for all $\alpha > 0$.

**Proposition 3:** For all $\alpha > 0$ and $r \in (0, 1)$, we have
\[
H(\alpha)(\mu) \leq \left(\frac{\alpha}{\alpha(1-r)}\right)^{2\alpha} \|\mu\|_r.
\]

**Remark:** The bound above is quite loose. For example, for $\alpha = 1$ the AM-GM inequality readily yields, for any $r \in (0, 1)$,
\[
H(\mu) \leq \frac{r}{1-r} \ln \|\mu\|_r.
\]

It may be of interest to investigate bounds of the form
\[
H(\alpha)(\mu) \leq a(\alpha) \log b(\alpha) \|\mu\|_{c(\alpha)},
\]
for some functions $a, b, c : \alpha \mapsto (0, \infty)$.

**Proof:** We first claim that for $\alpha > 0$ and $r \in (0, 1)$,
\[
x \log^\alpha \frac{1}{x} < x^r, \quad x \in [0, 1].
\]

Indeed, the function $x \mapsto x^{1-r} \log^\alpha \left(\frac{1}{x}\right)$ is maximized at $x = e^\frac{1}{\alpha r}$, attaining the maximum value of $e^{-\alpha \left(\frac{1}{\alpha r}\right)}$. The latter is less than 1 whenever $\alpha < c(1-r)$. Likewise, whenever $\alpha < ce(1-r)$, we have $x \log^\alpha \left(\frac{1}{x}\right) < e^\alpha x^r$, and so $H(\alpha)(\mu) < c^{2\alpha} \|\mu\|_r$. Choosing $c = \frac{1}{\pi(1-r)}$ proves the claim. □

**C. Resolution of [22] Conjectures**

In this section, we give a complete resolution of the conjectures posed by [22].

1) **Conjecture 10.1:** [22, Conjecture 10.1] posits that for $d = 2$, $\max_{\mu \in \Delta_N} H(\alpha)(\mu)$ has two maximizers $\pi_1(\alpha) = \left(\frac{1}{2} + x(\alpha), \frac{1}{2} - x(\alpha)\right)$ and $\pi_2(\alpha) = \left(\frac{1}{2} - x(\alpha), \frac{1}{2} + x(\alpha)\right)$ for some value $x(\alpha)$ such that $x(2) = \frac{1}{2\alpha} \sqrt{e-4}$ and $x(\alpha)$ is strictly increasing as $\alpha \to \infty$ and $\lim_{\alpha \to \infty} x(\alpha) = \frac{1}{2}$.

By Lemma 6, there are at most three maximizers. Since $H(\alpha)(e^{-\alpha}, 1-e^{-\alpha}) > \left(\frac{\alpha}{2}\right) \log(2) > \log^2(2)$, the uniform distribution is not a maximizer. So, including permutations, there are exactly two maximizers.

Let $(u^*_1, v^*_1)$ be the increasingly-ordered maximizing distribution. We cannot have $u^*_1 < e^{-\alpha}$, because this would only decrease $H(\alpha)$ as compared to $H(\alpha)(e^{-\alpha}, 1-e^{-\alpha})$. By Lemma 6, $u^*_1 \leq e^{-\alpha-1}$, and similarly $v^*_1 \in [1-e^{-(\alpha+1)}, 1-e^{-\alpha}]$. By convexity and monotonicity of $\log \frac{1}{x}$ on $[0, 1]$, the difference $\left|\log \left(\frac{1}{u^*_{\alpha+1}}\right) - \log \left(\frac{1}{u^*_{\alpha}}\right)\right|$ shrinks by more than the difference $\left|\log \left(\frac{1}{v^*_{\alpha+1}}\right) - \log \left(\frac{1}{v^*_{\alpha}}\right)\right|$. Consequently, $u^*_1 < u^*_1 < u^*_1 < v^*_1$ holds for sufficiently large $\alpha$.

Furthermore, $e^{-\left(\alpha-1\right)} \to 0$ as $\alpha \to \infty$, and so $\lim_{\alpha \to \infty} x(\alpha) = \frac{1}{2}$.

To find the value of $x(2)$, set $x := x(2)$ and find the critical points of $H(\alpha)(x + \frac{1}{2}, x - \frac{1}{2}) := \frac{1}{2} - x \log^2 \left(\frac{1}{2} - x\right) + \left(x + \frac{1}{2}\right) \log^2 \left(x + \frac{1}{2}\right)$.

Differentiating and factoring, we get
\[
\frac{d}{dx} H(\alpha)(x + \frac{1}{2}, x - \frac{1}{2}) = -\left(\log \left(-x + \frac{1}{2}\right) - \log \left(x + \frac{1}{2}\right)\right) \cdot \left(2 + \log \left(-x + \frac{1}{2}\right) + \log \left(x + \frac{1}{2}\right)\right) = 0.
\]

Now $x = 0$ is a solution which we know is not the maximum and we also get $x = \pm \frac{1}{2}\sqrt{e-2}$, which exactly what [22] conjectured.

2) **Conjecture 10.2:** [22, Conjecture 10.2] posits that for $\pi_1(\alpha), \pi_2(\alpha)$ as above and $\alpha \geq 2$, we have $H(\alpha)(\pi_1(\alpha)) = H(\alpha)(\pi_2(\alpha)) > (\log 2)^\alpha$ and moreover, this quantity is strictly increasing and unbounded as $\alpha \to \infty$.

Since $\log(2) < \frac{\alpha}{2}$, by Proposition 2, $H(\alpha)(\pi_1(\alpha)) = H(\alpha)(\pi_2(\alpha)) > (\log 2)^\alpha$ and unbounded.

3) **Conjecture 10.3:** [22, Conjecture 10.3] posits that $H(\alpha)$ has local maxima for $d > 2$ and $\alpha > 2$. By Lemma 6, the only maxima are the uniform distribution and the $d$ permutations of $\{u, v \in (\alpha^{-1}, \alpha^2) : (1-d)u + v \log^2(v), \text{ should be the outer exist with } v \in \text{ interior of interval.}\}$; so there are either 1 (e.g. $\alpha = d$) or $d + 1$ local maxima.

4) **Conjecture 10.4:** For $d, \alpha \in \mathbb{N}$, define $h_{d, \alpha} := \max_{\mu \in \Delta_N} H(\alpha)(\mu)$. [22, Conjecture 10.4] posits that $h_{d, \alpha} > h_{d, \alpha + 1}$ and that $\lim_{\alpha \to \infty} h_{d, \alpha} = \infty$. In light of the lower bound in Proposition 2, the latter claim (i.e., unboundedness) is immediate.

For $d > e^\alpha$, by Proposition 2, $\max_{\mu \in \Delta_N} H(\alpha)(\mu) \leq 2 \log^\alpha d$ and $\log^{\alpha+1} d \leq \max_{\mu \in \Delta_N} H(\alpha+1)(\mu)$. We find, therefore, that for $d > e^2$, $\max_{\mu \in \Delta_N} H(\alpha)(\mu) \leq \max_{\mu \in \Delta_N} H(\alpha+1)(\mu)$.

But since the conjecture takes interest in the case of $\alpha$ tending to infinity, let us focus on $e^\alpha \geq d - 1$.

By Lemma 6, $u^*_\alpha := \arg \max_{\mu \in \Delta_N} H(\alpha)(\mu)$, is either uniform or takes two distinct values $v \in (e^{-\alpha-1}, 1)$ and $u := \frac{1}{\alpha} e^{-\alpha}$ in $[0, e^{-\alpha-1})$.

For $x \in [0, \frac{1}{2}]$, $\log(1/x) \geq 1$, so $h^{(\alpha+1)}(x) \geq h(\alpha)(x)$. So if $u, v \in [0, \frac{1}{2}]$, then $H(\alpha)(u^*_\alpha) < H(\alpha+1)(u^*_\alpha) \leq \max_{\mu \in \Delta_N} H(\alpha+1)(\mu)$.

So assume instead that $v \in (\frac{1}{2}, 1]$. In this case, we can bound the difference $h(\alpha)(v) - h^{(\alpha+1)}(v) \leq h(\alpha)(v) \leq \frac{1}{2}$.

Since $e^\alpha \geq d - 1$, $u > e^{-\alpha}$ and must lie in $[e^{-\alpha}, e^{-\alpha+1}]$.

But for this entire interval, $x \in [e^{-\alpha}, e^{-\alpha+1}]$ has $h^{(\alpha+1)}(x) - h(\alpha)(x) \geq \frac{1}{2} \geq h(\alpha)(v)$. In order to see this, it suffices, since $h(\alpha)$ and $h^{(\alpha+1)}$ are both decreasing on $[e^{-\alpha}, e^{-\alpha+1}]$.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
to show that $h^{(\alpha+1)}(e^{-(\alpha-1)}) - h^{(\alpha)}(e^{-\alpha}) > \frac{1}{\alpha}$. This can be seen by observing $e^{(\alpha - 1)\alpha+1} - \alpha > e^{\alpha-1}$ for all $\alpha \geq 3$.

It follows, therefore, that for $\alpha \in \mathbb{N}$, when $d > e^\alpha$, for all $d \geq 8$, or for $e^\alpha \geq d - 1$, for all $\alpha \geq 3$, the former claim (monotonicity) holds, i.e.

$$\max_{\mu \in \Delta_d} H^{(\alpha)}(\mu) < \max_{\mu \in \Delta_d} H^{(\alpha+1)}(\mu).$$

(This can also be generalized to any $\beta(\neq \alpha + 1)$ for sufficiently large $d$ or $\alpha$ respectively, if one so desired).

5) Our Conjecture: We close the section with a conjecture of our own.

**Conjecture 1:** For $e^\alpha < d < \infty$, we have $\max_{\mu \in \Delta_d} H_{\alpha}(\mu) = \log \alpha d$ and moreover, the maximum is achieved by the uniform distribution over $[d]$.

**APPENDIX**

A. Comparison of Tail-Vs-Moment Assumptions

Expanding upon the observation of [4] that moment of information assumptions are “weaker (and somewhat more natural)” than tail decay rates, we make some concrete comparisons between the two.

Let us list a number of conditions one might impose $\mu$:

- **$A_1(\alpha)$:** Finite $\alpha$-moment of information [4]:
  
  For some $\alpha > 1$, $\mathbb{E} \left( \log^\alpha \frac{1}{\mu(X)} \right) < \infty$.

- **$B_1(\alpha, M_\alpha)$:** Bounded $\alpha$-moment of information:
  
  For some $\alpha > 1$, $\exists M_\alpha > 0$, $\mathbb{E} \left( \log^\alpha \frac{1}{\mu(X)} \right) < M_\alpha$.

- **$A_2(\beta)$:** Superlinear $\beta$ tail decay: For some $\beta > 1, \mu(i) = \mathcal{O}\left(\frac{1}{i^\beta}\right)$.

- **$B_2(\beta, \bar{\omega}, \bar{\gamma})$:** Controlled superlinear $\beta$ tail decay [4]:
  
  For some $\beta > 1, \bar{\omega}, \bar{\gamma} > 0$ such that $\frac{i^\beta}{\bar{\omega}} \leq \mu(i) \leq \frac{i^\beta}{\bar{\gamma}}$.

- **$A_3(\gamma)$:** Superlinearly $\gamma$-logarithmic tail decay:
  
  For some $\gamma > 1, \mu(i) = \mathcal{O}\left(\frac{1}{\log^\gamma i}\right)$.

- **$B_3(\gamma, \bar{\omega}, \bar{\gamma})$:** Controlled superlinearly $\gamma$-logarithmic tail decay [4]:
  
  For some $\gamma > 1, \mathbb{E}_{\bar{\omega}, \bar{\gamma}} \log^\gamma \mu(i) > 0$ such that $\frac{i^\gamma}{\bar{\omega} \log^\gamma i} \leq \mu(i) \leq \frac{i^\gamma}{\bar{\gamma} \log^\gamma i}$.

**Proposition 4:** The following implications hold:

- (a) $A_3(\gamma), \gamma > 2 \implies A_1(\alpha), \forall \alpha < \gamma - 1$.
- (b) $A_1(\alpha), \alpha > 1 \implies A_3(\gamma), \gamma < \alpha + 1$.
- (c) $B_1(\alpha, M_\alpha) \implies A_1(\alpha)$.
- (d) $A_2(\beta) \implies A_3(\gamma), \forall \gamma > 1$.
- (e) $A_2(\beta) \implies A_1(\gamma), \forall \alpha > 1$.
- (f) $B_2(\beta, \bar{\omega}, \bar{\gamma}) \implies A_2(\beta)$.
- (g) $B_3(\gamma, \bar{\omega}, \bar{\gamma}) \implies A_3(\gamma)$.
- (h) $B_2(\beta, \bar{\omega}, \bar{\gamma}) \implies B_1(\alpha, M_\alpha)$ with $M_\alpha = \max_{\alpha \in \mathbb{N}} (\alpha, \bar{\gamma})$.
- (i) $B_3(\gamma, \bar{\omega}, \bar{\gamma}), \gamma > 2 \implies B_1(\alpha, M_\alpha), \forall \alpha < \gamma - 1$.

**Remark:** We start by noticing that

$$\sum_{i \in \Omega} \frac{\mu(i) \log^\alpha \frac{1}{\mu(i)}}{\mu(i) \leq e^{-\alpha}} \leq \alpha^\alpha + \sum_{\mu(i) \leq e^{-\alpha}} \mu(i) \log^\alpha \frac{1}{\mu(i)}.$$

and that on $[0, e^{-\alpha}]$, it holds that $x \log \frac{1}{x} \frac{1}{\alpha}$ is increasing. Since $\alpha^\alpha$ is finite, the convergence of the series is primarily governed by what happens or small probabilities.

**Proof (c), (d), (f), (g):** Immediate.

**Proof (a):** Suppose that assumption $A_3(\gamma)$ holds. Then $\exists N \in \mathbb{N}, C > 0 \text{ such that for any } i \geq N, \mu(i) \leq C \frac{1}{\log^\gamma i}$. We focus on the rightmost term of (23).

$$\sum_{i > N} \frac{\mu(i) \log^\alpha \frac{1}{\mu(i)}}{\mu(i) \leq e^{-\alpha}} \leq N e^{-\alpha} \alpha^\alpha + \sum_{\mu(i) \leq e^{-\alpha}} \frac{\mu(i) \log^\alpha \frac{1}{\mu(i)}}{\mu(i) \leq e^{-\alpha}} \leq Ne^{-\alpha} \alpha^\alpha + C \sum_{i > N} \frac{1}{\log^\gamma i} \log \frac{1}{i} \leq \frac{1}{\log^\gamma i} \log \frac{1}{C} \log \frac{1}{i} \leq \frac{\log(1 + \log i + \log 1/C^\alpha)}{i \log^\gamma i}.$$

Since $\log i$ dominates both $\log \log i$ and $\log 1/C$, the series converges whenever $\sum_{i \in \mathbb{N}} \frac{1}{\log^\gamma i}$ converges, which occurs exactly for $\gamma > 1$. Let $\alpha > 1$. We decompose the expression of $\mathbb{E} \left( \log^\alpha \frac{1}{\mu(X)} \right)$:

$$\sum_{i \in \mathbb{N}} \frac{\mu(i) \log^\alpha \frac{1}{\mu(i)}}{i \log^\gamma i} = \sum_{i \leq N} \frac{\mu(i) \log^\alpha \frac{1}{\mu(i)}}{i \log^\gamma i} + \sum_{\mu(i) > e^{-\alpha}} \frac{\mu(i) \log^\alpha \frac{1}{\mu(i)}}{i \log^\gamma i} + \sum_{\mu(i) > e^{-\alpha}} \frac{\mu(i) \log^\alpha \frac{1}{\mu(i)}}{i \log^\gamma i} \text{ increasing on } [0, e^{-\alpha}]$$

where

$$\sum_{i \in \mathbb{N}} \frac{1}{\log^\gamma i} \log \frac{1}{\mu(i) \leq e^{-\alpha}} \leq \sum_{\mu(i) \geq e^{-\alpha}} \frac{1}{\log^\gamma i} \log \frac{1}{\mu(i) \leq e^{-\alpha}} \leq S_{\alpha, \beta} \leq \infty$$

at most $e^\alpha$ elements

$$\leq (Ne^{-\alpha} + 1) \alpha^\alpha + C \alpha^\beta S_{\alpha, \beta},$$

such that for any $\alpha > 1$, there exists $M_\alpha < \infty$ that bounds the $\alpha$-moment of information. Notice that although existence is
guaranteed, $N, C$ depend on the unknown $\mu$. The asymptotic nature of assumption $A_2(\beta)$ is therefore not enough to specify what $M_\alpha$ is.

\textbf{Proof (h):} Starting from (23),
\[
\sum_{i \in \Omega} \mu(i) \log^\alpha \frac{1}{\mu(i)} \leq \alpha^+ \sum_{i \in \Omega} \mu(i) \log^\alpha \frac{1}{\mu(i)} \leq \alpha^+ \sum_{i \in \Omega} \frac{\alpha}{\gamma \delta} \log^\alpha \frac{i^\beta}{\gamma},
\]
which is upper bounded by a converging series, whose value is entirely defined by $\alpha, \beta, \gamma, \delta$.

\textbf{Proof (i):} Follows the arguments of the proof for (a) and the proof of (h). The series converges exactly when $\alpha < \gamma - 1$, and if it does, the value of the converging series is a function of $\alpha, \gamma, \delta$.

\textbf{Acknowledgment}

The authors would like to thank Ioannis Kontoyiannis, Igal Sason, and Sergio Verdú for enlightening conversations. They also thank the anonymous referees for helpful suggestions and corrections. The present version of this paper features a sharper bound in Theorem 1, which is now tight, as shown in the (new) Theorem 2. Also added was the discussion on moments vs. tails (Section A) and the complete resolution of the Jürgensen-Matthes conjectures (Section VII-C).

\textbf{References}

[1] J. Acharya, H. Das, A. Orlitsky, and A. T. Suresh, “A unified maximum likelihood approach for estimating symmetric properties of discrete distributions,” in Proc. Int. Conf. Mach. Learn., Aug. 2017, pp. 11–21.

[2] J. Acharya, S. Bhadane, P. Indyk, and Z. Sun, “Estimating entropy of distributions in constant space,” in Proc. 33rd Int. Conf. Neural Inf. Process. Syst., Red Hook, NY, USA: Curran Associates, 2019, pp. 1–12.

[3] A. Antos and I. Kontoyiannis, “Convergence properties of functional estimates for discrete distributions,” Random Struct. Algorithms, vol. 19, nos. 3–4, pp. 163–193, Oct. 2001.

[4] A. Antos and I. Kontoyiannis, “Estimating the entropy of discrete distributions,” in Proc. IEEE Int. Symp. Inf. Theory, Jan. 2001, p. 45.

[5] K. M. R. Audenaert, “A sharp continuity estimate for the von Neumann entropy,” J. Phys. A: Math. Theor., vol. 40, no. 28, pp. 8127–8136, Jul. 2007.

[6] D. Berend and A. Kontorovich, “A sharp estimate of the binomial mean absolute deviation with applications,” Statist. Probab. Lett., vol. 83, no. 4, pp. 1254–1259, Apr. 2013.

[7] D. Berend, A. Kontorovich, and G. Zadkanski, “The expected missing mass under an entropy constraint,” Entropy, vol. 19, no. 7, p. 315, 2017, doi: 10.3390/e19070315.

[8] M. Braitbar and A. Samorodnitsky, “Approximating entropy from sublinear samples,” in Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms (SODA), New Orleans, LA, USA, 2007, pp. 366–375.

[9] D. Brink, “A (probably) exact solution to the birthday problem,” Ramanujan J., vol. 28, no. 2, pp. 223–238, Jun. 2012.

[10] D. Cohen, A. Kontorovich, and G. Wolfers, “Learning discrete distributions with infinite support,” in Proc. Neural Inf. Process. Syst. (NIPS), 2020, pp. 3942–3951.

[11] D. Cohen, A. Kontorovich, A. Koolyk, and G. Wolfers, “Dimension-free empirical entropy estimation,” in Proc. Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 13911–13923.

[12] T. M. Cover, Elements of Information Theory, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[13] K. Fukuchi and J. Sakuma, “Minimax optimal estimators for additive scalar functions of discrete distributions,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jun. 2017, pp. 2103–2107, doi: 10.1109/ISIT.2017.8006900.

[14] K. Fukuchi and J. Sakuma, “Minimax optimal additive functional estimation with discrete distribution: Slow divergence speed case,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jun. 2018, pp. 1041–1045, doi: 10.1109/ISIT.2018.8437255.

[15] S. Golomb, “The information generating function of a probability distribution (Corresp.),” IEEE Trans. Inf. Theory, vol. IT-12, no. 1, pp. 75–77, Jan. 1966.

[16] Y. Han, J. Jiao, and T. Weissman, “Adaptive estimation of Shannon entropy,” in Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jun. 2015, pp. 1372–1376.

[17] Y. Hao and A. Orlitsky, “Data amplification: Instance-optimal property estimation,” in Proc. Int. Conf. Mach. Learn., Jul. 2020, pp. 4049–4059.

[18] Y. Hao, A. Orlitsky, A. T. Suresh, and Y. Wu, “Data amplification: A unified and competitive approach to property estimation,” in Proc. Adv. Neural Inf. Process. Syst., vol. 31, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett, Ed. Red Hook, NY, USA: Curran Associates, 2018, pp. 1–10.

[19] S.-W. Ho and R. W. Yeung, “The interplay between entropy and variational distance,” IEEE Trans. Inf. Theory, vol. 56, no. 12, pp. 5906–5929, Dec. 2010.

[20] J. Jiao, K. Venkat, Y. Han, and T. Weissman, “Maximum likelihood estimation of functionals of discrete distributions,” IEEE Trans. Inf. Theory, vol. 63, no. 10, pp. 6774–6798, Oct. 2017.

[21] J. Jiao, K. Venkat, Y. Han, and T. Weissman, “Minimax estimation of functionals of discrete distributions,” IEEE Trans. Inf. Theory, vol. 61, no. 5, pp. 2835–2885, May 2015.

[22] H. Jürgensen and D. E. Matthews, “Entropy and higher moments of information,” J. Universal Comput. Sci., vol. 16, no. 5, pp. 749–794, 2010.

[23] N. Kasulitsch, “Why the theorem of Scheffé should be rather a theorem of Riesz,” Periodica Math. Hungarica, vol. 61, nos. 1–2, pp. 225–229, Sep. 2010, doi: 10.1007/s10998-010-3225-6.

[24] H. E. Lieb and M. Loss, “Analysis,” in Graduate Studies in Mathematics, vol. 14, 2nd ed. Providence, RI, USA: American Mathematical Society, 2001, doi: 10.1090/gsm/014.

[25] P.-S. Loh, “Convexity,” Carnegie Mellon Univ., Pittsburgh, PA, USA, 2013. [Online]. Available: https://www.math.cmu.edu/~plohp/docs/math/mop2013/convexity-soln.pdf

[26] G. Lugosi and S. Mendelson, “Mean estimation and regression under heavy-tailed distributions: A survey,” Found. Comput. Math., vol. 19, no. 5, pp. 1145–1190, Oct. 2019.

[27] P. Massart, “The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality,” Ann. Probab., vol. 18, no. 3, pp. 1269–1283, Jul. 1990.

[28] L. Paninski, “Estimating entropy on M bins given fewer than M samples,” IEEE Trans. Inf. Theory, vol. 50, no. 9, pp. 2200–2203, Sep. 2004.

[29] L. Paninski, “Estimation of entropy and mutual information,” Neural Comput., vol. 15, no. 6, pp. 1191–1253, Jun. 2003.

[30] I. Pinelis, “Best possible bounds of the von Bahr-Esseen type,” Ann. Probab., vol. 18, no. 3, pp. 1269–1283, Jul. 1990.

[31] I. Pinelis, “Convexity and mutual information,” Neural Comput., vol. 15, no. 6, pp. 1191–1253, Jun. 2003.

[32] J. Pinelis, “Entropy bounds for discrete random variables via maximal coupling,” IEEE Trans. Inf. Theory, vol. 59, no. 11, pp. 7118–7131, Nov. 2013.

[33] H. Scheffe, “A useful convergence theorem for probability distributions,” Ann. Math. Statist., vol. 18, no. 3, pp. 434–438, Sep. 1947.

[34] J. Silva, “Shannon entropy estimation in heavy-tailed distributions: A survey,” in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2015, vol. 20, no. 6, p. 397, May 2018.

[35] S. Verdu, “Empirical estimation of information measures: A literature guide,” Entropy, vol. 21, no. 8, p. 720, Jul. 2019.
[39] B. V. Bahr and C.-G. Esseen, “Inequalities for the \( r \)th absolute moment of a sum of random variables, \( 1 \leq r \leq 2 \),” *Ann. Math. Statist.*, vol. 36, no. 1, pp. 299–303, 1965, doi: 10.1214/aoms/1177700291.

[40] Y. Wu and P. Yang, “Minimax rates of entropy estimation on large alphabets via best polynomial approximation,” *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3702–3720, Jun. 2016.

[41] A. J. Wyner and D. Foster, “On the lower limits of entropy estimation,” *IEEE Trans. Inf. Theory*, 2003, to be published.

[42] Z. Zhang, “Estimating mutual information via Kolmogorov distance,” *IEEE Trans. Inf. Theory*, vol. 53, no. 9, pp. 3280–3282, Sep. 2007.

Doron Cohen received the B.Sc. and M.Sc. degrees in computer science from the Ben-Gurion University of the Negev, Be’er Sheva, Israel, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree.

Aryeh Kontorovich received the bachelor’s degree in mathematics with a certificate in applied mathematics from Princeton University in 2001, and the M.Sc. and Ph.D. degrees from Carnegie Mellon University in 2007. After a Post-Doctoral Fellowship with the Weizmann Institute of Science, he joined the Department of Computer Science, Ben-Gurion University of the Negev, in 2009, where he is currently a Full Professor. He is also working as the Director of the Ben-Gurion University Data Science Research Center. His research interests include machine learning, with a focus on probability, statistics, Markov chains, and metric spaces.

Aaron Koolyk is currently pursuing the Ph.D. degree with The Hebrew University of Jerusalem, Israel.

Geoffrey Wolfer received the Engineering degree from the Ecole Centrale de Nantes, Nantes, France, the M.Sc.Eng. degree from Keio University, Yokohama, Japan, in 2013, and the Ph.D. degree in computer science from the Ben-Gurion University of the Negev, Be’er Sheva, Israel, in 2020. From 2020 to 2022, he was a JSPS International Fellow with the Department of Computer and Information Sciences, Tokyo University of Agriculture and Technology, Tokyo, Japan. He is currently a Special Post-Doctoral Researcher (SPDR) with the Center for AI Project, RIKEN, Tokyo.