Large alphabets: finite, infinite, and scaling models

(Invited Paper)

Mesrob I. Ohannessian
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139
Email: mesrob@gmail.com

Munther A. Dahleh
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, MA 02139
Email: dahleh@mit.edu

***Abstract*—How can we effectively model situations with large alphabets? On a pragmatic level, any engineered system, be it for inference, communication, or encryption, requires working with a finite number of symbols. Therefore, the most straightforward model is a finite alphabet. However, to emphasize the disproportionate size of the alphabet, one may want to compare its finite size with the length of data at hand. More generally, this gives rise to scaling models that strive to capture regimes of operation where one anticipates such imbalance. Large alphabets may also be idealized as infinite. The caveat then is that such generality strips away many of the convenient machinery of finite settings. However, some of it may be salvaged by refocusing the tasks of interest, such as by moving from sequence to pattern compression, or by minimally restricting the classes of infinite models, such as via tail properties. In this paper we present an overview of models for large alphabets, some recent results, and possible directions in this area.**

I. Introduction

As a probabilistic model of discrete data, we can often use a *source* $(\mathcal{A}, p)$ defined as an at most countable set $\mathcal{A}$ which we call the *alphabet*, of size denoted by $m = |\mathcal{A}|$, and a probability distribution p on all finite n-length sequences $x^n := x_1 x_2 \cdots x_n \in \mathcal{A}^n$. We will concern ourselves mostly with stationary memoryless sources, or independent and identically distributed sequences. In this case, with a slight abuse of notation, we let p be a probability distribution on $\mathcal{A}$, and let the sequence probabilities be described by the product distribution $p(x^n) = \prod_{i=1}^n p(x_i)$.

Such models are successfully employed with real world data, be it for file compression, encryption, or text classification. Most of the theory is built based on the assumption that in any pragmatic context the alphabet $\mathcal{A}$ must be finite. This leads to an array of powerful and elegant results in universal compression, entropy estimation, and statistical learning, to name a few. Yet, despite the fact that it is true that any engineered system has to operate with finitely many symbols, one is often faced with situations where the alphabet size is large, so large in fact that some of the asymptotic bounds given by theory do not apply in particular instances. One is then faced with the question: how can we effectively model situations with large alphabets?

Notation: Throughout the paper, we use the notation $f \sim g$ to mean $f/g \to 1$. We also use the subscripts $_p$ and $_{\text{a.s.}}$ to indicate in probability and almost sure convergence respectively, when quantities are random.

II. Finite Models and Entropy Estimation

The simplest model for a large alphabet is one with finitely many elements, i.e. $m < \infty$. The qualifier ‘large’ is meaningful only in a relative sense. Whether the alphabet is too large will often depend on a specific task. Moreover, what is large for a given sequence length n might not be so with an order of magnitude more data. Lastly, if p places very little probability on a vast number of symbols, such a source may have a large alphabet technically, but may be indistinguishable from a smaller alphabet source for the given number of samples.

To account for those various nuisances in assessing the size of the alphabet, the natural approach is to restrict some, and vary others. A common perspective is to fix the task, and assess performance universally over classes of finite sources $(\mathcal{A}, p)$ defined in terms of a compromise between n and m. Given a non-decreasing function $g : \mathbb{N} \to \mathbb{N}$, we can write such a compromise abstractly as a *class of finite sources* indexed by sample size:

$$\mathcal{S}\left(g(n)\right) = \left\{(\mathcal{A}, p) : |\mathcal{A}| = m \leq g(n)\right\}.$$

We illustrate this with some prior work for the task of entropy estimation.

A. Alphabet Size Linear in the Sequence Length

In the context of neuroscience, mutual information estimation with continuous distributions can be performed via discretization. The question is, how finely can one discretize, i.e. how large can the alphabet be, yet permit consistent estimation of the entropy? In [1], Paninski considered the class of sources, for $\gamma > 0$:

$$\mathcal{S}(\gamma n) = \left\{(\mathcal{A}, p) : m \leq \gamma n\right\},$$

which directly encodes a linear relationship between the alphabet size and the number of samples. He then showed, non-constructively, that there exists a single estimator $\hat{H}_n$ for the entropy $H(p)$, such that for any fixed $\gamma > 0$, as $n \to \infty$ (rewording [1], Theorem 1):

$$\sup_{(\mathcal{A},p) \in \mathcal{S}(\gamma n)} \mathbf{E}\left[\left(\hat{H}_n - H(p)\right)^2\right] \to 0. \tag{1}$$

B. Alphabet Size Super-Linear in the Sequence Length

As a corollary to this uniform consistency result, Paninski also showed ([1], Corollary 2) that there exists a sequence $\gamma_n \to \infty$, such that the convergence of (1) holds uniformly over the larger class $\mathcal{S}(\gamma_n n)$. However, γ_n cannot be $\Omega(n^\alpha)$ for any $\alpha > 0$.

In [2], Valiant and Valiant proposed a tighter characterization, showing that, one could essentially choose any $\gamma_n = o(\log n)$. Their original result can be written as follows (by rewording [2], Proposition 1). Given the class:

$$\mathcal{S}(\delta n \log n) = \{(\mathcal{A}, p) : m \leq \delta n \log n\},$$

there exists an estimator $\hat{H}_n$ for the entropy, and a function $h(\delta)$ that is $\mathcal{O}(\sqrt{\delta}|\log \delta|)$ as $\delta \to 0$, such that if $\epsilon > h(\delta)$ then:

$$\sup_{(\mathcal{A},p)\in\mathcal{S}(\delta n \log n)} \mathbb{P}\left\{\left|\hat{H}_n - H(p)\right| > \epsilon\right\} \leq e^{-n^{0.04}}. \qquad (2)$$

In [2], this base result is recast as a *sample complexity* statement, that is: for a given ϵ, how large of a sequence length n is it sufficient to have, as a function of m and ϵ, in order to approximate entropy to within ϵ with high probability.

Here, we instead place Equation (2) in a form that is more directly comparable with the consistency result of Equation (1). In particular, for $\delta_n \downarrow 0$, no matter how slowly, we have that for *all* $\epsilon > 0$, as $n \to \infty$:

$$\sup_{(\mathcal{A},p)\in\mathcal{S}(\delta_n n \log n)} \mathbb{P}\left\{\left|\hat{H}_n - H(p)\right| > \epsilon\right\} \to 0.$$

This reinterpretation shows that Equation (1) is paralleled for $\delta_n = \gamma/\log(n)$. However, we have more leeway, e.g. by choosing $\delta_n = 1/\log\log(n)$, or any $\delta_n = o(1)$.

Valiant and Valiant also establish an effective lower bound ([2], Corollary 2), that suggests that this $\gamma_n = o(\log n)$ rate is tight. Furthermore, their approach works for more general functionals of the distribution, as long as they are symmetric with respect to relabeling, and continuous with respect to a metric (technically on the multiset of, or sorted, probabilities) that they refer to as *relative earthmover distance*, defined as:

$$R(p,q) = \sup_{f:|f'(x)|\leq 1/x} \left|\sum_{a\in\mathcal{A}} p(a) f(p(a)) - \sum_{a\in\mathcal{A}} q(a) f(q(a))\right|.$$

Shannon entropy is a special case, it corresponds to one of the sums with $f(x) = -\log(x)$, which satisfies $|f'(x)| \leq 1/x$. Therefore, we immediately have the (Lipschitz) continuity $|H(p) - H(q)| \leq R(p,q)$.

III. Scaling Models

Finite models for large alphabets, such as the ones we presented, and which are described as uniform consistency over classes of finite sources are adequate for directly capturing alphabet size versus sample size compromises. They do not place any other constraints on the source, and do not allow statements for particular sample sequences. To extend such characterizations, one can use models that achieve description of large alphabets by a sequential scaling, similarly to Kolmogorov asymptotics in high-dimensional statistics.

We define a *scaling model* as a sequence of sources:

$$\{(\mathcal{A}_n, p_n)\}_{n=1,2,\cdots},$$

in addition to a property of the sequence. It is understood that the sample sequence of length n is drawn from p_n. We denote this random sequence by $X_{n,1}\cdots X_{n,n}$. The property may be defined as pointwise constraints, as in the finite models, but may more generally be a sequential property.

To illustrate some of the benefits of this perspective, let us simply use the pointwise constraint $(\mathcal{A}_n, p_n) \in \mathcal{S}(n \log\log n)$. By (2), it follows that for any $\epsilon > 0$, we have a summable convergence in probability, and therefore by the Borel-Cantelli lemma, we have:

$$\left|\hat{H}_n - H(p_n)\right| \to_{\text{a.s.}} 0, \qquad (3)$$

which is a statement that is less elegant to formalize outside of the scaling framework. Furthermore, this shows that finite models are subsumed by scaling models. We can indeed think of a scaling model as defining an appropriate probability space using a sequence of classes of finite sources, as suggested for example in [1].

A. Rare-Events Sources

More generally, it is convenient to work with scaling models directly and to use sequential properties. This is done notably by Wagner, Viswanath, and Kulkarni, whom we refer to as WVK henceforth, in [3]. Let P_n be the law of $np_n(X_{n,1})$, i.e. the following probability measure on $\mathbb{R}^+$:

$$P_n(\mathrm{d}x) = \sum_{a\in\mathcal{A}_n} p_n(a)\delta_{np_n(a)}(\mathrm{d}x), \qquad (4)$$

where δ_x is a Dirac delta at x. WVK call this measure the 'shadow', and they define a *rare-events source* as a scaling model where the shadows converge weakly:

$$P_n \Rightarrow P, \qquad (5)$$

for some probability measure P on $\mathbb{R}^+$.

This definition focuses on the probability distributions p_n. Unlike finite models and the example we gave (preceding Equation (3)) that uses a finite model as a constraint, this scaling model construction does not impose that every alphabet $\mathcal{A}_n$ be finite. Yet, the weak convergence of Equation (5) does capture a situation where the effective alphabet size, where most symbols lie, is roughly linear in the sample size. We make this precise in the following proposition.

Proposition 1. *Consider a rare-events source $\{(\mathcal{A}_n, p_n)\}$ with limiting distribution P such that $P(0) = 0$. Then for any choice of $\check{c}_n \downarrow 0$ and of $\hat{c}_n \uparrow \infty$, the subsets $\mathcal{B}_n = \{a \in \mathcal{A}_n : \check{c}_n \leq np_n(a) \leq \hat{c}_n\}$ are an effective alphabet, in that:*

$$p_n(\mathcal{B}_n) \to 1.$$

Furthermore, we have that

$$o(n) = \frac{np_n(\mathcal{B}_n)}{\hat{c}_n} \leq |\mathcal{B}_n| \leq \frac{np_n(\mathcal{B}_n)}{\check{c}_n} = \omega(n).$$

Proof: Let $\mathcal{D}_n = \{a \in \mathcal{A}_n : np_n(a) < \check{c}_n\}$. Choose $\epsilon > 0$, and find n_0 such that for all $n > n_0$, $\check{c}_n < \epsilon$. It follows that for all $n > n_0$:

$$p_n(\mathcal{D}_n) \leq \int_0^\epsilon \mathrm{d}P_n.$$

Therefore $\limsup_{n\to\infty} p_n(\mathcal{D}_n) \leq \int_0^\epsilon \mathrm{d}P$, and since ϵ was arbitrary, $p_n(\mathcal{D}_n) \to 0$. A similar argument shows that $p_n(\{a \in \mathcal{A}_n : np_n(a) > \hat{c}_n\}) \to 0$. The bounds on the cardinality of $\mathcal{B}_n$ are straightforward. ■

The fact that all the symbols in this effective alphabet have probability roughly on the order of $\frac{1}{n}$ justifies the name 'rare-events'.

B. Rare Probability Estimation

WVK showed that in the regime of a rare-events source one can consistently estimate the total probabilities of rare events. To make this more precise, denote by $B_{n,r}$ the subsets of the alphabet of symbols that appear exactly r times in the sample sequence. In general, we think of r as much smaller than the sample size n, but in the rare-events regime this is true for all r. Note that the case $r = 0$ corresponds to the subset of symbols which do not appear in the sequence. We then define the *rare probabilities* as:

$$M_{n,r} := p_n(B_{n,r}).$$

In particular, $M_{n,0}$ denotes the missing mass, the probability of unseen symbols. The estimation of $M_{n,r}$ is sometimes called the *Good-Turing estimation problem*, in reference to the pioneering work of [4], who gives due credit to Turing. Their solution to this estimation problem, the *Good-Turing estimator*, is:

$$G_{n,r} := \frac{(r+1)K_{n,r+1}}{n}, \tag{6}$$

where $K_{n,r} := |B_{n,r}|$ is the number of distinct symbols appearing exactly r times in the sample. The study of the numbers $K_{n,r}$ themselves, and the number $K_n := \sum_{r\geq 1} K_{n,r}$ is known as the occupancy problem, in reference to the classical urn schemes.

In the rare events regime, WVK show in [3] that both $M_{n,r}$ and $G_{n,r}$ satisfy the same Poisson limit, in a strong sense:

$$M_{n,r} \sim_{\text{a.s.}} G_{n,r} \sim_{\text{a.s.}} \int_{\mathbb{R}^+} \frac{x^r e^{-x}}{r!} \mathrm{d}P(x). \tag{7}$$

In fact, L_1 convergence also follows. This behavior is particularly interesting, considering that the empirical distribution does not converge in the same way as it does in the finite setting. However, the restriction to particular subsets $B_{n,r}$ essentially regularizes the problem, and makes results such as (7) possible.

C. Entropy and General Functional Estimation

Using simply the weak convergence constraint of Equation (5), entropy estimation is rather ill-posed because of very small probabilities. To remedy the problem, and in light of Proposition 1, WVK further constrain the rare-events source to satisfy, for all n:

$$\check{c} \leq np_n(\mathcal{A}_n) \leq \hat{c}, \tag{8}$$

where $\check{c}$ and $\hat{c}$ are unknown constants. These constraints place the scaling model closer to the finite setting, since it follows that:

$$\frac{n}{\hat{c}} \leq |\mathcal{A}_n| \leq \frac{n}{\check{c}}. \tag{9}$$

WVK give an explicit construction of entropy in this particular framework, using the strong consistency of the Good-Turing estimates, and appropriate power series expansions.

In [5], we showed that it is possible to use the Good-Turing estimator differently, by casting it as a pseudo-empirical measure issued from the Poisson mixture of Equation (7). We can then use mixing density estimation techniques to construct a measure:

$$\tilde{P}_n(X_{n,1} \cdots X_{n,n}) \Rightarrow_{\text{a.s.}} P.$$

This in itself is sufficient for estimating bounded continuous functionals of the form:

$$Y_n := \int f(x)\ \mathrm{d}P_n(x),$$

which converges to $\int f(x)\ \mathrm{d}P(x)$ by weak convergence. If the functional is generally unbounded, but bounded on $[\check{c}, \hat{c}]$ in the restricted model of Equation (8), then if a bound D is known for the interval, as in $[\check{c}, \hat{c}] \subset D$, then estimation can be performed by tapering f:

$$f_D(x) \equiv \begin{cases} f(D^{-1}) & x < D^{-1}, \\ f(x) & x \in [D^{-1}, D], \\ f(D) & x > D. \end{cases}$$

More precisely, ([5], Lemma 2):

Proposition 2. *If $\tilde{P}_n \Rightarrow_{\text{a.s.}} P$ as $n \to \infty$ and $[\check{c}, \hat{c}] \subset D$, then*

$$\hat{Y}_n := \int_{\mathbb{R}^+} f_D(x) \mathrm{d}\tilde{P}_n(x)$$

in a consistent estimator of Y_n, in the sense that

$$\left|\hat{Y}_n - Y_n\right| \to_{\text{a.s.}} 0.$$

For example, entropy is of the form:

$$H(p_n) = \int -\log(x)\, \mathrm{d}P_n(x) + \log n,$$

and since the $\log n$ term is deterministic, we can apply Proposition 2 to estimate entropy consistently, given a construction of $\tilde{P}_n$ and a bound D. When the functional is generally unbounded, and no bounds are known for $[\check{c}, \hat{c}]$, then estimation is still possible in this framework provided that we can characterize the rate of the convergence $\tilde{P}_n \Rightarrow P$, say in terms of the Wasserstein metric.

Lastly, it is worth noting that we can also relate this scaling framework to the finite setting, as suggested by (9). In particular, we can once again use the Valiant and Valiant estimator leading to Equation (2), in order to obtain the same consistency result as in Equation (3).

IV. Universal Compression and Rare Probability Estimation in Infinite Models

It is also possible to idealize large alphabets as being infinite, but fixed. We first consider the problem of universal compression in this setting, and then outline some results in the related problem of probability estimation.

We recall the relevant definitions in universal compression. The *worst case redundancy* for n-length sequences over a class $\mathcal{S}$ of memoryless sources is defined as:

$$\mathcal{R}(n,\mathcal{S}) = \inf_q \sup_{(\mathcal{A},p)\in\mathcal{S}} \sup_{x^n\in\mathcal{A}^n} \log \frac{p(x^n)}{q(x^n)},$$

which quantifies within how many nats any single algorithm can compress a source in $\mathcal{S}$, as compared to the compression of an algorithm designed with exact knowledge of the source. If, as $n \to \infty$,

$$\frac{1}{n}\mathcal{R}(n,\mathcal{S}) \to 0,$$

we say that the per-symbol redundancy vanishes, which is the notion of universality that we seek: the source is asymptotically compressed as though it were known.

When the alphabet size is bounded, universality is immediate. But for infinite sources without further constraints, there are many negative results that have stymied research in this area. In particular, in [6], Kieffer defined a weaker notion of universality, which he showed not to hold for general stationary and ergodic sources. Weak universality was later shown to fail even for memoryless sources by Györfi, Páli and van der Meulen in [7]. This means that, for general $\mathcal{S}$, we cannot expect to have vanishing per-symbol redundancy.

A. Universal Compression of Patterns

In [8], Orlitsky, Santhanam, and Zhang, whom we refer to as OSZ henceforth, recover universality results by shifting attention from the compression of sequences to the compression of patterns. More precisely, given a sequence $x^n = x_1 \cdots x_n$, we define its *pattern* to be a sequence of positive integers:

$$\Psi(x^n) = \iota(x_1)\cdots\iota(x_n),$$

where $\iota(\cdot)$ is a sequence-dependent function that maps each symbol of the alphabet that appeared in the sequence to its index of first appearance.

In analogy to the sequence case, we can define the worst case redundancy for n-length *patterns* over a class $\mathcal{S}$ of memoryless sources as:

$$\mathcal{R}_\Psi(n,\mathcal{S}) = \inf_q \sup_{(\mathcal{A},p)\in\mathcal{S}} \sup_{x^n\in\mathcal{A}^n} \log \frac{p(\Psi(x^n))}{q(\Psi(x^n))},$$

where we have restricted our attention on the induced probability over patterns. To emphasize the symmetry over Ψ, it is possible to write this redundancy after transforming all quantities to pattern-specific form, as is done in [8]. The surprising result of OSZ was that, without further restrictions on $\mathcal{S}$, one has vanishing per-symbol pattern redundancy:

$$\frac{1}{n}\mathcal{R}_\Psi(n,\mathcal{S}) \to 0,$$

for any $\mathcal{S}$, even with alphabets that are infinite. Furthermore, it is possible to achieve this in a computationally efficient manner, albeit at a suboptimal rate.

It is worth drawing some parallels with probability estimation in the rare-events regime. Indeed, the work of WVK was spawned by questions arising from OSZ. In particular, the type or empirical measure does not converge in the rare-events regime, which is tantamount to the impossibility results in universal compression. Yet, when attention is restricted to the probability of the subsets $B_{n,r}$, consistent estimation becomes possible, as evidenced by Equation (7). Indeed, the occupancy numbers $K_{n,r} = |B_{n,r}|$, each referred to as the *prevalence* of multiplicity r, and collectively as the *profile* of the pattern in [8], are intimately related to the pattern itself. In particular any two patterns with the same profile have the same probability, just as any two sequences with the same type do.

B. Universal Compression with Tail Properties

Although universal compression of sequences is not possible for general infinite alphabets, it is possible to restrict $\mathcal{S}$ and obtain universality within the restricted class. Kieffer [6] gives certain conditions for weak universality. More recently, in [9], Boucheron, Garivier, and Gassiat studied universal compression over classes of infinite sources defined by envelope conditions. These conditions effectively model tail behavior. Of particular interest is the result of vanishing per-symbol redundancy for power-law and exponential tails. We do not elaborate on their results, but rather use it as prelude to the following section.

C. Rare Probability Estimation with Heavy-Tails

In the scaling model of the rare-events sources proposed by WVK, we have seen that the rare probability estimation problem is strongly consistent,by Equation (7), and this is in fact achieved by the Good-Turing estimator of (6). The question we ask is therefore, given a fixed arbitrary (possibly infinite-alphabet) source $(\mathcal{A},p)$ from a class $\mathcal{S}$, when is it possible to estimate the rare probabilities $M_{n,r}$ consistently?

It is important to note that for a fixed source, $\mathbf{E}[M_{n,r}] \to_{\text{a.s.}} 0$, for fixed $r \geq 0$, thus $M_{n,r} \to_p 0$ and talking about consistency using additive error is not meaningful. Therefore we define consistency multiplicatively. We call an estimator $\hat{M}_{n,r}$ *strongly consistent* over a class $\mathcal{S}$ of sources, if for every $(\mathcal{A},p)\in\mathcal{S}$:

$$\hat{M}_{n,r} \sim M_{n,r}, \text{ as } n \to \infty.$$

where the convergence is either in probability or almost surely.

Such strong consistency is not to be taken for granted. In fact, despite the various additive error and concentration results, such as [10], [11], [12], [13], the Good-Turing estimator is not strongly consistent over arbitrary classes of sources $\mathcal{S}$. In particular, we show in [14] (Lemma 3.2) that $G_{n,0}$, the missing mass estimator, is not strongly consistent even for a geometric distribution. This is perhaps related to the redundancy gap that is observed in the work of OSZ, when using predictors based on the Good-Turing estimator.

Going back to the task of identifying sufficient conditions for estimating rare probabilities, we focus on characterizing classes of distributions $\mathcal{S}$ based on a tail property. We use Karamata's theory of regular variation [15], suggested originally by Karlin [16]. We first introduce the following counting measure on $[0,1]$, using the notation of [17]:

$$\nu(\mathrm{d}x) := \sum_{a\in\mathcal{A}} \delta_{p(a)}(\mathrm{d}x).$$

(An almost identical function is referred to as the 'histogram of a distribution' in [2], and it is also related to the 'shadow' of Equation (4)). Using ν, we define the following function, which is a cumulative count of all symbols having no less than a certain probability mass:

$$\vec{\nu}(x) := \nu[x,1].$$

Definition. *Following [16], we say that $(\mathcal{A},p)$ is* regularly varying *with* regular variation index $\alpha \in (0,1)$*, if the following holds:*

$$\vec{\nu}(x) \sim x^{-\alpha}\ell(1/x), \quad \text{as} \quad x \downarrow 0,$$

where $\ell(t)$ is a slowly varying function, in the sense that for all $c>0$, $\ell(ct)\sim\ell(t)$ as $t\to\infty$.

This definition can be extended to the cases of $\alpha = 0$ and $\alpha = 1$, as is done in [17] for example. Regular variation is thus an elegant way of describing tail regularity in discrete distributions. For the case $\alpha \in (0,1)$, it captures the heavy-tailed property rather generally, without restriction to pure power laws such as the Zipf-Mandelbrot distributions. Let's denote the collection of all such regularly varying heavy-tailed distributions by $\mathcal{S}_{\mathrm{RV}}$. We then have the following result (reworded from [14], Theorem 3.15):

Proposition 3. *For any $(\mathcal{A},p)$ in $\mathcal{S}_{\mathrm{RV}}$, the Good-Turing estimator is strongly consistent:*

$$G_{n,r} \sim_{\mathrm{a.s.}} M_{n,r}, \text{ as } n\to\infty.$$

Proof: The proof relies on a new multiplicative concentration result under regularly varying heavy tails ([14], Theorem 3.13), from which one can deduce strong laws for the rare probabilities: $M_{n,r} \sim_{\mathrm{a.s.}} \mathbf{E}[M_{n,r}]$. It is easy to show that (and is in fact the basis of the Good-Turing estimator that):

$$\mathbf{E}[M_{n,r}] = \frac{r+1}{n+1}\mathbf{E}[K_{n+1,r+1}].$$

Compounding these with Karlin's (see [16] and [17]) strong laws for the occupancy numbers, which establish that $K_{n,r} \sim_{\mathrm{a.s.}} \mathbf{E}[K_{n,r}]$ and $\mathbf{E}[K_{n,r}] \sim \mathbf{E}[K_{n+1,r}]$, the strong consistency result follows. ■

It is interesting that heavy-tailed sources are natural models for both infinite alphabet universal compression and rare event probability estimation. After identifying $\mathcal{S}_{\mathrm{RV}}$ as a sufficient regime in which strongly consistent rare probability estimation is possible, one can develop estimators that go beyond that of Good and Turing, as we do in [14]. In particular, regular variation opens the platform to techniques from extreme value theory (see [18]), which have been successfully used in rare event estimation in continuous settings.

V. Conclusion

In this paper we surveyed various approaches (by no means exhaustive) to modeling large alphabets, striving to present them from a single vantage point. We highlighted particular estimation and communication problems, and showed three main modeling paradigms: finite, scaling, and infinite.

What would be arguably most useful is a clear understanding and encoding of the various compromises that are at play: the task, the sample size, the size of the alphabet if it is finite, or distribution tail properties if it is infinite. A potentially fruitful approach would be to mutually map various large alphabet models, show whether statements for certain tasks are necessary or sufficient for others as well, and develop efficient algorithms for simultaneous large data and alphabet regimes.

The importance of relating these various models is in clarifying the fundamental limits of the very pragmatic question of how much one can do when working with alphabets that are disproportionate in size to data.

References

[1] L. Paninski, "Estimating entropy on m bins given fewer than m samples," *IEEE Trans. on Information Theory*, vol. 50, no. 9, pp. 2200–2203, 2004.

[2] G. Valiant and P. Valiant, "Estimating the unseen: an $n/log(n)$-sample estimator for entropy and support, shown optimal via new CLTs," in *43rd ACM Symposium on Theory of Computing, STOC*, 2011, pp. 1840–1847.

[3] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Probability estimation in the rare-events regime," *IEEE Trans. on Information Theory*, vol. 57, no. 6, pp. 3207–3229, 2011.

[4] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 16, pp. 237–264, 1953.

[5] M. I. Ohannessian, V. Y. F. Tan, and M. A. Dahleh, "Canonical estimation in a rare-events regime," in *49th Allerton Conference on Communication, Control, and Computing, Allerton*, 2011, pp. 679–682.

[6] J. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. on Information Theory*, vol. IT-24, no. 6, pp. 674–682, 1978.

[7] L. Györfi, I. Páli, and E. van der Meulen, "A unified approach to weak universal source coding," *IEEE Trans. on Information Theory*, vol. 40, no. 1, pp. 267–271, 1994.

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

[9] S. Boucheron, A. Garivier, and E. Gassiat, "Coding on countably infinite alphabets," *IEEE Trans. on Information Theory*, vol. 55, no. 1, pp. 358–373, 2009.

[10] H. E. Robbins, "Estimating the total probability of the unobserved outcomes of an experiment," *Annals of Mathematical Statistics*, vol. 39, no. 1, pp. 256–257, 1968.

[11] W. W. Esty, "A normal limit law for a nonparametric estimator of the coverage of a random sample," *The Annals of Statistics*, vol. 11, no. 3, pp. 905–912, 1983.

[12] D. McAllester and R. E. Schapire, "On the convergence rate of Good-Turing estimators," in *13th Annual Conference on Computational Learning Theory*, 2000.

[13] D. McAllester and L. Ortiz, "Concentration inequalities for the missing mass and for histogram rule error," *Journal of Machine Learning Research*, vol. 4, pp. 895–911, 2003.

[14] M. I. Ohannessian, "On inference about rare events," Ph.D. dissertation, Massachusetts Institute of Technology, 2012.

[15] J. Karamata, "Sur un mode de croissance régulière. Théorèmes fondamenteaux," *Bulletin de la Société Mathématique de France*, vol. 61, pp. 55–62, 1933.

[16] S. Karlin, "Central limit theorems for certain infinite urn schemes," *J. of Mathematics and Mechanics*, vol. 17, no. 4, pp. 373–401, 1967.

[17] A. Gnedin, B. Hansen, and J. Pitman, "Notes on the occupancy problem with infinitely many boxes: general asymptotics and power laws," *Probability Surveys*, vol. 4, pp. 146–171, 2007.

[18] J. Beirlant, Y. Goegebeur, J. Segers, and J. Teugels, *Statistics of extremes: theory and applications*. Wiley, 2004.