Probability Estimation in the Rare-Events Regime

Aaron B. Wagner, Pramod Viswanath, and Sanjeev R. Kulkarni

Abstract

We address the problem of estimating the probability of an observed string that is drawn i.i.d. from an unknown distribution. Motivated by models of natural language, we consider the regime in which the length of the observed string and the size of the underlying alphabet are comparably large. In this regime, the maximum likelihood distribution tends to overestimate the probability of the observed letters, so the Good-Turing probability estimator is typically used instead. We show that when used to estimate the sequence probability, the Good-Turing estimator is not consistent in this regime. We then introduce a novel sequence probability estimator that is consistent. This estimator also yields consistent estimators for other quantities of interest and a consistent universal classifier.

I. Introduction

Existing research on probability estimation and lossless compression focuses almost exclusively on one asymptotic regime: the source alphabet and probability distribution are fixed, even if they are unknown, and the amount of data is permitted to tend to infinity. This mathematical scaling captures the practical scenario in which enough data is available so that every possible source symbol is observed many times. This practically-important asymptotic facilitates the use of typicality techniques that are the cornerstone of results in information theory. An analogous asymptotic is often employed when studying reliable communication over noisy channels.

Despite the ubiquity and success of this approach, however it is not appropriate in all situations. In particular, it harbors an implicit assumption that may fail in practice. If the probability distribution that generates a discrete memoryless source is held fixed, and the amount of observed data (i.e., the block length) is allowed to tend to infinity, then asymptotically the distribution can be estimated perfectly from the data, even if nothing about the distribution is known *a priori*. In short, the conventional asymptotic tacitly assumes that the distribution can be learned from the data. In practice, however, this may not be the case.

A. Modeling Natural Language

Consider, for example, what is arguably the most fundamental of data sources, natural language. Any realistic model of natural language must capture the statistical dependence among nearby letters. Perhaps the simplest model is to assume that this dependence is Markovian. To be concrete, one could assume that each English letter depends only on the previous three letters, and then estimate the transition probabilities from a large text corpus. One realization of this source model is the following [1, p. 109]

```
The generated job providual better trand the displayed
code, abovery upondults well the coderst in thestical it
do hock bothe merg.
```

Quantitative measures show that that even this third-order Markov approximation is inadequate: the entropy of this source is 2.8 bits per symbol [1, p. 111], while real English (ignoring punctuation and case) is estimated to have an entropy of 1.3 bits per symbol [2] or lower [3, p. 50] [1, p. 138]. This model evidently fails to capture much of the structure of the language.

School of Electrical and Computer Engineering, Cornell University. Email: wagner@ece.cornell.edu.

Department of Electrical and Computer Engineering and Coordinated Science Laboratory, University of Illinois at Urbana-Champaign. Email: pramodv@uiuc.edu.

Department of Electrical Engineering, Princeton University. Email: kulkarni@princeton.edu.

This could be rectified by using a higher-order Markov model, but higher-order models present their own difficulties. Intuition suggests that the order of the Markov model would need to be about 10 letters to capture most of the dependence present in the language. But a 10th-order Markov model over an alphabet of size 100 has 100^{11} different transition probabilities that must be learned, however, which would require many exabytes of data[1]. Since text-processing systems usually operate on much smaller documents, it is unreasonable to assume that they can learn the underlying distribution from the observed data. Thus the conventional asymptotic is not appropriate for English text.

Of course, many 10-tuples of characters never occur in real English, so the 10th-order transition probability matrix has significant sparsity that lowers its effective dimension. The most convenient way of harnessing this sparsity is to use *words* as the atomic source symbols. But since there are several hundred thousand English words, even a first-order word-level Markov model still requires several gigabytes of text to learn completely. Thus the conventional asymptotic is still inappropriate.

The poor fit between the usual information-theoretic models and natural language data can also be seen from word-count studies in linguistics. Under the conventional asymptotic, a string of source symbols with length n will be dominated by $\Theta(1)$ different symbols each appearing $\Theta(n)$ times. For natural language data, on the other hand, the set of words that appear order-n times[2] collectively comprise only about one quarter to one third of the overall document [4, p. 9]. Length-n documents are therefore dominated by words that appear fewer than order-n times.

The frequency with which words appear was studied by Zipf [5], who found that in a particular document, the kth most frequent word tended to have an empirical probability proportional to $1/k^p$, where p is near one. But later work [4, Sec. 1.3] has shown that the Zipf distribution and its improvements [6], [7], [8] misfit data in a systematic way. More importantly for us, if we assume the source has a Zipf-like distribution and let the sample size n tend to infinity, then the source will emit strings which are dominated by $\Theta(1)$ different symbols each appearing $\Theta(n)$ times. Evidently the problem is not with the choice of the source distribution, it is with the asymptotic itself.

B. Contributions of this Paper

We study probability estimation using a different scaling model that is better suited to natural language applications. We treat words as the atomic source symbols, so that the "alphabet" is large but intersymbol dependence is slight. In fact, we neglect intersymbol dependence in the present work and model words as drawn i.i.d. from a large alphabet. Rather than using a fixed distribution, however, and letting the block length tend to infinity, we suppose that the alphabet size and the source distribution *both* scale with the block length. In particular, we assume that the source that generates the length-n string has an alphabet of size $\Theta(n)$ with each symbol having probability $\Theta(1/n)$. This asymptotic allows us to focus on the regime of practical interest, in which the length of the document and the number of distinct words appearing within it are comparably large. We call this the *rare-events* regime.

In the rare-events regime, the empirical distribution does not converge to the true distribution as the block length tends to infinity. This fact makes even some basic probability estimation problems nontrivial. Yet, we show in this paper that many important quantities can be estimated consistently from the observed data. Specifically, we show that in the rare-events regime, when the underlying distribution is unknown,

- the total (sum) probability of the set of symbols appearing k times in the sequence can be consistently estimated from the observed sequence for each k, using the Good-Turing estimator [9],
- the normalized probability of the observed sequence can be consistently estimated from the sequence itself,
- the normalized entropy of the source can be consistently estimated from the sequence itself,

[1]The number of distinct English characters, including upper and lower case and punctuation symbols, is approximately 100.

[2]The set of these words is called the *linguistically closed class.*

- the relative entropy between the true and empirical distributions can be estimated consistently from the observed sequence, as can the relative entropy between the true distribution and the uniform distribution over the symbols that appear k times, for each k,
- the normalized probability of one sequence under the distribution that generated a second sequence can be consistently estimated, as can the relative entropy between the two distributions, using only the pair of sequences, and
- consistent universal hypothesis testing is possible.

Improved models and probability estimation techniques for natural languages could lead to better algorithms for text compression, optical character recognition, speech recognition, author identification, and subject classification. It should be added that improved text compression increases the security of cryptographic techniques in addition to reducing storage and transmission requirements, since any redundancy present in the plaintext message makes cryptographic schemes easier to compromise. Although we are motivated by problems in natural language processing, the rare-events model is also applicable in other application areas, such as digital video and audio, where pixels and samples, respectively, play the role of words.

C. Connections to the Literature

Large-alphabet sources are known to present special challenges in information theory. Kieffer [10] provided necessary and sufficient conditions for whether a class of sources is universally compressible under the conventional asymptotic. He noted that his result implies that the class of stationary and ergodic sources over an infinite alphabet is not universally compressible, in contrast to the finite-alphabet case. The impetus for the present paper came from the more-recent work of Orlitsky et al. [11], [12]. These authors called attention to the discrepancy between the asymptotic that is typically used in information theory and many realistic data sources, and showed that a function of the observed sequence called the *pattern* can be universally compressed, even when the alphabet is infinite and the source distribution is arbitrary. Our focus is on estimation instead of compression, and we exploit the rare-events structure of natural language instead of assuming an arbitrary source distribution.

The source model that we study is essentially the "Large Number of Rare Events" (LNRE) model introduced by Khmaladze [13] (see also [14]). By studying word counts in large documents, Baayen [4] argues that the LNRE model is well suited to natural language sources. Detection and estimation problems involving LNRE sources have been considered [15], [16], but these have not addressed the estimation of important information-theoretic quantities such as sequence probabilities, entropies, and divergences. Paninski [17] proves the existence of a consistent entropy estimator in a similar regime. This paper provides a constructive demonstration of such an estimator; moreover, we provide explicit estimators for a range of other important quantities.

Generating symbols from an i.i.d. source is equivalent to dropping balls into bins: the bins represent the source symbols and the balls represent positions in the string. Thus this work is connected to the extensive literature on random allocations and occupancy problems (e.g. [18], [19]). In the terminology of balls and bins, our asymptotic amounts to assuming that the number of balls and bins tend to infinity at the same rate. There is a literature on random allocations in this regime (see [18], [19], [20], [21] and the references therein), but it focuses mainly on central limit and large deviations characterizations of the number of bins containing a given number of balls. It does not address the information-theoretic questions studied here.

In collaboration with Turing, Good [9] introduced a probability estimator that turns out to be well-suited to the rare-events regime. Good was motivated by the problem of estimating the probability of a symbol selected randomly from the set of symbols appearing k times in the string, for a given k. Good motivates the Good-Turing estimator via a calculation of its bias; other early theoretical work on the Good-Turing estimator also focused on its bias [22], [23]. Recent work has been directed toward developing confidence intervals for the estimates using central limit theorems [24], [25] and concentration inequalities [26],

[27]. Orlitsky et al. [12] studied what they call the *pattern redundancy* of the Good-Turing estimator and showed that it is near optimal but can be improved. None of these works, however, has shown that the estimator is consistent.

We show that the Good-Turing estimator is consistent for rare-event sources[3]. We consider the problem of estimating the total probability of all symbols that appear k times in the observed string for each nonnegative integer k. For $k = 0$, this is the total probability of the unseen symbols, a quantity that has received particular attention [22], [28], [29]. Estimating the total probability of all symbols with the same empirical frequency is a natural approach because these symbols cannot be distinguished using the observed data. Although the total probabilities are themselves random, we show that in the rare-events regime, they converge to a deterministic limit, which we characterize. Note that if the alphabet was small and the block length was large, then estimating the total probabilities would reduce to estimating the probability of the individual symbols because it is unlikely that multiple symbols will have the same empirical frequency.

D. Outline

The rare-events model is described in detail in the next section. In Section III, we discuss the Good-Turing total probability estimator and show that it is consistent for rare-events sources. Section IV shows how the Good-Turing estimator can be used to consistently estimate a number of other important quantities, including the probability of the observed sequence and the entropy of the source. In Section V, we extend the rare-events model to pairs of sources, and in Section VI we show that an extension of the Good-Turing total probability estimator can be used to consistently estimate the probability of one sequence under the distribution that generated a second sequence. This result has implications for universal hypothesis testing in the rare-events regime, which are discussed in Section VII. Finally, in Section VIII, we study the finite-n behavior of our estimators via simulation. The proofs of the results in Sections II through VI are given in Appendices A through E, respectively, with the exception of those that are brief.

II. The Rare-Events Model

Let A_n be a sequence of finite alphabets. For each n, let p_n be a probability distribution on A_n satisfying

$$\frac{\check{c}}{n} \leq p_n(a) \leq \frac{\hat{c}}{n} \tag{1}$$

for all $a \in A_n$, where $\check{c}$ and $\hat{c}$ are fixed positive constants that are independent of n. For each n, we observe a random string $\mathbf{X}$ of length[4] n drawn i.i.d. from A_n according to p_n. We abuse notation slightly and use X_n to refer to a generic random variable with distribution p_n and X_i to refer to the ith variable from $\mathbf{X}$.

Note that both the alphabet and the underlying distribution vary with n. Note also that by assumption (1), each element of A_n has probability $\Theta(1/n)$ and thus will appear $\Theta(1)$ times on average in the string. In fact, for any fixed k, the probability of any given symbol appearing k times in the string is bounded away from 0 and 1 as $n \to \infty$. In words, every letter is *rare*. The number of distinct symbols in the string will grow linearly with n as a result. While there are other, less restrictive ways of requiring that all symbols appear "rarely," or not at all, the condition in (1) is particularly useful [19, p. 6]. We do not assume that p_n or even the constants $\check{c}$ and $\hat{c}$ are known.

Our focus will be on quantities such as $p_n(\mathbf{X})$ and $H(p_n)$ that are invariant under a relabeling of the symbols in A_n. It is therefore convenient to consider the multiset of probabilities assigned by p_n. It is also convenient to normalize these probabilities so that they are $\Theta(1)$.

[3]By consistent, we mean that the estimator converges to the true value with probability one as the block length tends to infinity. This is sometimes called strong consistency.

[4]We do not index $\mathbf{X}$ by the block length n since it should be clear from the context.

Definition 1. *Let X_n be a random variable on A_n with distribution p_n. The* shadow *of p_n, denoted by P_n, is defined to be the distribution of the random variable $n \cdot p_n(X_n)$.*

Example 1. If $A_1 = \{a, b, c\}$, and

$$p_1(a) = p_1(b) = \frac{1}{2} \cdot p_1(c),$$

then the shadow, P_1, is uniform over $\{1/4, 1/2\}$. If p_n itself is uniform, then the shadow is a point mass.

Note that P_n is a probability distribution on $C := [\check{c}, \hat{c}]$, and that the entropy of p_n can be expressed as[5]

$$H(p_n) = -\int_C \log \frac{x}{n}\, dP_n(x). \tag{2}$$

In order to prove consistency results, we assume that the shadows converge weakly.

Definition 2. *A rare-events source is a sequence (A_n, p_n) of alphabets and distributions such that (1) holds for some positive $\check{c}$ and $\hat{c}$ and the shadows $\{P_n\}$ converge weakly to a distribution P.*

Example 2. If p_n is a uniform distribution over an alphabet of size n, then the shadow is a point mass at 1 for each n and hence converges in distribution. More complicated examples can be constructed by quantizing a fixed density more and more finely as follows. Let $f(x)$ be a density on $[0, 1]$ that is continuous a.e. such that

$$\check{c} \le f(x) \le \hat{c}.$$

Let X have density f and let p_n be the distribution of $\lceil nX \rceil$. Then p_n is a distribution on $\{1, \ldots, n\}$, and we obtain a rare-events source with the limiting shadow P being the distribution of $f(X)$. This example can be easily modified so that the cardinality of A_n is βn for some $\beta \neq 1$.

A. Important Limits

Our goal is to estimate the probability of the observed sequence, the entropy of the source, and other quantities using only the sequence itself. We first show that the quantities of interest converge to limits that depend on the limiting shadow P; this will also serve as a preview of the quantities to be estimated.

For each nonnegative integer k, let $B_{n,k}$ denote the random set of symbols in A_n that appear exactly k times in $\mathbf{X}$. We call

$$\gamma_{n,k} := p_n(B_{n,k})$$

the *total probability* of symbols appearing k times. We view $\gamma_{n,k}$ as a random probability distribution on the nonnegative integers. For a rare-events source, this distribution converges almost surely to a deterministic Poisson mixture.

Proposition 1. *The random distribution $\gamma_{n,k}$ converges to*

$$\lambda_k := \int_C \frac{x^k e^{-x}}{k!}\, dP(x) \quad k = 0, 1, 2, \ldots$$

in L^1 almost surely as $n \to \infty$.

The proofs of the results in this section combine moment calculations, usually involving a Poisson approximation to a binomial distribution, with concentration results. It should be mentioned that Proposition 1 above, and Proposition 7 and Theorem 1, which appear later, do not require the assumption that $\check{c} \le np_n(a) \le \hat{c}$ for all a and n [30]. Our proofs of the other results in this paper do rely on this assumption, however.

[5]Throughout we use natural logarithms.

Recall that the classical (finite-alphabet, fixed-distribution) asymptotic equipartition property (AEP) asserts that

$$\lim_{n\to\infty} \frac{1}{n} \log \mu(\mathbf{W}) = -H(\mu) \quad \text{a.s.}, \tag{3}$$

where $\mathbf{W}$ is an i.i.d. sequence drawn according to μ. Loosely speaking, (3) says that the probability of the observed sequence, $\mu(\mathbf{W})$, is approximately

$$\exp(-nH(\mu)).$$

In the rare events regime, one expects the probability of an observed sequence to be approximately

$$\left(\frac{e^{-h}}{n}\right)^n$$

for some constant h. Indeed, in the rare events regime the following AEP holds true.

Proposition 2. *For any rare-events source,*

$$\lim_{n\to\infty} \frac{1}{n} \log p_n(\mathbf{X}) + \log n = \int_C \log x \; dP(x) \quad \textit{a.s.}$$

It will be useful later to decompose both the sequence probability and the limit in Proposition 2 according to $B_{n,k}$. Write

$$\begin{aligned} \frac{1}{n} \log p_n(\mathbf{X}) + \log n &= \frac{1}{n} \sum_{i=1}^{n} \log(np_n(X_i)) \\ &= \sum_{k=1}^{n} \frac{k}{n} \sum_{a\in B_{n,k}} \log(np_n(a)). \end{aligned} \tag{4}$$

Proposition 3. *For any $k \geq 1$,*

$$\frac{k}{n} \sum_{a\in B_{n,k}} \log(np_n(a)) \to \int_C \frac{x^{k-1}e^{-x}}{(k-1)!} \log x \; dP(x) \quad \textit{a.s.}$$

Next consider the entropy of the source. From (2), we have

$$H(p_n) - \log n = -\int_C \log x \; dP_n(x).$$

The following characterization of the growth rate of the entropy is immediate.

Proposition 4. *For any rare-events source,*

$$\lim_{n\to\infty} H(p_n) - \log n = -\int_C \log x \; dP(x).$$

Proof: By hypothesis, P_n converges weakly to P, and $\log x$ is bounded and continuous over C. □

Consider next the relative entropy between the true distribution, p_n, and the empirical distribution, $p_{\mathbf{X}}$, of $\mathbf{X}$. The empirical distribution is the maximum likelihood estimate for p_n given $\mathbf{X}$, and it is natural to ask how far this estimate is from the true distribution, as measured by the relative entropy. In the conventional asymptotic, this relative entropy tends to zero as the block length tends to infinity. For a rare-events source, we have the following non-zero limit.

Proposition 5. *For any rare-events source,*

$$\begin{aligned} \lim_{n\to\infty} D(p_{\mathbf{X}}||p_n) &= \int_C \sum_{k=0}^{\infty} \frac{x^k e^{-x}}{k!} \log \frac{k+1}{x} \; dP(x) \quad \textit{a.s.} \\ &\geq e^{-\hat{c}}. \end{aligned}$$

Finally, consider again the total probabilities. Given a consistent estimator for the total probability of all symbols appearing k times, there is a natural estimator for the constituent probabilities of these symbols: we simply divide the estimated total probability by $|B_{n,k}|$. This estimate will be good only if the true distribution, p_n, restricted to $B_{n,k}$, is nearly uniform, and we would like to know how far p_n deviates from a uniform distribution when conditioned on $B_{n,k}$. The next result shows that the relative entropy between the two distributions converges to a limit that depends on P.

Proposition 6. *Let*

$$D_{n,k} = \sum_{a \in B_{n,k}} \frac{p_n(a)}{p_n(B_{n,k})} \log \frac{p_n(a)|B_{n,k}|}{p_n(B_{n,k})}.$$

denote the (random) relative entropy between the true distribution on $B_{n,k}$ and the uniform distribution over this set. Then for any rare-events source and any $k \geq 1$,

$$\lim_{n\to\infty} D_{n,k} = \frac{1}{\lambda_k} \int_C \frac{e^{-x}x^k}{k!} \log x \; dP(x) + \log \frac{\lambda_{k-1}}{k\lambda_k} \quad \textit{a.s.}$$

III. Good-Turing Consistency

We first show that the Good-Turing total probability estimator is consistent. This result will serve as a basis for the estimators to follow.

The Good-Turing estimator is traditionally viewed as an estimator for the probabilities of the individual symbols. Let $\varphi_{n,k} = |B_{n,k}|$ denote the number of symbols that appear exactly k times in the observed sequence. The basic Good-Turing estimator assigns probability

$$\frac{(k+1)\varphi_{n,k+1}}{n\varphi_{n,k}} \tag{5}$$

to each symbol that appears $k \leq n-1$ times [9]. The case $k = n$ must be handled separately, but this case is unimportant since under our model it is unlikely that the string will consist of one symbol repeated n times.

Actually, Good introduces (5) as an estimate of the probability of a symbol chosen uniformly at random from $B_{n,k}$. Good points out that this estimation problem is related to the problem of estimating the total probability of all symbols appearing k times, $\gamma_{n,k}$, because the $\varphi_{n,k}$ in the denominator can be interpreted as merely dividing the total probability equally among the $\varphi_{n,k}$ symbols on $B_{n,k}$. Thus the *Good-Turing total probability estimator* assigns probability

$$\phi_{n,k} := \frac{(k+1)\varphi_{n,k+1}}{n} \quad k = 0, 1, \ldots, n-1$$

to the *set* of symbols that have appeared k times. As a convention, we shall always assign zero probability to the set of symbols that appear n times

$$\phi_{n,n} := 0.$$

Like $\gamma_{n,k}$, $\phi_{n,k}$ is a random probability distribution on the nonnegative integers.

As an estimator for $\gamma_{n,k}$, $\phi_{n,k}$ is not ideal. For one thing, $\phi_{n,k}$ can be positive even when $B_{n,k}$ is empty and $\gamma_{n,k}$ is clearly zero. A similar problem arises when estimating the probabilities of individual symbols, and modifications to the basic Good-Turing estimator have been proposed to avoid it [9]. But we shall show that even the basic form of the Good-Turing estimator is consistent for total probability in the rare-events regime. The key is to establish a convergence result for the Good-Turing estimator that is analogous to Proposition 1 for the total probabilities.

Proposition 7. *The random distribution $\phi_{n,k}$ converges to λ_k in L^1 almost surely as $n \to \infty$.*

The proof of Proposition 7 parallels that of Proposition 1 in the previous section. In particular, we first show that the mean of $\phi_{n,k}$ converges to λ_k and then establish concentration around the mean. The desired consistency follows from this result and Proposition 1.

Theorem 1. *The Good-Turing total probability estimator is consistent in L^1, i.e.,*

$$\lim_{n\to\infty} \sum_{k=0}^{n} |\gamma_{n,k} - \phi_{n,k}| = 0 \quad \textit{a.s.}$$

Proof: We have

$$\sum_{k=0}^{n} |\gamma_{n,k} - \phi_{n,k}| \leq \sum_{k=0}^{n} |\gamma_{n,k} - \lambda_k| + \sum_{k=0}^{n} |\lambda_k - \phi_{n,k}|.$$

We now let $n \to \infty$ and invoke Propositions 1 and 7. □

For the case in which the underlying distribution p_n is uniform over an alphabet of size βn, Dupuis et al. [21] have determined the large-deviations behavior of vectors $(\varphi_{n,1}, \ldots, \varphi_{n,K})$. It would be desirable to extend their result to non-uniform rare-events sources, and also to determine the large-deviations behavior of the entire vector $(\varphi_{n,1}, \ldots, \varphi_{n,n})$.

IV. Single-Sequence Estimators

Next we turn to the problem of estimating the quantities of interest using the observed sequence. Recall that these quantities are

(*i*) the probability of the observed sequence,
(*ii*) the entropy of the underlying distribution,
(*iii*) the relative entropy between the empirical distribution and the true distribution, and
(*iv*) the relative entropy between the true distribution and the uniform distribution over all symbols that appear k times in the observed string, for each k.

The Good-Turing total probability estimator provides a natural starting point for the design of these estimators. We show that a naive application of the Good-Turing estimator does not yield a consistent estimator of the sequence probability, but that a more sophisticated application of the Good-Turing estimator indeed works. We then use this sequence probability estimate to obtain a consistent estimator for quantities (*ii*)–(*iv*).

A. Naive Good-Turing is not Consistent

Before discussing the new estimator, it is instructive to see how a naive application of the Good-Turing total probability estimator fails to yield a consistent sequence probability estimator. The naive approach is to first estimate the probability of each symbol and then multiply these probabilities accordingly. If we multiply the individual probability estimates in (5), we obtain the following estimate for the probability of the observed sequence

$$\prod_{k=1}^{n-1} \left(\frac{(k+1)\varphi_{n,k+1}}{n\varphi_{n,k}} \right)^{k\varphi_{n,k}}.$$

This in turn suggests the following estimator for the limit in Proposition 2

$$\sum_{k=1}^{n-1} \frac{k\varphi_{n,k}}{n} \log \left(\frac{(k+1)\varphi_{n,k+1}}{\varphi_{n,k}} \right). \tag{6}$$

This estimator is problematic, however, because for the largest k for which $\varphi_{n,k} > 0$,

$$\frac{(k+1)\varphi_{n,k+1}}{\varphi_{n,k}} = 0,$$

which means that the corresponding term in (6) equals $-\infty$. Various "smoothing" techniques have been introduced to address this and related problems with the estimator [9]. Our approach will be to truncate the summation at a large but fixed threshold, K

$$\sum_{k=1}^{K} \frac{k\varphi_{n,k}}{n} \log \left(\frac{(k+1)\varphi_{n,k+1}}{\varphi_{n,k}} \right).$$

In the rare events regime, with probability one it will eventually happen that $\varphi_{n,k} > 0$ for all $k = 1, \ldots, K$, thus obviating the problem.

By Proposition 7, this estimator will converge to

$$\sum_{k=1}^{K} \lambda_{k-1} \log \frac{k\lambda_k}{\lambda_{k-1}}. \tag{7}$$

We next show that this quantity need not tend to the correct limit (given in Proposition 2) as K tends to infinity.

Example 3. Consider the case in which A_n is the set $\{1, 2, \ldots, 3n\}$. Suppose that p_n assigns probability $1/(4n)$ to the first $2n$ elements and probability $1/(2n)$ to the remaining n. The limiting scaled shadow P will place mass $1/2$ on each of the points $1/4$ and $1/2$. From Proposition 2, the limiting normalized probability of $\mathbf{X}$ is $-(1/2)\log 8$. By (7), the naive estimate converges to

$$\frac{1}{2} \sum_{k=1}^{K} \frac{e^{-1/4}(1/4)^{k-1}}{(k-1)!} \left(1 + e^{-1/4} 2^{k-1}\right) \cdot \log \left(\frac{\sqrt{8}(1 + e^{-1/4} 2^k)}{4(1 + e^{-1/4} 2^{k-1})} \right)$$
$$+ \frac{1}{2} \sum_{k=1}^{K} \frac{e^{-1/4}(1/4)^{k-1}}{(k-1)!} \left(1 + e^{-1/4} 2^{k-1}\right) \log \frac{1}{\sqrt{8}}.$$

As K tends to infinity, the second sum converges to the correct answer, $-(1/2)\log 8$. But one can verify that every term in the first sum is strictly positive. Thus a naive application of the Good-Turing estimator is not consistent in this example. Note that simple modifications of the Good-Turing scheme such as that of Orlitsky et al. [12, Eq. (17)] will not rectify this.

The problem is that, according to Proposition 7, the Good-Turing estimator is estimating the sum, or equivalently the arithmetic mean, of the probabilities of the symbols appearing k times in $\mathbf{X}$. Estimating the sequence probability, on the other hand, amounts to estimating the product, or equivalently the geometric mean, of these probabilities. If p_n is uniform, then the arithmetic and geometric means coincide, and one can verify that the naive sequence probability estimator is consistent. In the above example, p_n is not uniform, and the naive estimator converges to the wrong value.

B. A Consistent Sequence-Probability Estimator

To create a consistent estimator, we write the normalized sequence probability as in (4)

$$\frac{1}{n} \log p_n(\mathbf{X}) + \log n = \sum_{k=1}^{n} \frac{k}{n} \sum_{a \in B_{n,k}} \log(n p_n(a)).$$

Thus it suffices to create a consistent estimator for the quantity

$$\frac{k}{n} \sum_{a \in B_{n,k}} \log(n p_n(a)) \tag{8}$$

for each k. Our approach is the following. From Propositions 3 and 7, we have

$$\lim_{n\to\infty} \frac{k}{n} \sum_{a\in B_{n,k}} \log(np_n(a)) = \int_C \frac{x^{k-1}e^{-x}}{(k-1)!} \log x \; dP(x) \quad \text{a.s.} \tag{9}$$

and

$$\lim_{n\to\infty} \phi_{n,k} = \int_C \frac{x^k e^{-x}}{k!} \; dP(x) = \lambda_k \quad \text{a.s.}$$

Our approach will be to express the right-hand side of (9) in terms of the λ_k. We will then "plug-in" $\phi_{n,k}$, which is only a function of the observed sequence, for λ_k to obtain an estimator. We begin by expanding $\log x$ as a Taylor series about a constant c

$$\log x = \log c - \sum_{m=1}^{\infty} \frac{(1-x/c)^m}{m} \qquad 0 < x \le 2c,$$

where the convergence is uniform over compact sets in $(0, 2c)$. By the binomial theorem, this can be written as

$$\log x = \log c - \sum_{m=1}^{\infty} \sum_{\ell=0}^{m} \frac{1}{m} \binom{m}{\ell} \left(-\frac{x}{c}\right)^{\ell} \qquad 0 < x \le 2c.$$

If we substitute this expression into (9) and formally swap the integral and infinite sum, we obtain

$$\begin{aligned}
\int_C \frac{x^{k-1}e^{-x}}{(k-1)!} \log x \; dP(x) &= \lambda_{k-1} \log c - \sum_{m=1}^{\infty} \sum_{\ell=0}^{m} \frac{1}{m} \binom{m}{\ell} (-c)^{-\ell} \int_C \frac{x^{k+\ell-1}e^{-x}}{(k-1)!} dP(x) \\
&= \lambda_{k-1} \log c - \sum_{m=1}^{\infty} \sum_{\ell=0}^{m} \frac{1}{m} \binom{m}{\ell} (-c)^{-\ell} \frac{(k+\ell-1)!}{(k-1)!} \lambda_{k+\ell-1} \\
&\approx \phi_{n,k-1} \log c - \sum_{m=1}^{\infty} \sum_{\ell=0}^{m} \frac{1}{m} \binom{m}{\ell} (-c)^{-\ell} \frac{(k+\ell-1)!}{(k-1)!} \phi_{n,k+\ell-1},
\end{aligned}$$

which is essentially our estimator. Two practical questions arise, namely, how many terms to include in the infinite sum and how to choose the constant c. Including more terms in the sum obviously provides for a better approximation of $\log x$, but the rate of convergence of $\phi_{n,k}$ slows as k increases. We show that we obtain a consistent estimator by having the number of terms grow very slowly with n

$$N = \lfloor (\log n)^{\epsilon_1} \rfloor,$$

where ϵ_1 is a constant in $(0, 1)$. Note that we suppress the dependence of N on n. In practice this choice amounts to including only the first few terms. Regarding the choice of c, if $\hat{c}$ were known, then we could choose c to guarantee that $[\check{c}, \hat{c}] \subseteq (0, 2c)$ so that the series expansion is uniformly convergent over C. Since we are not assuming that $\hat{c}$ is known, we choose c to grow with n to guarantee that $2c > \hat{c}$ eventually. There is a tension inherent in choosing the speed with which c grows, however. We desire rapid growth so that $(0, 2c)$ quickly envelopes C. But once this occurs, slower growth will yield better convergence of the power series over C. It turns out that if

$$c_n = \lfloor (\log n)^{\epsilon_2} \rfloor$$

where $0 < \epsilon_2 < \epsilon_1$, then the power series converges uniformly over compact sets, as shown next.

Lemma 1. *The function*

$$\log c_n - \sum_{m=1}^{N} \frac{1}{m} \left(1 - \frac{x}{c_n}\right)^m$$

converges to $\log x$ *uniformly on compact subsets of* $(0, \infty)$.

Our estimator for the quantity in (8) is the following

Definition 3. *For* $1 \leq k \leq n - N$*, let*

$$\zeta_{n,k} = \log(c_n)\phi_{n,k-1} - \sum_{m=1}^{N} \sum_{\ell=0}^{m} \frac{(-c_n)^{-\ell}}{m} \binom{m}{\ell} \frac{(k+\ell-1)!}{(k-1)!} \phi_{n,k+\ell-1}. \tag{10}$$

The next result shows that this estimator is consistent.

Theorem 2. *For any rare-events source and any* $k \geq 1$*,*

$$\lim_{n\to\infty} \zeta_{n,k} = \int_C \frac{x^{k-1}e^{-x}}{(k-1)!} \log x \; dP(x) \quad \textit{a.s.},$$

and hence

$$\zeta_{n,k} - \frac{k}{n} \sum_{a \in B_{n,k}} \log(np_n(a)) \to 0 \quad \textit{a.s.}$$

Our end goal is to estimate the sequence probability, and Theorem 2 and Proposition 2 together indicate that a natural estimator is $\sum_k \zeta_{n,k}$. Choosing the number of terms to include in this sum presents a similar tradeoff to the choice of the number of power series terms to include in $\zeta_{n,k}$ itself. We show that a consistent estimator can be obtained by again using $N = \lfloor (\log n)^{\epsilon_1} \rfloor$.

Definition 4.

$$\zeta_n = \sum_{k=1}^{N} \zeta_{n,k}. \tag{11}$$

Theorem 3. *For any rare-events source,*

$$\lim_{n\to\infty} \zeta_n = \int_C \log x \; dP(x) \quad \textit{a.s.},$$

and hence

$$\frac{1}{n} \log p_n(\mathbf{X}) + \log n - \zeta_n \to 0 \quad \textit{a.s.}$$

Since the estimator is an alternating sum with large constants, its numerical stability is unclear *a priori*. In Section VIII, we show via simulation that the estimator is stable and exhibits reasonable convergence properties. We also numerically optimize the ϵ_1 and ϵ_2 parameters.

C. A Consistent Estimator for Entropy and Relative Entropy

The sequence probability estimator can also be used to estimate the entropy of the source, the relative entropy between the true and empirical distributions, and the relative entropy between the true and uniform distributions over the symbols appearing k times. Recall that the entropy of p_n can be expressed as

$$-\int_C \log \frac{x}{n} \; dP_n(x).$$

Thus $H(p_n) - \log n$ converges to

$$-\int_C \log x \; dP(x).$$

Theorem 4. *For any rare-events source,*

$$H(p_n) - \log n + \zeta_n \to 0 \quad \textit{a.s.}$$

We turn next to the problem of estimating the relative entropy between the true distribution and the empirical distribution. It is well known that the probability of the observed sequence is given by [31, Theorem 11.1.2]

$$p_n(\mathbf{X}) = \exp(-n(D(p_{\mathbf{X}}||p_n) + H(p_{\mathbf{X}}))). \tag{12}$$

Thus our estimator for the sequence probability can be combined with the entropy of the empirical distribution to yield an estimator for $D(p_{\mathbf{X}}||p_n)$.

Theorem 5. *For any rare-events source,*

$$D(p_{\mathbf{X}}||p_n) + \zeta_n + H(p_{\mathbf{X}}) - \log n \to 0 \quad \textit{a.s.}$$

Proof: By (12),

$$D(p_{\mathbf{X}}||p_n) + \zeta_n + H(p_{\mathbf{X}}) - \log n = -\frac{1}{n}\log p_n(\mathbf{X}) - \log n + \zeta_n$$

which tends to zero almost surely by Theorem 3. □

The final result in this section shows that we can consistently estimate the relative entropy between the true and uniform distribution over the symbols appearing k times.

Theorem 6. *For any rare-events source and for any $k \geq 1$,*

$$D_{n,k} - \frac{\zeta_{n,k+1}}{\phi_{n,k}} - \log\frac{\phi_{n,k-1}}{k\phi_{n,k}} \to 0 \quad \textit{a.s.}$$

V. Two-Sequence Model

The results up to this point have addressed a single source in isolation. Many problems in natural language processing and information theory, such as hypothesis testing and mismatched compression, require considering multiple sources simultaneously.

We now suppose that for each n, we have a pair of probability measures on A_n, p_n and q_n, satisfying

$$\frac{\check{c}}{n} \leq \min(p_n(a), q_n(a)) \leq \max(p_n(a), q_n(a)) \leq \frac{\hat{c}}{n} \tag{13}$$

for all $a \in A_n$ and all n for some positive constants $\check{c}$ and $\hat{c}$. We observe two strings of length n. The first, $\mathbf{X}$, is drawn i.i.d. from A_n according to p_n. The second, $\mathbf{Y}$, is drawn i.i.d. according to q_n. We assume that the two strings are statistically independent.

Let P_n denote the distribution of

$$(np_n(X_n), nq_n(X_n)),$$

where X_n is drawn according to p_n. Likewise, let Q_n denote the distribution of

$$(np_n(Y_n), nq_n(Y_n)),$$

where Y_n is drawn according to q_n.

Note that both P_n and Q_n are probability measures on $C^2 := [\check{c}, \hat{c}] \times [\check{c}, \hat{c}]$. It follows from the definitions that P_n and Q_n are absolutely continuous with respect to each other with Radon-Nikodym derivative

$$\frac{dQ_n}{dP_n}(x, y) = \frac{y}{x}. \tag{14}$$

Note that the relative entropy between q_n and p_n is given by

$$D(q_n||p_n) = \int_{C^2} \log\frac{y}{x}\, dQ_n(x, y). \tag{15}$$

We shall again assume that P_n converges in distribution to a probability measure P on C^2. Since P_n and Q_n are related by (14), this implies that Q_n converges to Q satisfying

$$\frac{dQ}{dP}(x, y) = \frac{y}{x}.$$

Definition 5. *A rare events two-source is a sequence (A_n, p_n, q_n) of alphabets and distributions satisfying (13) such that P_n converges weakly to a distribution P.*

A. Important Limits

In the next section, we shall construct a consistent estimator for $p_n(\mathbf{Y})$, that is, the probability of the sequence generated using q_n under p_n, using only $\mathbf{X}$ and $\mathbf{Y}$. This problem arises in detection, where one must determine the likelihood of a given realization under multiple probability distributions. As in the single-sequence setup, this probability converges if it is suitably normalized.

Proposition 8. *For any rare-events two-source,*

$$\frac{1}{n}\log p_n(\mathbf{Y}) + \log n = \int_{C^2} \log x \; dQ(x,y) \quad \textit{a.s.} \tag{16}$$

An analogous result obviously holds for $q_n(\mathbf{X})$.

We shall also construct a consistent estimator for the relative entropy between q_n and p_n from the sequences $\mathbf{X}$ and $\mathbf{Y}$, a quantity that is related to $p_n(\mathbf{Y})$.

Proposition 9. *For any rare-events two-source,*

$$\begin{aligned}\lim_{n\to\infty} D(q_n||p_n) &= \int_{C^2} \log\frac{y}{x}\; dQ(x,y) \\ &= \int_{C^2} \frac{y}{x}\log\frac{y}{x}\; dP(x,y).\end{aligned}$$

The proof immediately follows from (15), since $\log y/x$ is bounded and continuous over C^2. An analogous result holds for $D(p_n||q_n)$.

VI. Two-Sequence Estimators

We turn to the problem of estimating $p_n(\mathbf{Y})$ and $D(q_n||p_n)$ using the sequences $\mathbf{X}$ and $\mathbf{Y}$. In the single-sequence setting, our starting point was the Good-Turing total probability estimator. For pairs of sequences, we require a similar "engine."

Let $\varphi_{n,k,j}$ denote the number of symbols that appear k times in $\mathbf{X}$ and j times in $\mathbf{Y}$. Then

$$\psi_{n,k} := \sum_{j=1}^{n} \frac{j\varphi_{n,k,j}}{n}$$

is the fraction of $\mathbf{Y}$ taken up by symbols appearing k times in $\mathbf{X}$. It turns out that $\psi_{n,k}$ obeys a convergence result that is similar to the one for $\phi_{n,k}$.

Proposition 10. *For any rare-events two-source and for any k,*

$$\lim_{n\to\infty} \psi_{n,k} = \int_{C^2} \frac{x^k e^{-x}}{k!}\; dQ(x,y) \quad \textit{a.s.} \tag{17}$$

Comparing Propositions 8 and 10, we see that to estimate $p_n(\mathbf{Y})$ we need a way of estimating the integral in (16) from the Poisson mixture given in (17). But this is equivalent to the problem of estimating the single-sequence probability from $\phi_{n,k}$, which was solved in Section IV. We simply replace $\phi_{n,k}$ with $\psi_{n,k}$ in the definition of $\zeta_{n,k}$ and ζ_n.

Definition 6. *Define*

$$\xi_n = \log(c_n)\sum_{k=1}^{N}\psi_{n,k-1} - \sum_{k=1}^{N}\sum_{m=1}^{N}\sum_{\ell=0}^{m}\frac{(-c_n)^{-\ell}}{m}\binom{m}{\ell}\frac{(k+\ell-1)!}{(k-1)!}\psi_{n,k+\ell-1}.$$

Theorem 7. *For any rare-events two-source,*

$$\lim_{n\to\infty}\xi_n = \int_{C^2}\log x\; dQ(x,y) \quad \textit{a.s.} \tag{18}$$

and hence

$$\lim_{n\to\infty} \frac{1}{n} \log p_n(\mathbf{Y}) + \log n - \xi_n = 0 \quad a.s. \tag{19}$$

By combining ξ_n with our single-sequence estimator, we can consistently estimate the relative entropy between q_n and p_n. Let us redefine $\varphi_{n,k}$ to be the number of symbols appearing k times in $\mathbf{Y}$, and let $\zeta_{n,k}$ and ζ_n be the estimators in (10) and (11) as before.

Theorem 8. *For any rare-events two-source,*

$$\lim_{n\to\infty} |D(q_n||p_n) + \xi_n - \zeta_n| = 0 \quad a.s. \tag{20}$$

VII. Universal Hypothesis Testing

The ξ_n estimator also provides a consistent decision rule for universal hypothesis testing. In particular, it shows that consistent universal hypothesis testing for rare-events sources is possible. Suppose that we again observe $\mathbf{X}$ and $\mathbf{Y}$, which we now view as training sequences. In addition, we observe a test sequence, say $\mathbf{Z}$, which is generated i.i.d. from the distribution r_n. We assume that $\mathbf{Z}$ is independent of $\mathbf{X}$ and $\mathbf{Y}$ and that either $r_n = p_n$ for all n or $r_n = q_n$ for all n. The problem is to determine which of these two possibilities is in effect using only the sequences $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$.

Under the conventional asymptotic, the two possible distributions can be learned from the test sequences in the large n limit, (say, from their empirical distributions), and a standard likelihood ratio test can be employed. This can be easily shown to be a consistent decision rule, although there are other schemes with superior error exponents [32], [33]. For rare-events sources, finding a consistent decision rule is less simple.

Using Theorem 7, one can estimate $p_n(\mathbf{Z})$ and $q_n(\mathbf{Z})$ and by comparing the two, determine which of the two distributions is more likely to have generated $\mathbf{Z}$. The next result shows that this decision rule is consistent.

Lemma 2.

$$\int_{C^2} \log x \; dQ(x,y) \leq \int_{C^2} \log y \; dQ(x,y),$$

with equality if and only if $P = Q$, i.e.,

$$P((x,y) : x = y) = Q((x,y) : x = y) = 1.$$

Proof: Since

$$\log \frac{x}{y} \leq \frac{x}{y} - 1$$

with equality if and only if $x = y$, we have

$$\int_{C^2} \log \frac{y}{x} \; dQ(x,y) \geq \int_{C^2} \left(1 - \frac{x}{y}\right) \; dQ(x,y)$$

with equality if and only if

$$Q((x,y) : x = y) = 1.$$

Now

$$\begin{aligned} \int_{C^2} \left(1 - \frac{x}{y}\right) \; dQ(x,y) &= \int_{C^2} \; dQ(x,y) - \int_{C^2} \frac{x}{y} \; dQ(x,y) \\ &= 1 - \int_{C^2} \; dP(x,y) = 0. \end{aligned}$$

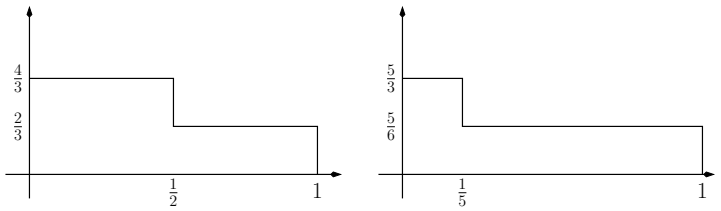
But

$$Q((x,y) : x = y) = 1$$

if and only if

$$P((x,y) : x = y) = 1$$

if and only if $P = Q$. □



(a) First simulated distribution (b) Second simulated distribution

Fig. 1. Densities used to simulate the single-sequence and two-sequence estimators. The probabilities of the individual symbols are assigned from these densities as in Example 2.

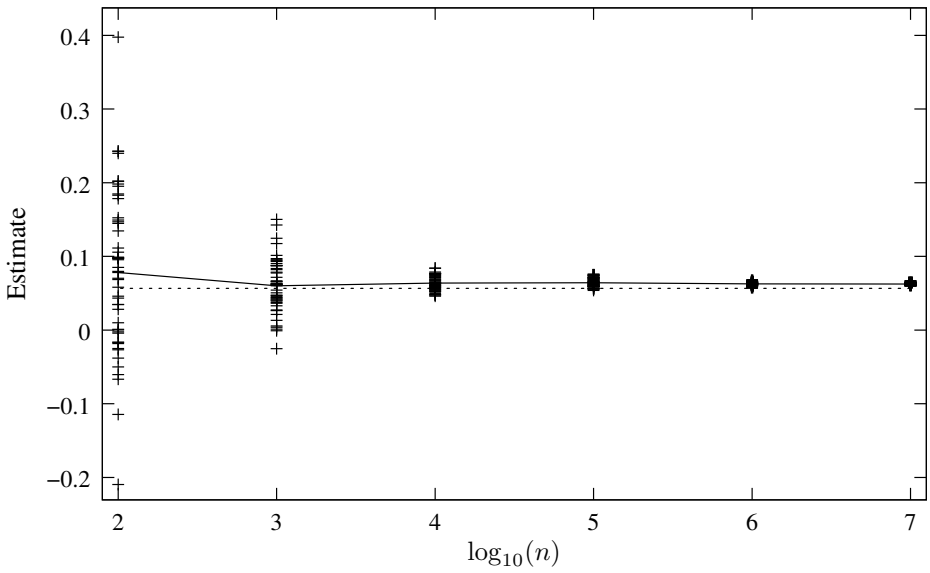


Fig. 2. Simulation of single-sequence probability estimator for the source distribution in Fig. 1(a). The dotted line indicates the true limiting sequence probability, as computed using Proposition 2. For each value of n, the estimator was computed on 50 independent realizations of the source. Each estimate is noted by a '+' on the graph, and the mean estimate is indicated by the solid line. ($\epsilon_1 = 0.99$, $\epsilon_2 = 0.5$)

VIII. Simulation Results

We next simulate the one- and two-sequence probability estimators to determine their convergence properties, test their numerical stability, and optimize the parameters ϵ_1 and ϵ_2. Consider the single-sequence estimator (ζ_n) and suppose that the source is generated via Example 2 using the bi-uniform density in Fig. 1(a). Fig 2 shows the simulated performance of the estimator for $n = 10^m$ where m varies from 2 to 7. Fig. 3 shows the results of a similar simulation using the distribution in Fig. 1(b). Both simulations show reasonable convergence and numerical stability.

The plots were generated using the parameters $\epsilon_1 = 0.99$ and $\epsilon_2 = 0.5$. Increasing ϵ_1 tends to reduce the bias of the estimator while increasing its variance, as illustrated in Fig. 4. As the source distribution becomes more "peaky," this effect becomes more pronounced, which in turn makes the range of acceptable ϵ_1 smaller[6]. The choice $\epsilon = 0.99$ provides both a small bias and small variance in most cases. The estimator is relatively insensitive to the choice of ϵ_2.

Fig. 5 shows the performance of the two-sequence estimator (ξ_n) in which $\mathbf{X}$ and $\mathbf{Y}$ are chosen according to the densities in Fig. 1(b) and 1(a), respectively. Again the densities are mapped to rare-events sources using the sampling approach in Example 2. We see that the same values of ϵ_1 and ϵ_2 also work for the two-sequence estimator.

[6]Also note that as the distribution becomes more peaky, the variance of $p_n(\mathbf{X})$ increases, which makes this quantity more difficult to estimate.

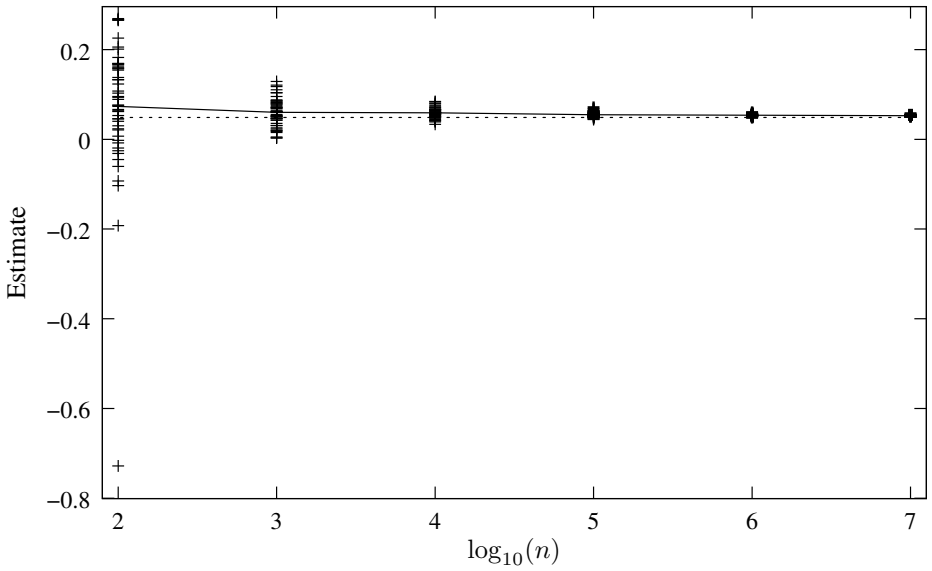


Fig. 3. Simulation of the single-sequence probability estimator for the source distribution in Fig. 1(b). ($\epsilon_1 = 0.99$, $\epsilon_2 = 0.5$)

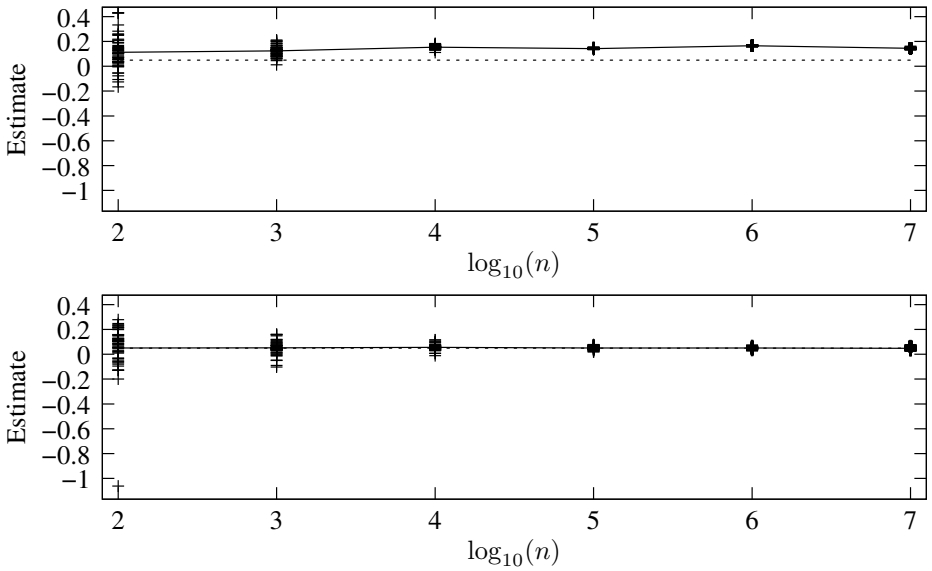


Fig. 4. Effect of ϵ_1 on estimator performance for the distribution in Fig. 1(b). The upper plot uses $\epsilon_1 = 0.6$ and the lower plot uses $\epsilon_1 = 1.1$. Increasing ϵ_1 tends to reduce the bias of the estimate while increasing its variance. Although our results do not guarantee convergence for the $\epsilon_1 = 1.1$, the lower plot is useful for observing the effect of increasing ϵ_1.

IX. Acknowledgment

The authors thank Thitidej Tularak for writing the simulation code and Benjamin Kelly for calling their attention to some of the references. This research was supported by the National Science Foundation under grant CCF-0830496.

Appendix A
Single-Sequence Limits

The proofs of the results in Section II tend to follow a common pattern. We first compute the expectation of the relevant quantity and show that it converges to the desired limit. We then show concentration around the mean to establish almost sure convergence. For the expectation calculations, it is convenient to make

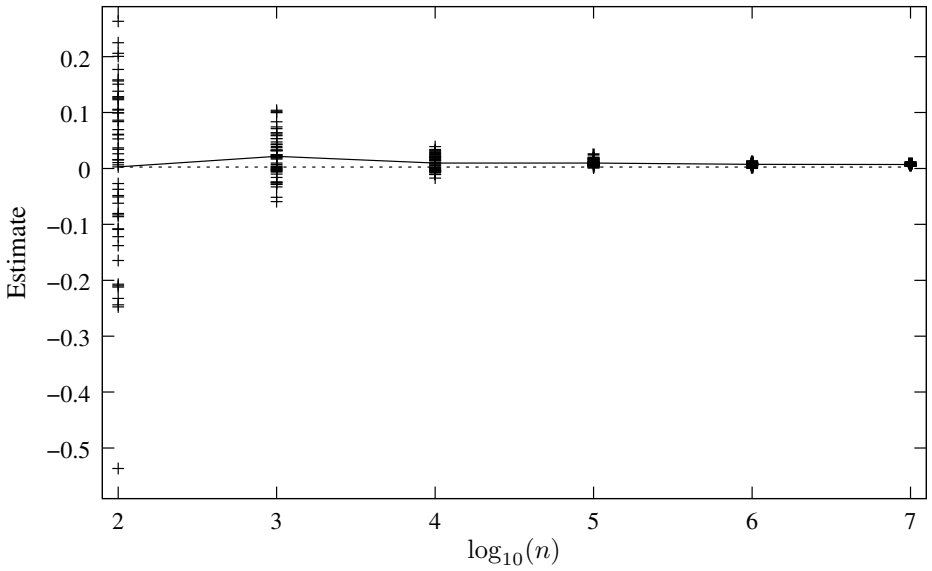


Fig. 5. Simulation of the two-sequence probability estimator. ($\epsilon_1 = 0.99$, $\epsilon_2 = 0.5$)

several definitions. Let

$$g_k^n(x) = \binom{n}{k} \left(\frac{x}{n}\right)^k \left(1 - \frac{x}{n}\right)^{n-k}$$

and

$$g_k(x) = \frac{x^k \exp(-x)}{k!}.$$

Since

$$\binom{n}{k} \frac{1}{n^k} \to \frac{1}{k!} \quad \text{as } n \to \infty$$

and

$$\left(1 + \frac{x_n}{n}\right)^n \to \exp(x) \quad \text{if } x_n \to x,$$

it follows that for all sequences $x_n \to x$, $g_k^n(x_n) \to g_k(x)$. Note also that $g_k^n(x) \leq 1$ if $0 \leq x \leq n$ by the binomial theorem. In several of the proofs we will use the abbreviation

$$c := \max(|\log(\check{c})|, |\log(\hat{c})|).$$

Lemma 3. *For all nonnegative integers k,*

$$\lim_{n\to\infty} E[\gamma_{n,k}] = \lambda_k.$$

Proof: For any $k \geq 0$,

$$\begin{aligned} E[\gamma_{n,k}] &= \sum_{a \in A_n} \binom{n}{k} p_n(a)^k (1 - p_n(a))^{n-k} p_n(a) \\ &= \sum_{a \in A_n} g_k^n(np_n(a)) p_n(a) \\ &= E[g_k^n(np_n(X_n))], \end{aligned}$$

where X_n has distribution p_n. Since $np_n(X_n)$ converges in distribution to a random variable W with distribution P, we can create a sequence of random variables $\{W_n\}_{n=1}^{\infty}$ such that W_n has the same distribution as $np_n(X_n)$ and W_n converges to W almost surely [34, Theorem 4.30]. Then

$$g_k^n(W_n) \to g_k(W) \quad \text{a.s.}$$

Since $g_k^n(W_n) \leq 1$ a.s., the bounded convergence theorem implies that

$$\begin{aligned}\lim_{n\to\infty} E[g_k^n(W_n)] &= E[g_k(W)] \\ &= \int_0^\infty g_k(x)\, dP(x) = \lambda_k.\end{aligned}$$

□

Lemma 4. *For all nonnegative integers k,*

$$\lim_{n\to\infty} |\gamma_{n,k} - E[\gamma_{n,k}]| = 0 \quad \textit{a.s.}$$

Proof: If we change one symbol in the underlying sequence, then $\gamma_{n,k}$ can change by at most $2\hat{c}/n$. By McDiarmid's [35] form of the Azuma-Hoeffding concentration inequality (available as [36, Corollary 2.4.14]), it follows that for all $\tau > 0$

$$\Pr\left(|\gamma_{n,k} - E[\gamma_{n,k}]| \geq \tau\right) \leq 2\exp\left[-\frac{n\tau^2}{8\hat{c}^2}\right].$$

Since the right-hand side is summable over n, the Borel Cantelli lemma implies the result. □

Proof of Proposition 1: It follows from Lemmas 3 and 4 that for each k,

$$\lim_{n\to\infty} \gamma_{n,k} = \lambda_k \quad \text{a.s.}$$

That is, the random distribution $\gamma_{n,k}$ converges pointwise to λ_k with probability one. The strengthening to L^1 convergence follows from Scheffé's theorem [37, Theorem 16.12], but we provide a proof since it is brief. Observe that with probability one,

$$\begin{aligned}0 &= \sum_{k=0}^\infty [\lambda_k - \gamma_{n,k}] \\ &= \sum_{k=0}^\infty [\lambda_k - \gamma_{n,k}]^+ - \sum_{k=0}^\infty [\lambda_k - \gamma_{n,k}]^-,\end{aligned}$$

where $[\cdot]^+$ and $[\cdot]^-$ represent the positive and negative parts, respectively. Thus

$$\sum_{k=0}^\infty |\lambda_k - \gamma_{n,k}| = 2\sum_{k=0}^\infty [\lambda_k - \gamma_{n,k}]^+ \quad \text{a.s.}$$

But $[\lambda_k - \gamma_{n,k}]^+$ converges pointwise to 0 a.s. and is less than or equal to λ_k. The dominated convergence theorem then implies that

$$\lim_{n\to\infty} \sum_{k=0}^\infty [\lambda_k - \gamma_{n,k}]^+ = 0 \quad \text{a.s.}$$

□

Lemma 5. (a) *For any $k \geq 1$,*

$$\lim_{n\to\infty} E\left[\frac{k}{n} \sum_{a\in B_{n,k}} \log(np_n(a))\right] = \int_C \frac{x^{k-1}e^{-x}}{(k-1)!} \log x\, dP(x).$$

(b)

$$\lim_{n\to\infty} E\left[\frac{1}{n}\log p_n(\mathbf{X})\right] + \log n = \int_C \log x\, dP(x).$$

Proof: For any $k \geq 1$,

$$E\left[\frac{k}{n}\sum_{a\in B_{n,k}}\log(np_n(a))\right] = \frac{k}{n}\sum_{a\in A_n}\binom{n}{k}(p_n(a))^k(1-p_n(a))^{n-k}\log(np_n(a))$$
$$= \sum_{a\in A_n} g_{k-1}^{n-1}((n-1)p_n(a))\log(np_n(a))p_n(a)$$
$$= E\left[g_{k-1}^{n-1}((n-1)p_n(X_n))\log(np_n(X_n))\right].$$

As in the proof of Lemma 3, it follows that

$$\lim_{n\to\infty} E\left[g_{k-1}^{n-1}((n-1)p_n(X_n))\log(np_n(X_n))\right]$$
$$= \int_C \frac{e^{-x}x^{k-1}}{(k-1)!}\log x\; dP(x),$$

which establishes (a). Now

$$\frac{1}{n}\log p_n(\mathbf{X}) + \log n = \frac{1}{n}\sum_{i=1}^{n}\log(np_n(X_i)).$$

Thus

$$E\left[\frac{1}{n}\log p_n(\mathbf{X})\right] + \log n = E[\log(np_n(X_n))]$$
$$= \int_C \log x\; dP_n(x)$$
$$\to \int_C \log x\; dP(x),$$

where the convergence follows because $P_n \to P$ weakly and $\log x$ is bounded and continuous over C. □

Lemma 6. (a) *For any $k \geq 1$,*

$$\frac{k}{n}\sum_{a\in B_{n,k}}\log(np_n(a)) - \lim_{n\to\infty} E\left[\frac{k}{n}\sum_{a\in B_{n,k}}\log(np_n(a))\right] = 0 \quad \textit{a.s.}$$

(b)

$$\lim_{n\to\infty}\frac{1}{n}\log p_n(\mathbf{X}) - E\left[\frac{1}{n}\log p_n(\mathbf{X})\right] = 0 \quad \textit{a.s.}$$

Proof: If one symbol in the underlying i.i.d. sequence is altered, then

$$\frac{k}{n}\sum_{a\in B_{n,k}}\log(np_n(a))$$

can change by at most $2kc/n$ and

$$\frac{1}{n}\log p_n(\mathbf{X}) + \log n$$

can change by at most $2c/n$. Concentration and almost sure convergence then follows as in the proof of Lemma 4. □

Propositions 2 and 3 follow immediately from Lemmas 5 and 6. Before proving Proposition 5, we prove a convergence result for the empirical entropy.

Lemma 7. *For any rare-events source,*

$$H(p_{\mathbf{X}}) - \log n \to -\int_C \sum_{k=1}^{\infty} \frac{e^{-x}x^k}{k!} \log(k+1)\, dP(x) \quad \textit{a.s.}$$

Proof: We have

$$\begin{aligned} H(p_{\mathbf{X}}) - \log n &= -\sum_{k=1}^{n} \frac{\varphi_{n,k} \cdot k}{n} \log k \\ &= -\sum_{k=1}^{n-1} \phi_{n,k} \log(k+1). \end{aligned}$$

Then for all sufficiently large n,

$$\left| \sum_{k=0}^{n-1} \phi_{n,k} \log(k+1) - \sum_{k=0}^{n-1} E[\phi_{n,k}] \log(k+1) \right| \le \sum_{k=0}^{n-1} |\phi_{n,k} - E[\phi_{n,k}]| \cdot n^{1/3},$$

which tends to zero a.s. by Lemma 11 to follow. Thus it suffices to show that

$$\sum_{k=0}^{n-1} E[\phi_{n,k}] \log(k+1) \to \int_C \sum_{k=0}^{\infty} \frac{e^{-x}x^k}{k!} \log(k+1)\, dP(x). \tag{21}$$

As in the proof of Lemma 10 to follow,

$$E[\phi_{n,k}] = \int_C g_k^{n-1}\left(\frac{n-1}{n}x\right)\, dP_n(x).$$

Since

$$\sum_{k=0}^{n-1} \phi_{n,k} = 1 \quad \text{a.s.}$$

we have that

$$\sum_{k=0}^{n-1} E[\phi_{n,k}] = \sum_{k=0}^{n-1} \int_C g_k^{n-1}\left(\frac{n-1}{n}x\right)\, dP_n(x) = 1.$$

Therefore we can create a random variable W_n with distribution

$$\Pr(W_n = k) = E[\phi_{n,k}].$$

By Lemma 10 to follow, $\{W_n\}$ converges in distribution to a random variable W with distribution

$$\Pr(W = k) = \int_C \frac{e^{-x}x^k}{k!}\, dP(x).$$

Since W_n is a mixture of binomials,

$$\begin{aligned} E[W_n] &= \sum_{k=0}^{n-1} \int_C k g_k^{n-1}\left(\frac{n-1}{n}x\right)\, dP_n(x) \\ &= \int_C \sum_{k=0}^{n-1} k g_k^{n-1}\left(\frac{n-1}{n}x\right)\, dP_n(x) \\ &= \frac{n-1}{n} \int_C x dP_n(x) \\ &\to \int_C x\, dP(x) \end{aligned}$$

where the convergence follows because the identity function x is bounded and continuous over C and P_n converges weakly to P. But by monotone convergence,

$$\begin{aligned} E[W] &= \sum_{k=0}^{\infty} \int_C k \frac{e^{-x}x^k}{k!}\, dP(x) \\ &= \int_C \sum_{k=0}^{\infty} k \frac{e^{-x}x^k}{k!}\, dP(x) \\ &= \int_C x\, dP(x). \end{aligned}$$

Thus $E[W_n] \to E[W]$, and the sequence $\{W_n\}$ is uniformly integrable [37, Theorem 16.14], which implies that $\log(W_n + 1)$ is uniformly integrable [38, Ex. 4.5.1], and hence

$$E[\log(W_n + 1)] \to E[\log(W + 1)].$$

But

$$E[\log(W_n + 1)] = \sum_{k=0}^{n-1} E[\phi_{n,k}] \log(k + 1)$$

and

$$E[\log(W + 1)] = \sum_{k=0}^{\infty} \int_C \frac{e^{-x}x^k}{k!} dP(x) \log(k + 1)$$

which by monotone convergence equals

$$\int_C \sum_{k=0}^{\infty} \frac{e^{-x}x^k}{k!} \log(k + 1)\, dP(x).$$

This establishes (21) and hence the lemma. □

Proof of Proposition 5: We have

$$\begin{aligned} D(p_{\mathbf{X}}||p_n) &= -H(p_{\mathbf{X}}) - \frac{1}{n} \log p_n(\mathbf{X}) \\ &= -\left[H(p_{\mathbf{X}}) - \log n\right] - \left[\frac{1}{n} \log p_n(\mathbf{X}) + \log n\right]. \end{aligned}$$

The convergence then follows from Lemma 7 and Proposition 2. To show the $e^{-\hat{c}}$ lower bound, we use the fact that $\log x \le x - 1$,

$$\begin{aligned} \int_C \sum_{k=0}^{\infty} \frac{x^k e^{-x}}{k!} \log \frac{x}{k+1}\, dP(x) &\le \int_C \sum_{k=0}^{\infty} \frac{x^k e^{-x}}{k!} \left(\frac{x}{k+1} - 1\right)\, dP(x) \\ &= \int_C \sum_{k=0}^{\infty} \frac{x^{k+1} e^{-x}}{(k+1)!}\, dP(x) - 1 \\ &= -\int_C e^{-x}\, dP(x) \\ &\le -e^{-\hat{c}}. \end{aligned}$$

□

Lemma 8. *For any $k \ge 0$,*

$$\lim_{n\to\infty} E\left[\sum_{a \in B_{n,k}} p_n(a) \log(n p_n(a))\right] = \int_C \frac{e^{-x}x^k}{k!} \log x\, dP(x).$$

Proof: We have

$$E\left[\sum_{a\in B_{n,k}} p_n(a)\log(np_n(a))\right] = \sum_{a\in A_n}\binom{n}{k}(p_n(a))^{k+1}(1-p_n(a))^{n-k}\log(np_n(a))$$
$$= \sum_{a\in A_n} g_k^n(np_n(a))\log(np_n(a))p_n(a)$$
$$= E[g_k^n(np_n(X_n))\log(np_n(X_n))].$$

As in the proof of Lemma 3, we have

$$\lim_{n\to\infty} E[g_k^n(np_n(X_n))\log(np_n(X_n))] = \int_C \frac{e^{-x}x^k}{k!}\log x\; dP(x).$$

□

Lemma 9. *For any $k \geq 0$,*

$$\lim_{n\to\infty}\sum_{a\in B_{n,k}} p_n(a)\log(np_n(a)) - E\left[\sum_{a\in B_{n,k}} p_n(a)\log(np_n(a))\right] = 0.$$

Proof: If one symbol in the underlying i.i.d. sequence is altered, then

$$\sum_{a\in B_{n,k}} p_n(a)\log(np_n(a))$$

can change by at most $2c\hat{c}/n$. Concentration and a.s. convergence then follow as in the proof of Lemma 4. □

Proof of Proposition 6: We have

$$D_{n,k} = \frac{1}{p_n(B_{n,k})}\sum_{a\in B_{n,k}} p_n(a)\log(np_n(a)) + \log\frac{\varphi_{n,k}}{np_n(B_{n,k})}.$$

By Lemmas 8 and 9,

$$\lim_{n\to\infty}\sum_{a\in B_{n,k}} p_n(a)\log(np_n(a)) = \int_C \frac{e^{-x}x^k}{k!}\log x\; dP(x) \quad \text{a.s.}$$

By Proposition 1,

$$\lim_{n\to\infty} p_n(B_{n,k}) = \lambda_k \quad \text{a.s.},$$

and by Proposition 7

$$\lim_{n\to\infty}\frac{\varphi_{n,k}}{n} = \frac{\lambda_{k-1}}{k} \quad \text{a.s.}$$

The result follows. □

APPENDIX B
GOOD-TURING ESTIMATOR

Lemma 10. *For any rare-events source,*

$$\lim_{n\to\infty}\sum_{k=0}^{\infty}|E[\phi_{n,k}]-\lambda_k|=0.$$

Proof: For any $0\le k\le n-1$,

$$\phi_{n,k}=\sum_{a\in A_n}\frac{k+1}{n}1(a\in B_{n,k+1}).$$

Thus

$$\begin{aligned}E[\phi_{n,k}]&=\sum_{a\in A_n}\frac{k+1}{n}\binom{n}{k+1}(p_n(a))^{k+1}(1-p_n(a))^{n-k-1}\\&=\sum_{a\in A_n}\binom{n-1}{k}(p_n(a))^{k}(1-p_n(a))^{n-k-1}p_n(a)\\&=\sum_{a\in A_n}g_k^{n-1}((n-1)p_n(a))p_n(a)\\&=E[g_k^{n-1}((n-1)p_n(X_n))].\end{aligned}$$

The reasoning in the proof of Lemma 3 can then be used to show that $E[\phi_{n,k}]\to\lambda_k$. The strengthening to L^1 convergence follows from Scheffé's theorem. □

Lemma 11. *For any $\delta>0$,*

$$\lim_{n\to\infty}\sum_{k=0}^{n-1}|\phi_{n,k}-E[\phi_{n,k}]|\cdot n^{1/2-\delta}=0\quad a.s.$$

Remark: The $n^{1/2-\delta}$ factor is not necessarily the largest possible, but it is sufficient for our purposes.

Proof: We have

$$\begin{aligned}&\sum_{k=0}^{n-1}|\phi_{n,k}-E[\phi_{n,k}]|\cdot n^{1/2-\delta}\\&\qquad\le\sum_{k=0}^{\lceil n^{\delta/4}\rceil}|\phi_{n,k}-E[\phi_{n,k}]|\cdot n^{1/2-\delta}+\sum_{k=\lceil n^{\delta/4}\rceil+1}^{n-1}\phi_{n,k}\cdot n^{1/2-\delta}+\sum_{k=\lceil n^{\delta/4}\rceil+1}^{n-1}E[\phi_{n,k}]\cdot n^{1/2-\delta}.\end{aligned}\quad(22)$$

We will show that each of these terms tends to zero in turn. For the first term, observe that if we alter one symbol in the underlying i.i.d. sequence, then $\phi_{n,k}$ will change by at most $2(k+1)/n$. As in the proof of Lemma 4, McDiarmid's form of the Azuma-Hoeffding concentration inequality implies that

$$\Pr\left(|\phi_{n,k}-E[\phi_{n,k}]|>\epsilon\right)\le 2\exp\left(-\frac{\epsilon^2 n}{8(n^{\delta/4}+2)^2}\right)$$

for all $k\le\lceil n^{\delta/4}\rceil$. Then by the union bound,

$$\begin{aligned}\Pr\left(\sum_{k=0}^{\lceil n^{\delta/4}\rceil}|\phi_{n,k}-E[\phi_{n,k}]|\cdot n^{1/2-\delta}>\epsilon\right)&\le\sum_{k=0}^{\lceil n^{\delta/4}\rceil}\Pr\left(|\phi_{n,k}-E[\phi_{n,k}]|\cdot n^{1/2-\delta}>\frac{\epsilon}{\lceil n^{\delta/2}\rceil}\right)\\&\le 2(n^{\delta/4}+1)\exp\left(-\frac{\epsilon^2 n^{2\delta}}{8(n^{\delta/4}+2)^2(n^{\delta/2}+1)^2}\right).\end{aligned}$$

Since the right-hand side is summable, the first term in (22) tends to zero a.s. by the Borel-Cantelli lemma. To handle the second term, note that since $\varphi_{n,k}$ is integer valued, it suffices to show that

$$\lim_{n\to\infty} \sup_{\lceil n^{\delta/4}\rceil<k\le n-1} \phi_{n,k} = 0 \quad \text{a.s.} \tag{23}$$

To show this, observe that

$$\Pr(\varphi_{n,k} > 0) \le \sum_{a\in A_n} \binom{n}{k} (p_n(a))^k (1-p_n(a))^{n-k}.$$

Now if $k > \lceil n^{\delta/4}\rceil$, then for all sufficiently large n, we can upper bound the right-hand side by taking $k = \lceil n^{\delta/4}\rceil$ [31, Eq. 11.47], yielding

$$\begin{aligned}\Pr(\varphi_{n,k} > 0) &\le \sum_{a\in A_n} \binom{n}{\lceil n^{\delta/4}\rceil} (p_n(a))^{\lceil n^{\delta/4}\rceil} (1-p_n(a))^{n-\lceil n^{\delta/4}\rceil} \\ &\le \frac{n}{\check{c}} \binom{n}{\lceil n^{\delta/4}\rceil} \left(\frac{\hat{c}}{n}\right)^{\lceil n^{\delta/4}\rceil}.\end{aligned}$$

Using Feller's bounds on the Stirling approximation, Orlitsky et al. [11, Lemma 4] have shown that

$$\binom{n}{k} \le \frac{e^{1/(12n)}}{\sqrt{2\pi}} \sqrt{\frac{n}{k(n-k)}} \left(\frac{en}{k}\right)^k. \tag{24}$$

Thus

$$\begin{aligned}\Pr(\varphi_{n,k} > 0) &\le \frac{n}{\check{c}} \frac{e}{\sqrt{2\pi}} \sqrt{\frac{n}{\lceil n^{\delta/4}\rceil (n-\lceil n^{\delta/4}\rceil)}} \left(\frac{e\hat{c}}{\lceil n^{\delta/4}\rceil}\right)^{\lceil n^{\delta/4}\rceil} \\ &\le \frac{n^{3/2}e}{\check{c}\sqrt{2\pi}} \left(\frac{e\hat{c}}{\lceil n^{\delta/4}\rceil}\right)^{\lceil n^{\delta/4}\rceil}.\end{aligned}$$

By the union bound, this implies that for all sufficiently large n,

$$\begin{aligned}\Pr\left(\sup_{\lceil n^{\delta/4}\rceil<k\le n-1} \varphi_{n,k} > 0\right) &\le \frac{n^{5/2}e}{\check{c}\sqrt{2\pi}} \left(\frac{e\hat{c}}{\lceil n^{\delta/4}\rceil}\right)^{\lceil n^{\delta/4}\rceil} \\ &\le \frac{n^{5/2}e}{\check{c}\sqrt{2\pi}} \left(\frac{e\hat{c}}{n^{\delta/4}}\right)^{n^{\delta/4}}\end{aligned}$$

Since the right-hand side is summable, it follows that

$$\lim_{n\to\infty} \sup_{\lceil n^{\delta/4}\rceil<k\le n-1} \varphi_{n,k} = 0 \quad \text{a.s.}$$

which implies (23). Finally, turning to the third term in (22), we have

$$\begin{aligned}E[\phi_{n,k}] &= \sum_{a\in A_n} \binom{n-1}{k} (p_n(a))^k (1-p_n(a))^{n-k-1} p_n(a) \\ &\le \sum_{a\in A_n} \binom{n-1}{k} (p_n(a))^k (1-p_n(a))^{n-k-1}.\end{aligned}$$

If $k > \lceil n^{\delta/4} \rceil$, then the right-hand side can be upper bounded by taking $k = \lceil n^{\delta/4} \rceil$. Thus

$$\begin{aligned} E\left[\phi_{n,k}\right] &\leq \sum_{a \in A_n} \binom{n-1}{\lceil n^{\delta/4} \rceil} (p_n(a))^{\lceil n^{\delta/4} \rceil} \\ &\leq \sum_{a \in A_n} \binom{n-1}{\lceil n^{\delta/4} \rceil} \left(\frac{\hat{c}}{n}\right)^{\lceil n^{\delta/4} \rceil} \\ &\leq \frac{n}{\check{c}} \binom{n-1}{\lceil n^{\delta/4} \rceil} \left(\frac{\hat{c}}{n}\right)^{\lceil n^{\delta/4} \rceil}, \end{aligned}$$

which implies that

$$\sum_{k=\lceil n^{\delta/4} \rceil + 1}^{n-1} E[\phi_{n,k}] \cdot n^{1/2-\delta} \leq \frac{n^{5/2-\delta}}{\check{c}} \binom{n-1}{\lceil n^{\delta/4} \rceil} \left(\frac{\hat{c}}{n}\right)^{\lceil n^{\delta/4} \rceil}.$$

Using the Orlitsky et al. bound in (24) once again shows that the right-hand side tends to zero. This shows that the right-hand side in (22) tends to zero a.s. and completes the proof. □

Proof of Proposition 7: The result follows from Lemmas 10 and 11. □

APPENDIX C
SINGLE-SEQUENCE ESTIMATORS

Proof of Lemma 1: By the well-known Mercator series, for all y such that $-1 < y \leq 1$,

$$\log(1+y) = -\sum_{m=1}^{\infty} \frac{(-y)^m}{m}.$$

In particular,

$$\begin{aligned} \left| \log(1+y) + \sum_{m=1}^{N} \frac{(-y)^m}{m} \right| &= \left| \sum_{m=N+1}^{\infty} \frac{(-y)^m}{m} \right| \\ &\leq \sum_{m=N+1}^{\infty} |y|^m \\ &= \frac{|y|^{N+1}}{1-|y|}. \end{aligned}$$

Consider a compact set contained in an interval $[\check{\check{c}}, \hat{\hat{c}}]$ and write

$$\begin{aligned} \alpha_n &= \frac{\hat{\hat{c}}}{c_n} - 1 \\ \beta_n &= \frac{\check{\check{c}}}{c_n} - 1. \end{aligned}$$

Then we have

$$\sup_{\beta_n \leq y \leq \alpha_n} \left| \log(1+y) + \sum_{m=1}^{N} \frac{(-y)^m}{m} \right| \leq \sup_{\beta_n \leq y \leq \alpha_n} \frac{|y|^{N+1}}{1-|y|}.$$

For all sufficiently large n, $|\alpha_n| \leq |\beta_n|$, so

$$\sup_{\beta_n \leq y \leq \alpha_n} \frac{|y|^{N+1}}{1-|y|} = \frac{|\beta_n|^{N+1}}{1-|\beta_n|}.$$

Substituting x/c_n for $1+y$, this can be rewritten as

$$\sup_{\check{c}\le x\le\hat{c}}\left|\log\frac{x}{c_n}+\sum_{m=1}^{N}\frac{1}{m}\left(1-\frac{x}{c_n}\right)^m\right|\le\frac{|\beta_n|^{N+1}}{1-|\beta_n|}.$$

Thus it suffices to show that

$$\frac{|\beta_n|^{N+1}}{1-|\beta_n|}\to 0.$$

But for all sufficiently large n,

$$\begin{aligned}\log\frac{|\beta_n|^{N+1}}{1-|\beta_n|}&=(N+1)\log\left(1-\frac{\check{\check{c}}}{c_n}\right)-\log\left(\frac{\check{\check{c}}}{c_n}\right)\\&\le-(N+1)\left(\frac{\check{\check{c}}}{c_n}\right)-\log\left(\frac{\check{\check{c}}}{c_n}\right),\end{aligned}$$

which diverges to $-\infty$. □

The main task in this Appendix is to prove Theorems 2 and 3, and the main step in this task is to establish convergence of the expectation. We divide this step into a sequence of lemmas. Define the functions

$$\Gamma_{n,k}(x)=\frac{x^{k-1}e^{-x}}{(k-1)!}\log(c_n)-\sum_{m=1}^{N}\sum_{\ell=0}^{m}\frac{(-c_n)^{-\ell}}{m}\binom{m}{\ell}\frac{x^{k+\ell-1}e^{-x}}{(k-1)!}$$

and

$$\Gamma_n(x)=\sum_{k=1}^{N}\Gamma_k(x).$$

Lemma 12. (a) *For any $k\ge 1$, $\Gamma_{n,k}(x)$ converges to*

$$\frac{x^{k-1}e^{-x}}{(k-1)!}\log x$$

uniformly on compact sets as $n\to\infty$.

(b) *$\Gamma_n(x)$ converges to $\log x$ uniformly on compact sets as $n\to\infty$.*

Proof: By the binomial theorem,

$$\begin{aligned}\sum_{\ell=0}^{m}\frac{(-c_n)^{-\ell}}{m}\binom{m}{\ell}\frac{x^{k+\ell-1}e^{-x}}{(k-1)!}&=\sum_{\ell=0}^{m}\frac{1}{m}\left(-c_n\right)^{m-m-\ell}\binom{m}{\ell}\frac{x^{k+\ell-1}e^{-x}}{(k-1)!}\\&=\frac{1}{m}(-c_n)^{-m}\left(x-c_n\right)^m\frac{x^{k-1}e^{-x}}{(k-1)!}.\end{aligned}$$

Thus we have

$$\begin{aligned}\Gamma_{n,k}(x)&=\frac{x^{k-1}e^{-x}}{(k-1)!}\log(c_n)-\sum_{m=1}^{N}\frac{(-c_n)^{-m}}{m}(x-c_n)^m\frac{x^{k-1}e^{-x}}{(k-1)!}\\&=\left[\log(c_n)-\sum_{m=1}^{N}\frac{(-c_n)^{-m}}{m}(x-c_n)^m\right]\frac{x^{k-1}e^{-x}}{(k-1)!}\end{aligned}$$

and

$$\Gamma_n(x)=\left[\log(c_n)-\sum_{m=1}^{N}\frac{(-c_n)^{-m}}{m}(x-c_n)^m\right]\left[\sum_{k=1}^{N}\frac{x^{k-1}e^{-x}}{(k-1)!}\right].$$

But

$$\log(c_n) - \sum_{m=1}^{N} \frac{(-c_n)^{-m}}{m}(x-c_n)^m \to \log x$$

uniformly on compact sets by Lemma 1. This establishes (a). To show (b), note that

$$\sum_{k=1}^{N} \frac{x^{k-1}e^{-x}}{(k-1)!} \to 1$$

uniformly on compact sets as well. □

We next show a similar result but with the Poisson mass functions in $\Gamma_{n,k}(x)$ and $\Gamma_n(x)$ replaced by binomial mass functions. Define the function

$$\tilde{\Gamma}_{n,k}^{(h)}(x) = \binom{h}{k-1}\left(\frac{x}{h}\right)^{k-1}\left(1-\frac{x}{h}\right)^{h-k+1}\log(c_n) - \sum_{m=1}^{N}\sum_{\ell=0}^{m}\frac{(-c_n)^{-\ell}}{m}\binom{m}{\ell}\frac{(k+\ell-1)!}{(k-1)!}\binom{h}{k+\ell-1}\left(\frac{x}{h}\right)^{k+\ell-1}\left(1-\frac{x}{h}\right)^{h-k-\ell+1}$$

and

$$\tilde{\Gamma}_{n}^{(h)}(x) = \sum_{k=1}^{N}\tilde{\Gamma}_{n,k}^{(h)}(x).$$

Lemma 13. (a) *For any $k \geq 1$, $\tilde{\Gamma}_{n,k}^{(n)}(x)$ converges to*

$$\frac{x^{k-1}e^{-x}}{(k-1)!}\log x$$

uniformly on compact sets as $n \to \infty$.

(b) $\tilde{\Gamma}_{n}^{(n)}(x)$ *converges to* $\log x$ *uniformly on compact sets as* $n \to \infty$.

(c) *For any $k \geq 1$, $\tilde{\Gamma}_{n,k}^{(n-1)}(\frac{n-1}{n}x)$ converges to*

$$\frac{x^{k-1}e^{-x}}{(k-1)!}\log x$$

uniformly on compact sets as $n \to \infty$.

(d) $\tilde{\Gamma}_{n}^{(n-1)}(\frac{n-1}{n}x)$ *converges to* $\log x$ *uniformly on compact sets as* $n \to \infty$.

Proof: Consider the difference between $\Gamma_{n,k}(x)$ and $\tilde{\Gamma}_{n,k}^{(n)}(x)$,

$$|\Gamma_{n,k}(x) - \tilde{\Gamma}_{n,k}^{(n)}(x)| \leq \left|\frac{x^{k-1}e^{-x}}{(k-1)!} - \binom{n}{k-1}\left(\frac{x}{n}\right)^{k-1}\left(1-\frac{x}{n}\right)^{n-k+1}\right|\log(c_n) + \sum_{m=1}^{N}\sum_{\ell=0}^{m}\frac{c_n^{-\ell}}{m}\binom{m}{\ell}\frac{x^{k+\ell-1}}{(k-1)!}\left|e^{-x} - \frac{(k+\ell-1)!}{n^{k+\ell-1}}\binom{n}{k+\ell-1}\left(1-\frac{x}{n}\right)^{n-k-\ell+1}\right|.$$

Fix a compact set in $[\check{c}, \hat{c}]$. Applying Corollary 1 in Appendix F gives, for all sufficiently large n,

$$|\Gamma_{n,k}(x) - \tilde{\Gamma}_{n,k}^{(n)}(x)| \leq \frac{\log(c_n)}{\sqrt{n}}\frac{x^{k-1}}{(k-1)!} + \sum_{m=1}^{N}\sum_{\ell=0}^{m}\frac{c_n^{-\ell}}{m}\binom{m}{\ell}\frac{x^{k+\ell-1}}{(k-1)!}\frac{1}{\sqrt{n}}.$$

Note that the $c_n^{-\ell}$ factor in the second term can be omitted since $c_n \geq 1$ for all sufficiently large n. Applying the binomial theorem then yields

$$\begin{aligned}|\Gamma_{n,k}(x) - \tilde{\Gamma}_{n,k}^{(n)}(x)| &\leq \frac{\log(c_n)}{\sqrt{n}} \frac{x^{k-1}}{(k-1)!} + \sum_{m=1}^{N} \frac{1}{m}(1+x)^m \frac{x^{k-1}}{(k-1)!} \frac{1}{\sqrt{n}} \\ &\leq \frac{\log(c_n)}{\sqrt{n}} \frac{x^{k-1}}{(k-1)!} + \frac{N(1+x)^N x^{k-1}}{(k-1)!\sqrt{n}}.\end{aligned}$$

In particular, we have

$$\sup_{\check{c} \leq x \leq \hat{c}} |\Gamma_{n,k}(x) - \tilde{\Gamma}_{n,k}^{(n)}(x)| \leq \frac{\log(c_n)}{\sqrt{n}} \frac{\hat{c}^{k-1}}{(k-1)!} + \frac{N(1+\hat{c})^N \hat{c}^{k-1}}{(k-1)!\sqrt{n}}.$$

Corollary 2 in Appendix F implies that the right-hand side tends to zero as $n \to \infty$. This fact along with Lemma 12(a) establishes (a). Continuing,

$$\begin{aligned}\sup_{\check{c} \leq x \leq \hat{c}} |\Gamma_n(x) - \tilde{\Gamma}_n^{(n)}(x)| &\leq \sum_{k=1}^{N} \left[\frac{\log(c_n)}{\sqrt{n}} \frac{\hat{c}^{k-1}}{(k-1)!} + \frac{N(1+\hat{c})^N \hat{c}^{k-1}}{(k-1)!\sqrt{n}} \right] \\ &= \frac{\log(c_n)}{\sqrt{n}} \exp(\hat{c}) + \frac{N(1+\hat{c})^N \exp(\hat{c})}{\sqrt{n}},\end{aligned}$$

which also tends to zero as $n \to \infty$. This fact along with Lemma 12(b) establishes (b). To show (c), note that the proof of (a) also works for $\tilde{\Gamma}_{n,k}^{(n-1)}(x)$. In addition, $(n-1)x/n$ converges to x uniformly on compact sets. Since

$$\frac{x^{k-1} e^{-x} \log x}{(k-1)!}$$

is uniformly continuous on compact sets, (c) follows. The proof of (d) is analogous. □

Lemma 14. (a) *For any* $k \geq 1$,

$$\lim_{n \to \infty} E[\zeta_{n,k}] = \int_C \frac{x^{k-1} e^{-x}}{(k-1)!} \log x \, dP(x)$$

(b)

$$\lim_{n \to \infty} E[\zeta_n] = \int_C \log x \, dP(x)$$

Proof: Observe that

$$E\left[\phi_{n,k}\right] = \sum_{a \in A_n} \binom{n-1}{k} (p_n(a))^k (1 - p_n(a))^{n-1-k} p_n(a).$$

Thus we have

$$\begin{aligned}E[\zeta_{n,k}] = &\sum_{a \in A_n} \log(c_n) \binom{n-1}{k-1} (p_n(a))^{k-1} (1 - p_n(a))^{n-k} p_n(a) \\ &- \sum_{a \in A_n} \sum_{m=1}^{N} \sum_{\ell=0}^{m} \frac{(-c_n)^{-\ell}}{m} \binom{m}{\ell} \cdot \\ &\frac{(k+\ell-1)!}{(k-1)!} \binom{n-1}{k+\ell-1} (p_n(a))^{k+\ell-1} (1 - p_n(a))^{n-k-\ell} p_n(a),\end{aligned}$$

which can be rewritten as

$$E[\zeta_{n,k}] = \int_C \tilde{\Gamma}_{n,k}^{(n-1)}\left(\frac{(n-1)x}{n}\right)\, dP_n(x).$$

Similarly,

$$E[\zeta_n] = \int_C \tilde{\Gamma}_{n}^{(n-1)}\left(\frac{(n-1)x}{n}\right)\, dP_n(x).$$

Now by Lemma 13(c) and (d),

$$\tilde{\Gamma}_{n,k}^{(n-1)}\left(\frac{(n-1)x}{n}\right) \to \frac{x^{k-1}e^{-x}}{(k-1)!}\log x \text{ uniformly on } C$$
$$\tilde{\Gamma}_{n}^{(n-1)}\left(\frac{(n-1)x}{n}\right) \to \log x \text{ uniformly on } C.$$

Furthermore, $P_n \to P$ weakly. This implies (a) and (b). □

Lemma 15. (a) *For any* $k \geq 1$,

$$\lim_{n\to\infty} |\zeta_{n,k} - E[\zeta_{n,k}]| = 0 \quad \textit{a.s.}$$

(b)

$$\lim_{n\to\infty} |\zeta_{n} - E[\zeta_{n}]| = 0 \quad \textit{a.s.}$$

Proof: For both parts, it suffices to show that

$$\lim_{n\to\infty} \sum_{k=1}^{N} |\zeta_{n,k} - E[\zeta_{n,k}]| = 0 \quad \text{a.s.}$$

Now

$$\sum_{k=1}^{N} |\zeta_{n,k} - E[\zeta_{n,k}]| \leq \sum_{k=1}^{N}\sum_{m=1}^{N}\sum_{\ell=0}^{m} \frac{c_n^{-\ell}}{m}\binom{m}{\ell}\frac{(k+\ell-1)!}{(k-1)!}\cdot |\phi_{n,k+\ell-1} - E\left[\phi_{n,k+\ell-1}\right]|$$
$$+ \sum_{k=1}^{N} \log(c_n)\left|\phi_{n,k-1} - E\left[\phi_{n,k-1}\right]\right|.$$

Write

$$\eta_n = \sup_{0\leq k\leq 2N-1} \left|\phi_{n,k} - E\left[\phi_{n,k}\right]\right|.$$

Then

$$\sum_{k=1}^{N} |\zeta_{n,k} - E[\zeta_{n,k}]| \leq \sum_{k=1}^{N}\sum_{m=1}^{N}\sum_{\ell=0}^{m} \frac{c_n^{-\ell}}{m}\binom{m}{\ell}\frac{(k+\ell-1)!}{(k-1)!}\frac{1}{n^{1/3}}(\eta_n\cdot n^{1/3}) + \log(c_n)\cdot N\cdot\eta_n.$$

Now $\eta_n \cdot n^{1/3} \to 0$ a.s. and $\eta_n \cdot N \to 0$ a.s. by Lemma 11, and for all sufficiently large n,

$$\frac{1}{n^{1/3}}\sum_{k=1}^{N}\sum_{m=1}^{N}\sum_{\ell=0}^{m} \frac{c_n^{-\ell}}{m}\binom{m}{\ell}\frac{(k+\ell-1)!}{(k-1)!} \overset{(a)}{\leq} \frac{1}{n^{1/3}}\sum_{k=1}^{N}\sum_{m=1}^{N}\sum_{\ell=0}^{m}\binom{m}{\ell}\frac{(k+m-1)!}{(k-1)!}$$
$$= \frac{1}{n^{1/3}}\sum_{k=1}^{N}\sum_{m=1}^{N} 2^m\frac{(k+m-1)!}{(k-1)!}$$
$$\leq \frac{N^2 2^N (2N)!}{n^{1/3}},$$

where in (a) we have used the fact that $c_n \geq 1$ eventually. In Appendix F, we verify that this last quantity tends to zero (Lemma 24). □

Theorems 2 and 3 follow immediately from Lemmas 14 and 15 and Propositions 2 and 3.

Proof of Theorem 4: We have

$$H(p_n) - \log n + \zeta_n = \zeta_n - \int \log x \; dP_n(x),$$

but both terms on the right-hand side converge to

$$\int \log x \; dP(x),$$

the latter because $P_n \to P$ weakly and $\log x$ is bounded and continuous over C. □

Proof of Theorem 6: By the triangle inequality,

$$\begin{aligned}\left| D_{n,k} - \frac{\zeta_{n,k+1}}{\phi_{n,k}} - \log \frac{\phi_{n,k-1}}{k\phi_{n,k}} \right| \leq & \left| D_{n,k} - \frac{1}{\lambda_k} \int_C \frac{x^k e^{-x}}{k!} \log x \; dP(x) - \log \frac{\lambda_{k-1}}{k\lambda_k} \right| \\ & + \left| \frac{1}{\lambda_k} \int_C \frac{x^k e^{-x}}{k!} \log x \; dP(x) - \frac{\zeta_{n,k+1}}{\phi_{n,k}} \right| \\ & + \left| \log \frac{\lambda_{k-1}}{k\lambda_k} - \log \frac{\phi_{n,k-1}}{k\phi_{n,k}} \right|.\end{aligned}$$

But the three terms on the right-hand side tend to zero by Proposition 6, Theorem 2, and Proposition 7. □

APPENDIX D
TWO-SEQUENCE LIMITS

Lemma 16.

$$\lim_{n\to\infty} E\left[\frac{1}{n} \log p_n(\mathbf{Y})\right] + \log n = \int_{C^2} \log x \; dQ(x,y).$$

Proof: Let $B'_{n,k}$ denote the set of symbols in A_n that appear k times in $\mathbf{Y}$. Then we may write

$$\frac{1}{n} \log p_n(\mathbf{Y}) + \log n = \frac{1}{n} \sum_{k=1}^{n} \sum_{a \in B'_{n,k}} k \log(np_n(a)).$$

Then observe that for any $1 \leq k \leq n$,

$$\begin{aligned} E\left[\frac{k}{n} \sum_{a \in B'_{n,k}} \log(np_n(a))\right] &= \frac{k}{n} \sum_{a \in A_n} \binom{n}{k} (q_n(a))^k (1 - q_n(a))^{n-k} \log(np_n(a)) \\ &= \sum_{a \in A_n} \binom{n-1}{k-1} (q_n(a))^{k-1} (1 - q_n(a))^{n-k} \log(np_n(a)) q_n(a). \end{aligned}$$

Thus

$$\begin{aligned} E\left[\frac{1}{n} \sum_{k=1}^{n} \sum_{a \in B'_{n,k}} k \log(np_n(a))\right] &= \sum_{a \in A_n} \log(np_n(a)) q_n(a) \\ &= \int_{C^2} \log x \; dQ_n(x,y). \end{aligned}$$

Now $Q_n \to Q$ weakly and $\log x$ is bounded and continuous over C^2, so

$$\lim_{n\to\infty} \int_{C^2} \log x \; dQ_n(x,y) = \int_{C^2} \log x \; dQ(x,y).$$

□

Lemma 17.

$$\lim_{n\to\infty} \left| \frac{1}{n} \log(p_n(\mathbf{Y})) - E\left[\frac{1}{n}\log p_n(\mathbf{Y})\right] \right| = 0 \quad \textit{a.s.}$$

The proof of this result is virtually identical to that of Lemma 6 and is omitted. Proposition 8 follows immediately from this result and the previous one.

APPENDIX E
TWO-SEQUENCE ESTIMATORS

Lemma 18. *For any $k \geq 0$,*

$$\lim_{n\to\infty} E[\psi_{n,k}] = \int_{C^2} \frac{x^k e^{-x}}{k!} \; dQ(x,y).$$

Proof: For any $k \geq 0$,

$$\begin{aligned}
E\left[\psi_{n,k}\right] &= \sum_{a\in A_n} \sum_{j=1}^{n} \frac{j}{n} \binom{n}{k} (p_n(a))^k (1-p_n(a))^{n-k} \binom{n}{j} (q_n(a))^j (1-q_n(a))^{n-j} \\
&= \sum_{a\in A_n} \sum_{j=1}^{n} \binom{n}{k} (p_n(a))^k (1-p_n(a))^{n-k} \binom{n-1}{j-1} (q_n(a))^j (1-q_n(a))^{n-j} \\
&= \sum_{a\in A_n} \binom{n}{k} (p_n(a))^k (1-p_n(a))^{n-k} q_n(a) \qquad (25) \\
&= \sum_{a\in A_n} g_k^n(np_n(a)) q_n(a) \\
&= \int_{C^2} g_k^n(x) \; dQ_n(x,y).
\end{aligned}$$

The result then follows as in the proof of Lemma 3. □

Unlike the other quantities examined in this paper, McDiarmid's inequality is not strong enough to prove concentration for $\psi_{n,k}$. We proceed by creating a Doob martingale and applying the Azuma-Hoeffding inequality directly.

Lemma 19. *For any $\delta > 0$,*

$$\lim_{n\to\infty} \sup_{0\leq k\leq n} |\psi_{n,k} - E\left[\psi_{n,k}\right]| \cdot n^{1/2-\delta} = 0 \quad \textit{a.s.}$$

Proof: Fix k, and define the martingale $\{Z_i\}_{i=0}^{2n}$ by

$$\begin{aligned}
Z_0 &= E\left[\psi_{n,k}\right] \\
Z_i &= E\left[\psi_{n,k} | X_1, \ldots, X_i\right] \quad i = 1, \ldots, n, \quad \text{and} \\
Z_i &= E\left[\psi_{n,k} | \mathbf{X}, Y_1, \ldots, Y_{i-n}\right] \quad i = n+1, \ldots, 2n.
\end{aligned}$$

Fix $1 \leq i \leq n$, and let $\tilde{X}_i$ be identically distributed with X_i and independent of $(\mathbf{X}, \mathbf{Y})$. Let $B_{n,k}$ denote the set of symbols that appear k times in $\mathbf{X}$, and let $\tilde{B}_{n,k}$ denote the set of symbols that appear k times in $(X_1, \ldots, X_{i-1}, \tilde{X}_i, X_{i+1}, \ldots, X_n)$. Write

$$\psi_{n,k} = \frac{1}{n} \sum_{j=1}^{n} 1(Y_j \in B_{n,k}) = f_k(\mathbf{X}, \mathbf{Y}).$$

Then

$$\begin{aligned}|Z_i - Z_{i-1}| &= |E[f_k(\mathbf{X},\mathbf{Y})|X_1,\ldots,X_i] - E[f_k(\mathbf{X},\mathbf{Y})|X_1,\ldots,X_{i-1}]| \\ &= |E[f_k(\mathbf{X},\mathbf{Y})|X_1,\ldots,X_i] \\ &\qquad - E[f_k(X_1,\ldots,\tilde{X}_i,\ldots,X_n,\mathbf{Y})|X_1,\ldots X_i]| \\ &\le E[|f_k(\mathbf{X},\mathbf{Y}) - f_k(X_1,\ldots,\tilde{X}_i,\ldots,X_n,\mathbf{Y})||X_1,\ldots,X_i].\end{aligned}$$

Next note that

$$f_k(\mathbf{X},\mathbf{Y}) - f_k(X_1,\ldots,X_{i-1},\tilde{X}_i,X_{i+1},\ldots,X_n,\mathbf{Y}) = \frac{1}{n}\sum_{j=1}^{n}\left[1(Y_j \in B_{n,k}) - 1(Y_j \in \tilde{B}_{n,k})\right]$$

and

$$|1(Y_j \in B_{n,k}) - 1(Y_j \in \tilde{B}_{n,k})| \le 1(Y_j = X_i) + 1(Y_j = \tilde{X}_i).$$

Thus

$$|Z_i - Z_{i-1}| \le \frac{1}{n}\sum_{j=1}^{n} E[1(X_i = Y_j) + 1(\tilde{X}_i = Y_j)|X_1,\ldots,X_i].$$

But

$$E[1(X_i = Y_j)|X_1,\ldots,X_i] \le \frac{\hat{c}}{n}$$

and likewise

$$E[1(\tilde{X}_i = Y_j)|X_1,\ldots,X_i] \le \frac{\hat{c}}{n}.$$

Thus

$$|Z_i - Z_{i-1}| \le \frac{2\hat{c}}{n} \quad \text{a.s.}$$

for $i = 1,\ldots,n$. Similarly, fix $n+1 \le i \le 2n$, and let $\tilde{Y}_{i-n}$ be identically distributed with Y_{i-n} and independent of $(\mathbf{X},\mathbf{Y})$. Then

$$\begin{aligned}|Z_i - Z_{i-1}| &= |E[f_k(\mathbf{X},\mathbf{Y})|\mathbf{X},Y_1,\ldots,Y_{i-n}] \\ &\qquad - E[f_k(\mathbf{X},Y_1,\ldots,\tilde{Y}_{i-n},\ldots,Y_n)|\mathbf{X},Y_1,\ldots,Y_{i-n}]| \\ &\le E[|f_k(\mathbf{X},\mathbf{Y}) - f_k(\mathbf{X},Y_1,\ldots,\tilde{Y}_{i-n},\ldots,Y_n)||\mathbf{X},Y_1,\ldots,Y_{i-n}].\end{aligned}$$

But

$$|f_k(\mathbf{X},\mathbf{Y}) - f_k(\mathbf{X},Y_1,\ldots,\tilde{Y}_{i-n},\ldots,Y_n)| \le \frac{1}{n} \quad \text{a.s.}$$

Thus

$$|Z_i - Z_{i-1}| \le \frac{1}{n} \quad \text{a.s.}$$

for $i = n+1,\ldots,2n$. It follows that

$$|Z_i - Z_{i-1}| \le \frac{\max(1,2\hat{c})}{n} \quad \text{a.s.}$$

for all $i = 1,\ldots,2n$. Thus by the Azuma-Hoeffding inequality [36, Corollary 2.4.7],

$$\Pr\left(|\psi_{n,k} - E\left[\psi_{n,k}\right]| > \tau\right) \le 2\exp\left(-\frac{n\tau^2}{4\max(1,2\hat{c})^2}\right),$$

which implies

$$\Pr\left(|\psi_{n,k} - E\left[\psi_{n,k}\right]| \cdot n^{1/2-\delta} > \tau\right) \le 2\exp\left(-\frac{n^{2\delta}\tau^2}{4\max(1,\hat{c}^2)}\right).$$

By the union bound, this gives

$$\Pr\left(\sup_{0\le k\le n}|\psi_{n,k}-E\left[\psi_{n,k}\right]|\cdot n^{1/2-\delta}>\tau\right)\le 2(n+1)\exp\left(-\frac{n^{2\delta}\tau^2}{4\max(1,\hat{c}^2)}\right).$$

The result then follows by the Borel-Cantelli lemma. □

Proposition 10 is an immediate consequence of the previous two lemmas.

Lemma 20.

$$\lim_{n\to\infty}E[\xi_n]=\int_{C^2}\log x\;dQ(x,y).$$

Proof: From (25)

$$E[\psi_{n,k}]=\sum_{a\in A_n}\binom{n}{k}(p_n(a))^k(1-p_n(a))^{n-k}q_n(a).$$

Thus

$$\begin{aligned}E[\xi_n]&=\sum_{a\in A_n}\sum_{k=1}^{N}\binom{n}{k-1}(p_n(a))^{k-1}(1-p_n(a))^{n-k+1}\log(c_n)q_n(a)\\&\quad+\sum_{a\in A_n}\sum_{k=1}^{N}\sum_{m=1}^{N}\sum_{\ell=0}^{m}\frac{(-1)^{1-\ell}}{m}(c_n)^{-\ell}\binom{m}{\ell}\frac{(k+\ell-1)!}{(k-1)!}\\&\quad\cdot\binom{n}{k+\ell-1}(p_n(a))^{k+\ell-1}(1-p_n(a))^{n-k-\ell+1}q_n(a)\\&=\sum_{a\in A_n}\tilde{\Gamma}_n^{(n)}(np_n(a))q_n(a)\\&=\int_{C^2}\tilde{\Gamma}_n^{(n)}(x)\;dQ_n(x,y).\end{aligned}$$

Now since $\tilde{\Gamma}_n^{(n)}$ converges uniformly to $\log x$ on C^2 by Lemma 13(b) and $Q_n\to Q$ weakly, this converges to

$$\int_{C^2}\log x\;dQ(x,y).$$

□

Lemma 21.

$$\lim_{n\to\infty}|\xi_n-E[\xi_n]|=0\quad\textit{a.s.}$$

Proof: We have

$$\begin{aligned}|\xi_n-E[\xi_n]|&\le\sum_{k=1}^{N}|\psi_{n,k-1}-E\left[\psi_{n,k-1}\right]|\log(c_n)\\&\quad+\sum_{k=1}^{N}\sum_{m=1}^{N}\sum_{\ell=0}^{m}\frac{(c_n)^{-\ell}}{m}\binom{m}{\ell}\frac{(k+\ell-1)!}{(k-1)!}\left|\psi_{n,k+\ell-1}-E\left[\psi_{n,k+\ell-1}\right]\right|\\&\le\frac{N\log(c_n)}{n^{1/3}}\sup_{0\le k\le N-1}|\psi_{n,k}-E\left[\psi_{n,k}\right]|\cdot n^{1/3}\\&\quad+\sup_{0\le k\le 2N-1}\left[|\psi_{n,k}-E\left[\psi_{n,k}\right]|\right]\cdot n^{1/3}\cdot\\&\quad\frac{1}{n^{1/3}}\sum_{k=1}^{N}\sum_{m=1}^{N}\sum_{\ell=0}^{m}\frac{(c_n)^{-\ell}}{m}\binom{m}{\ell}\frac{(k+\ell-1)!}{(k-1)!}.\end{aligned}$$

Now

$$\lim_{n\to\infty} \sup_{0\le k\le n} |\psi_{n,k} - E[\psi_{n,k}]| \cdot n^{1/3} = 0 \quad \text{a.s.}$$

by Lemma 19. Also,

$$\frac{N\log(c_n)}{n^{1/3}} \to 0$$

and for all sufficiently large n,

$$\begin{aligned}\frac{1}{n^{1/3}}\sum_{k=1}^{N}\sum_{m=1}^{N}\sum_{\ell=0}^{m}\frac{(c_n)^{-\ell}}{m}\binom{m}{\ell}\frac{(k+\ell-1)!}{(k-1)!} &\overset{(a)}{\le} \frac{1}{n^{1/3}}\sum_{k=1}^{N}\sum_{m=1}^{N}\sum_{\ell=0}^{m}\binom{m}{\ell}\frac{(k+m-1)!}{(k-1)!} \\ &= \frac{1}{n^{1/3}}\sum_{k=1}^{N}\sum_{m=1}^{N}2^m\frac{(k+m-1)!}{(k-1)!} \\ &\le \frac{N^2 2^N (2N)!}{n^{1/3}},\end{aligned}$$

where in (a) we have used the fact that $c_n \ge 1$ eventually. The result then follows from Lemma 24 in Appendix F. □

Proof of Theorem 7: The previous two lemmas together imply (18). Then (18) and Proposition 8 together imply (19). □

Proof of Theorem 8: By Proposition 9,

$$D(q_n||p_n) \to \int_{C^2} \log y \; dQ(x,y) - \int_{C^2} \log x \; dQ(x,y).$$

But

$$\xi_n \to \int_{C^2} \log x \; dQ(x,y)$$

by Theorem 7 and

$$\zeta_n \to \int_{C^2} \log y \; dQ(x,y).$$

by Theorem 3. □

APPENDIX F
ANCILLARY RESULTS

Lemma 22. *For any $i \ge 0$, $j \ge 0$, and $n \ge 0$ such that $i + j \le n$,*

$$\left|\frac{(n-i)!}{(n-i-j)!n^j} - 1\right| \le \frac{j(i+j)}{n}.$$

Proof: Observe that

$$0 \le 1 - \frac{(n-i)!}{(n-i-j)!n^j} \le 1 - \left(\frac{n-i-j}{n}\right)^j.$$

Now if

$$f(x) = (1-x)^j - (1-jx),$$

then $f(0) = 0$ and $f'(x) \ge 0$ for all $0 \le x \le 1$. Thus $f(x) \ge 0$ if $0 \le x \le 1$, which implies

$$1 - \left(1 - \frac{i+j}{n}\right)^j \le \frac{j(i+j)}{n}.$$

□

Lemma 23. *For any n, $k \le n$, and $y \ge 0$,*

$$\sup_{0\le x\le y}\left|\exp(-x) - \left(1-\frac{x}{n}\right)^{n-k}\right| \le \frac{y^{n-k+1}e^y}{(n-k+1)!} + \frac{(k+1)y+y^2}{n}\cdot e^y.$$

Proof: By the binomial theorem,

$$\left(1-\frac{x}{n}\right)^{n-k} = \sum_{i=0}^{n-k}\binom{n-k}{i}\left(-\frac{x}{n}\right)^i.$$

Thus

$$\left|\exp(-x) - \left(1-\frac{x}{n}\right)^{n-k}\right| \le \sum_{i=0}^{n-k}\left|\binom{n-k}{i}\left(-\frac{x}{n}\right)^i - \frac{(-x)^i}{i!}\right| + \left|\sum_{i=n-k+1}^{\infty}\frac{(-x)^i}{i!}\right|.$$

By Taylor's theorem, if $0 \le x \le y$,

$$\left|\sum_{i=n-k+1}^{\infty}\frac{(-x)^i}{i!}\right| \le \frac{y^{n-k+1}e^y}{(n-k+1)!}.$$

And by Lemma 22,

$$\begin{aligned}\sum_{i=0}^{n-k}\left|\binom{n-k}{i}\left(-\frac{x}{n}\right)^i - \frac{(-x)^i}{i!}\right| &= \sum_{i=0}^{n-k}\frac{x^i}{i!}\left|\frac{(n-k)!}{(n-k-i)!n^i} - 1\right| \\ &\le \sum_{i=0}^{\infty}\frac{y^i}{i!}\frac{i(k+i)}{n} \\ &= \frac{(k+1)y+y^2}{n}\cdot e^y.\end{aligned}$$

□

Corollary 1. *For any compact set $\mathcal{C}$, any $\delta \in (0,1)$, all sufficiently large n, and all $k \le N$ and $\ell \le N$,*

$$\sup_{x\in\mathcal{C}}\left|\exp(-x) - \frac{n!}{(n-k-\ell)!n^{k+\ell}}\left(1-\frac{x}{n}\right)^{n-k-\ell}\right| \le \frac{1}{n^\delta}.$$

Proof: Suppose that $\mathcal{C}$ is bounded from above by $\hat{c}$. Then applying Lemmas (22) and (23) gives

$$\begin{aligned}\left|\exp(-x) - \frac{n!}{(n-k-\ell)!n^{k+\ell}}\left(1-\frac{x}{n}\right)^{n-k-\ell}\right| &\le \exp(-x)\left|1 - \frac{n!}{(n-k-\ell)!n^{k+\ell}}\right| \\ &\quad + \frac{n!}{(n-k-\ell)!n^{k+\ell}}\left|\exp(-x) - \left(1-\frac{x}{n}\right)^{n-k-\ell}\right| \\ &\le \frac{(k+\ell)^2}{n} + \frac{\hat{c}^{n-k-\ell+1}e^{\hat{c}}}{(n-k-\ell+1)!} + \frac{(k+\ell+1)\hat{c}+\hat{c}^2}{n}\cdot e^{\hat{c}} \\ &\le \frac{1}{n^\delta}\end{aligned}$$

for all sufficiently large n. □

Lemma 24. *For any $y > 0$ and any $\delta > 0$,*

$$\lim_{n\to\infty}\frac{N^2y^N(2N)!}{n^\delta} = 0.$$

Proof: It suffices to show the result when $y \geq 1$. We have

$$\frac{N^2 y^N (2N)!}{n^\delta} \leq \frac{((\log n)^{\epsilon_1})^2 y^{(\log n)^{\epsilon_1}} (2(\log n)^{\epsilon_1})^{2(\log n)^{\epsilon_1}}}{n^\delta}.$$

Writing x for $\log n$, it suffices to show that

$$\frac{(x^{\epsilon_1})^2 y^{x^{\epsilon_1}} (2x^{\epsilon_1})^{2x^{\epsilon_1}}}{\exp(\delta x)} \to 0$$

as $x \to \infty$, or, equivalently, that

$$f(x) = 2\epsilon_1 \log(x) + x^{\epsilon_1} \log y + 2x^{\epsilon_1} \log(2x^{\epsilon_1}) - \delta x \to -\infty.$$

But this holds since $f(x)/x \to -\delta$ as $x \to \infty$. □

Corollary 2. *For any $y > 0$,*

$$\frac{N(1+y)^N}{\sqrt{n}} \to 0.$$

REFERENCES

[1] R. W. Lucky, *Silicon Dreams: Information, Man, and Machine*. New York: St. Martin's Press, 1989.

[2] T. M. Cover and R. C. King, "A convergent gambling estimate of the entropy of English," *IEEE Trans. Inf. Theory*, vol. 24, no. 4, pp. 413–421, July 1978.

[3] C. E. Shannon, "Prediction and entropy of printed English," *Bell Syst. Tech. J.*, vol. 30, pp. 50–64, Jan. 1951.

[4] R. H. Baayen, *Word Frequency Distributions*. Dordrecht: Kluwer, 2001.

[5] G. K. Zipf, *Human behavior and the principle of least effort: An introduction to human ecology*. New York: Hafner, 1965.

[6] K. W. Church and W. A. Gale, "A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams," *Computer Speech and Language*, vol. 5, no. 1, pp. 19–54, Jan. 1991.

[7] W. A. Gale and G. Sampson, "Good-Turing frequency estimation without tears," *Journal of Quantitative Linguistics*, vol. 2, pp. 217–237, 1995.

[8] S. Naranan and V. Balasubrahmanyan, "Models for power law relations in linguistics and information science," *Journal of Quantitative Linguistics*, vol. 5, pp. 35–61, 1998.

[9] I. J. Good, "The population frequencies of species and the estimation of population parameters," *Biometrika*, vol. 40, no. 3/4, pp. 237–64, 1953.

[10] J. C. Kieffer, "A unified approach to weak universal source coding," *IEEE Trans. Inf. Theory*, vol. 24, no. 6, pp. 674–682, Nov. 1978.

[11] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Trans. Inf. Theory*, vol. 50, no. 7, pp. 1469–81, July 2004.

[12] ——, "Always Good Turing: Asymptotically optimal probability estimation," *Science*, vol. 302, pp. 427–31, Oct. 2003.

[13] E. V. Khmaladze, "The statistical analysis of a large number of rare events," Centrum Wiskunde & Informatica (CWI), Tech. Rep. MS-R8804, 1988.

[14] E. V. Khmaladze and R. Ya Chitashvili, "Statistical analysis of a large number of rare events and related problems," *Proc. A. Razmadze Math. Inst.*, vol. 92, pp. 196–245, 1989, in Russian.

[15] C. A. J. Klaassen and R. M. Mnatsakanov, "Consistent estimation of the structural distribution function," *Scand. J. Statist.*, vol. 27, pp. 733–46, 2000.

[16] G. I. Ivchenko and Y. I. Medvedev, "Decomposable statistics and hypothesis testing. The case of small samples." *Theor. Probability Appl.*, vol. 23, no. 4, pp. 764–775, 1978.

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

[18] N. L. Johnson and S. Kotz, *Urn Models and their Application: An Approach to Modern Discrete Probability Theory*. New York: John Wiley & Sons, 1977.

[19] V. F. Kolchin, B. A. Sevast'yanov, and V. P. Chistyakov, *Random Allocations*. Washington, D.C.: V. H. Winston & Sons, 1978.

[20] F. N. David and D. E. Barton, *Combinatorial Chance*. New York: Hafner, 1962.

[21] P. Dupuis, C. Nuzman, and P. Whiting, "Large deviation asymptotics for occupancy problems," *Ann. Probab.*, vol. 32, no. 3B, pp. 2765–2818, 2004.

[22] H. E. Robbins, "Estimating the total probability of the unobserved outcomes of an experiment," *Ann. of Math. Stat.*, vol. 39, no. 1, pp. 256–7, 1968.

[23] B. H. Juang and S. H. Lo, "On the bias of the Turing-Good estimate of probabilities," *IEEE Trans. Signal Processing*, vol. 42, no. 2, pp. 496–8, 1994.

[24] W. W. Esty, "The efficiency of Good's nonparametric coverage estimator," *Ann. Statist.*, vol. 14, no. 3, pp. 1257–60, 1986.

[25] C. X. Mao and B. G. Lindsay, "A Poisson model for the coverage problem with a genomic application," *Biometrika*, vol. 89, no. 3, pp. 669–81, 2002.

[26] D. McAllester and R. E. Schapire, "On the convergence rate of Good-Turing estimators," in *Proc. 13th Annu. Conference on Comput. Learning Theory*. Morgan Kaufmann, San Francisco, 2000, pp. 1–6.

[27] E. Drukh and Y. Mansour, "Concentration bounds for unigram language models," *J. Mach. Learn. Res.*, vol. 6, pp. 1231–1264, 2005.
[28] D. McAllester and L. Ortiz, "Concentration inequalities for the missing mass and for histogram rule error," *J. Mach. Learn. Res.*, vol. 4, pp. 895–911, 2003.
[29] C. Budianu, S. Ben-David, and L. Tong, "Estimation of the number of operating sensors in large-scale sensor networks with mobile access," *IEEE Trans. Signal Processing*, vol. 54, no. 5, pp. 1703–1715, May 2006.
[30] A. B. Wagner, P. Viswanath, and S. R. Kulkarni, "Strong consistency of the Good-Turing estimator," in *IEEE Int. Symp. Inf. Theor. Proc.*, July 2006, pp. 2526–30.
[31] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken: John Wiley & Sons, 2006.
[32] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Trans. Inf. Theory*, vol. 35, no. 2, pp. 401–8, Mar. 1989.
[33] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Trans. Inf. Theory*, vol. 34, no. 2, pp. 278–86, Mar. 1988.
[34] O. Kallenberg, *Foundations of Modern Probability*, 2nd ed. New York: Springer-Verlag, 2002.
[35] C. McDiarmid, "On the method of bounded differences," in *Surveys in Combinatorics*, ser. London Math. Soc. Lecture Note Ser., J. Siemons, Ed. Cambridge Univ. Press, 1989, vol. 141, pp. 148–188.
[36] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. New York: Springer-Verlag, 1998.
[37] P. Billingsley, *Probability and Measure*, 3rd ed. New York: John Wiley & Sons, 1995.
[38] R. Durrett, *Probability: Theory and Examples*, 2nd ed. Belmont, CA: Duxbury Press, 1996.