Determining the Number of Samples Required to Estimate Entropy in Natural Sequences

Andrew D. Back, *Member, IEEE*, Daniel Angus, *Member, IEEE* and
Janet Wiles, *Member, IEEE*
School of ITEE, The University of Queensland, Brisbane, QLD, 4072 Australia.

***Abstract*—Calculating the Shannon entropy for symbolic sequences has been widely considered in many fields. For descriptive statistical problems such as estimating the N-gram entropy of English language text, a common approach is to use as much data as possible to obtain progressively more accurate estimates. However in some instances, only short sequences may be available. This gives rise to the question of how many samples are needed to compute entropy. In this paper, we examine this problem and propose a method for estimating the number of samples required to compute Shannon entropy for a set of ranked symbolic "natural" events. The result is developed using a modified Zipf-Mandelbrot law and the Dvoretzky-Kiefer-Wolfowitz inequality, and we propose an approximation which yields an estimate for the minimum number of samples required to obtain an estimate of entropy with a given confidence level and degree of accuracy.**

I. Introduction

Consider a sequence of symbolic information given by $\Upsilon_i = [\upsilon_0, \ldots, \upsilon_{n-1}]^T$, where $\upsilon_j \in \Sigma^M$, and Σ^M is an alphabet or finite nonempty set with symbolic members and dimensionality M. Suppose we are interested in the information content in such a message sequence. One way to approach this problem is by measuring what is new or novel in a given sequence. If a string consists of symbols aaa, bbb, aaa - beyond the first few words, there is little novelty or 'surprise' about the message. On the other hand, if a string consists of symbols acb, bde, zxy, eqa - then it is evident that the message has a higher degree of novelty. The idea that the randomness of a message can give a measure of the information it conveys formed the basis of Shannon's entropy[1] theory which gives a means of assigning a value to the information carried within a message [1],[2]. The way in which Shannon formulated this principle is that, given a discrete random variable X of a sequence $X = X_1, \ldots, X_i, \ldots, X_K$, where $X_i = x \in \mathbf{X}^M$, that is, x_i may take on one of M distinct values, $\mathbf{X}^M$ is a set from which the members of the sequence are drawn, and hence x_i is in this sense symbolic, where each value occurs independently with probability $p\left(x_i\right)$, $i \in [1, M]$, then the single symbol Shannon entropy is defined as:

$$H_1(X) = -\sum_{i=1}^{M} p(x_i) \log_2(p\left(x_i\right)) \qquad (1)$$

This extends to the case where the probabilities of multiple symbols occurring together are taken into account. Hence, H_2 indicates the entropy from the probabilities of two symbols occurring consecutively. We define b_i as a block of $N-1$ symbols, and is hence referred to as an (*N-1*)-gram. Further we define the *N*-gram (b_i, x_j) where x_j is an arbitrary symbol following b_i with corresponding probability $p(b_i, x_j)$. The conditional probability of x_j occurring after b_i is given by $p(x_j|b_i) = p(b_i, x_j)/p(b_i)$. Hence the general *N*-gram entropy can be defined as

$$H_N(X|B) = -\sum_{i,j} p(b_i, x_j) \log_2(p(x_j|b_i)) \qquad (2)$$

which is a measure of the information due to the statistical probability of blocks of N consecutive symbols.

Shannon demonstrated the concept of entropy by applying it to English text, obtaining estimates of entropy by using a list of 1027 words which were taken from 100,000 words of English text [3]. Entropy has since been applied to a diverse range of problems including word entropy estimation [4], statistical keyword detection [5], phylogenetic diversity measurement [6], population biology [7], language assessment of Pictish symbols [8], facial recognition [9], and interpreting gene expression data in functional genomics for drug discovery [10].

One of the limitations of computing entropy accurately is the dependence on large amounts of data, even more so when computing *N*-gram entropy. As a concrete example, in language analytics, estimates of entropy based on letter, word and *N*-gram statistics have often relied on large data sets [11], [12]. The reliance on long data sequences to estimate the probability distributions used to calculate entropy and attempts to overcome this in coding schemes is discussed in [13] where they provide an estimate of letter entropy extrapolated for infinite text lengths.

Various approaches to estimating entropy over finite sample sizes have been considered. A method of computing the entropy of dynamical systems which corrects for statistical fluctuations of the sampled data over finite sample sizes has been proposed in [14]. Estimation techniques using small datasets have been proposed in [15], and a novel approach for calculating entropy using the idea of estimating probabilities from a quadratic function of the inverse number of symbol coincidences was proposed in [16]. An online approach for estimating entropy in limited resource environments was proposed in [17]. Entropy estimation over short symbolic

[1] For convenience we will generally refer to Shannon Entropy as simply entropy with the specific formulation evident from the context.

sequences was considered in the context of dynamical time series models based on logistic maps and correlated Markov chains, where an effective shortened sequence length was proposed which accounted for the correlation effect [18].

A question which naturally arises then is how many samples are required in order to obtain an accurate estimate of entropy according to some criteria? Answering this question may provide insight into problems where limited data is available and also for online analytical information theoretic models which seek to limit data, rather than a longer term descriptive statistical approaches. In this paper, a method is proposed for estimating the number of samples required to calculate entropy of a natural sequence. The proposed model is applied to some example cases, and the implications of this new approach and potential future work is discussed.

II. Estimating Samples required for Entropy

A. Shannon Entropy

To derive the estimation method proposed later in the paper, we frame entropy in a general sense as pertaining to a set of data of particular size at a given time. Hence, given a discrete random variable X_t of a finite sequence over time with discrete probability distribution $p_s(x,t)$ at time t, the Shannon entropy is defined in this case as:

$$H_s\left(X,t\right) = -\sum_{x\in X_t} p_s(x,t)\log_2\left(p_s(x,t)\right) \tag{3}$$

where the finite sample entropy[2] $H_s\left(X,t\right)$ is computed on probabilities $p_s(X_t)$ for $X_t = [x(t_0),\ldots,x(t_{N_s-1})]^T$ over N_s consecutive samples[3] where X_t defines an alphabet or finite nonempty set with M symbolic members observed at time t.

The usual approach to calculating entropy is by estimating $p_s(x)$. Assuming some theoretical, *true* values for the probabilities $p_s(x,t)$, the accuracy of $\widehat{H}_s(X,t)$ as an estimate of $H_s(X,t)$, is determined by the accuracy with which $\{p_s(x)\}$ is estimated by $\{\widehat{p}_s(x)\}$.

For small values of $p_s(x,t)$, $|\log_2 p_s(x)|$ becomes large and hence small $p_s(x,t)$ may contribute significantly to $H_s(X,t)$. Now, given a finite set of samples, the accuracy with which $\widehat{H}_s(X,t)$ can be computed will depend on the accuracy with which the empirical probability $\widehat{p}_0(x)$ can be computed, where the probability of the most infrequent event occurrence is defined[4] as

$$\widehat{p}_0(x) = \inf\left\{\widehat{p}(x,r)|x \in X, r\in[1,M]\right\} \tag{4}$$

where the reference to time t is implicit and omitted for clarity, r refers to the probabilistic rank of the symbolic events across the M probabilities (the alphabet size) which are computed from N samples and the empirical probability is defined as

$$\widehat{p}_0(x) = \frac{n}{N_0} \tag{5}$$

where N_0 is the total number of samples for M possible probabilistic events, and n is the number of occurrences of a particular event corresponding to the probability of the specific event being measured.

Since we are considering a set of probabilistic events, the more samples used, ie, the larger N_0 is, then this will result in a larger value for n, corresponding to each event. This in turn can be expected to lead to greater accuracy in computing $\widehat{p}_0(x)$. Hence, this raises questions of how large N_0 should be and is there a relationship between M and the number of samples N_0 required to obtain a specified degree of accuracy with some level of confidence? Intuitively, one would expect that the larger the alphabet size, then the greater the number of observations required. In the next sections, we develop a method for determining the number of samples required to estimate the entropy of natural sequences derived from a given alphabet.

B. Dvoretzky-Kiefer-Wolfowitz Inequality

Given a finite set of randomly sampled iid (independent and identically distributed) observations $X_1,\ldots,X_n$ for which there exists an unknown true distribution function $F(\lambda)$ where[5]

$$F(\lambda) = P\left\{X_j \leqslant \lambda\right\} \tag{6}$$

and an empirical distribution function[6] is available, defined by

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n}\mathbf{1}_{\{X_j\leqslant x\}} \tag{7}$$

where $\mathbf{1}_{\{X_j\leqslant x\}}$ is the indicator function defined as

$$\mathbf{1}_{\{X_j\leqslant x\}} = \begin{cases} 1 & \text{if } X_j \leqslant x \\ 0 & \text{otherwise}\end{cases} \tag{8}$$

The Dvoretzky-Kiefer-Wolfowitz (DKW) inequality [20] extends earlier asymptotic results by Kolmogorov and Smirnov [21] and provides a probabilistic bound on the difference between the empirical and true distributions. A tighter probabilistic bound was obtained by Masaart [22], which allows the DKW inequality to be expressed as

$$P\left\{\sup_{x\in\mathbb{R}}\left|\widehat{F}_n(x) - F(x)\right| > \epsilon\right\} \leq 2e^{-2n\epsilon^2} \tag{9}$$

Hence, using this inequality, for every $\epsilon > 0$, and confidence level specified by ζ', with $\zeta = (1-\zeta')$, $\zeta > 0$, it is possible

[2] A sample refers to the discrete measurements or observations of probabilistic events. The sample size in this context is defined as the number of samples in a finite set of data extracted from some larger set. Note that while Shannon entropy is commonly specified using base 2, it is also possible to formulate entropy using any base. Taking it to the base of the alphabet size would have the advantage that the entropy would be normalized [19].

[3] As a notational convenience, we designate $t_p = t - p$, where $x(t-1)$ indicates the value of a variable x one sample before $x(t)$ which in some cases is indicated as x_t. Samples refer to the values of the discrete random variable events, eg $x(t_0),\ldots,x(t_{p-1})$. The sample size in this context is defined as the number of samples in a finite set of data extracted from some larger set. Our interest here is particularly on entropy defined over a finite set of samples, designated initially as $H_s(X,t)$.

[4] We expect there to be a greatest lower bound on the set of probabilities, but a precise minimum does not necessarily exist.

[5] Following historical convention, we will at times use notation where $P(x)$ refers to probabilities, typically distribution probabilities associated with an event, and $p_i(x)$ refers also to probabilities, where $p_i(x) = P(X = x_i)$.

[6] In this section we introduce the notation $\widehat{p}(x)$ to represent empirical probability.

to calculate $N(\epsilon, \zeta)$ such that if $n \geq N(\epsilon, \zeta)$ [23], then

$$P\left\{\sup_{x\in\mathbb{R}}\left|\widehat{F}_n(x) - F(x)\right| > \epsilon\right\} \leq 1 - \zeta'. \tag{10}$$

where

$$1 - \zeta' = 2e^{-2n\epsilon_0^2} \tag{11}$$

A novel application of the DKW inequality to determining a probabilistic upper bound on the entropy of an unknown distribution based on a sample from that distribution was given recently by Learned-Miller & DeStefano [24]. In the work presented here, we consider the number of samples required to obtain a specified degree of accuracy with a given confidence level. Hence, suppose we wish to determine n, such that with some ζ', the maximum difference between $\widehat{F}_n(x)$ and $F(x)$ is ϵ, then it follows that a solution may be found as

$$P\left\{\sup_{x\in\mathbb{R}}\left|\widehat{F}_n(x) - F(x)\right| < \epsilon\right\} \leq \zeta'. \tag{12}$$

Now, the DKW inequality specified a bound for the difference in distribution functions $\widehat{F}_n(x)$ and $F(x)$. For discrete random variables with discrete probabilities, we have

$$\begin{aligned} \widehat{F}_n(r) &= P(X \leqslant r) \qquad (13)\\ &= \sum_{k=0}^{r} \widehat{p}(k) \qquad (14)\end{aligned}$$

Hence, we obtain

$$P\left\{\sup_{r\in\mathbb{N}}\left|\sum_{k=0}^{r}\left(\widehat{p}(k) - p(k)\right)\right| < \epsilon_0\right\} \leq \zeta' \tag{15}$$

Now, rearranging (11), we obtain

$$n = \frac{1}{2\epsilon_0^2}\ln\left(\frac{2}{1-\zeta'}\right). \tag{16}$$

Through consideration of $H_s(X, t)$, it is clear that ϵ_0 can be specified by the number of samples required to discriminate between the two closest probabilities used in the entropy calculation. This can be expressed in terms of the DKW inequality as

$$P\left\{\sup_{r\in\mathbb{N}}\left|\widehat{p}_n(r-1) - \widehat{p}(r)\right| < \epsilon_r\right\} \leq \zeta' \tag{17}$$

The next task is then to determine ϵ_r. Hence, given a finite set of randomly sampled iid observations $X_1, \ldots, X_n$ for which it is assumed that there exists a corresponding set of monotonic relative frequencies to give the estimated probabilities for a set of M possible events, i.e. $\widehat{p}(x_1), \ldots, \widehat{p}(x_M)$, then define

$$\Delta_0 = \inf_i\left\{\left|\widehat{p}(x_i) - \widehat{p}(x_{i+1})\right|\right\} \qquad i = 1, \ldots, M-1 \tag{18}$$

Now, it follows that $\epsilon_r < \Delta_0$, moreover, it can be observed that if $\epsilon_r = \Delta_0/2$, then at worst, we cannot still reliably discriminate between $p_s(r)$ and $p_s(r-1)$, since for this case, $\sup\left\{\widehat{p}_c(r)\right\} = \inf\left\{\widehat{p}_c(r-1)\right\}$, where

$$\widehat{p}_c(r) \in \mathbb{R} : \widehat{p}(r) - \epsilon_r \leq \widehat{p}_c(r) \leq \widehat{p}(r) + \epsilon_r \tag{19}$$

Hence, consider a rule for determining the entropy probabilities as a function of rank, alphabet size and some parametrization, i.e. $\widehat{p}(r) = f(r, M; \theta)$. From (4), then for a number of observations N_0 and expected number of events n_0, with the empirical probability defined in (5), an expression is required for

$$\begin{aligned} \epsilon_r &= f\left(\widehat{p}(r)\right), \quad r \in [1, M] \qquad (20)\\ &= f\left(r, M\right) \qquad (21)\end{aligned}$$

which will be considered in the next section.

C. Probabilistic Event Model

For various natural sequences, the probability of information events can generally be ranked into monotonically decreasing order. This phenomenon has been examined extensively, in particular, it was demonstrated by Zipf's early work that the frequency of ranked words in a text occur in such a way that they can be described by a power law [25].

For natural language, it has been shown that Zipf's law approximates the distribution of probabilities of letter or words across a corpus of sufficient size for the larger probabilities [26]. The universality of Zipf's law has been challenged and in particular, it has been shown to arise as a result of the choice of rank as an independent variable [27],[28]. Nevertheless, Zipfian laws have been proven to be useful as a means of statistically characterizing the observed behaviour of symbolic sequences of data [29].

For the purpose of the development here, we do not rely on Zipf's law to provide a universal model of human language or other natural sequences (see for example, the discussions in [28],[30]). Instead, it provides a convenient statistical model which enables the transformation between the ranking of symbolic events and an estimate of their expected probabilities. Hence, it is useful in forming a model of symbolic information transmission which is organized on the basis of sentences made by words in interaction with each other (this may be considered in a general sense of natural sequences, not just human language) [31]. Thus, Zipfian based models can be useful as a means of viewing the probabilistic rankings of the symbols employed in natural sequences.

For the calculation of entropy, the accuracy will depend on the accuracy of calculating the set of probabilities. It follows that we might expect the accuracy of the probability calculations will be determined by the most infrequent event occurrences. Therefore, the number of samples required to estimate the probability of the least frequent event $p_0(r)$ determines the number of samples required to estimate the entropy for the corresponding set of data. We proceed by imposing a probabilistic model of the symbolic events for a given sequence. Since we are dealing with natural sequences of symbolic data, we consider a Zipfian model approach. The most basic form of Zipf's law models the frequency rank r of a word[7], i.e. the r-th most frequent word, by a simple

[7] A word or N-gram is not necessarily referring to human language, but indicates a specific set of sequentially occurring symbols.

inverse power law, such that the frequency of a word $f(r)$ scales according to an equation which is approximately

$$f(r) \propto \frac{1}{r^{\alpha}} \tag{22}$$

where a constant of proportionality dependent on the particular corpus may be introduced, [27] and where typically $\alpha \approx 1$. Thus, if $p_i(x)$ follows a Zipfian law, then $p_0(x) \propto 1/M$ and $p_i(x) = \varphi f(r)$.

Numerous other variations of this general law have been proposed to provide more accurate representations, including the Bradford Law [32], Lotka Law [33]. A better known approach is the Zipf-Mandelbrot law [34] which we consider below.

Given symbols $x \in \Sigma^{M+1}$ from an alphabet of size $M+1$ which includes a blank space w_s then for any random word of length L, given by $v_k(L) = \{w_s, x_1, \ldots, x_L, w_s\}$, $k = 1, \ldots, M^L$ the frequency of occurrence is determined as

$$p_i\left(L\right) = \frac{\lambda}{\left(M+1\right)^{L+2}} \qquad i = 1, \ldots, M^L \tag{23}$$

then Li showed that λ can be determined via the summation of all probabilities of such words [27], hence

$$\sum_{L=1}^{\infty} M^L p_i\left(L\right) = 1 \tag{24}$$

$$\sum_{L=1}^{\infty} M^L \frac{\lambda}{\left(M+1\right)^{L+2}} = \frac{\lambda M}{\left(M+1\right)^2} \tag{25}$$

$$\lambda = \frac{\left(M+1\right)^2}{M} \tag{26}$$

which leads to

$$p_i\left(L\right) = \frac{1}{M\left(M+1\right)^L} \qquad i = 1, \ldots, M^L \tag{27}$$

Now, defining the rank of a given word $v_k(L)$ as $r(L)$, then after performing an exponential transformation from the word length to word rank, the probability of occurrence of a given word in terms of rank can be defined as [29],[35]:

$$p(r) = \frac{\gamma}{(r+\beta)^{\alpha}} \tag{28}$$

where, for iid samples, Li showed the constants can be computed as [27]:

$$\begin{aligned} \alpha &= \frac{\log_2(M+1)}{\log_2(M)} \\ \beta &= \frac{M}{M+1} \\ \gamma &= \frac{M^{\alpha-1}}{(M-1)^{\alpha}} \end{aligned} \tag{29}$$

We introduce a normalization step as follows. Since

$$\sum_{i=1}^{M} p(i) = 1, \quad \sum_{i=1}^{\infty} \frac{\gamma}{(r+\beta)^{\alpha}} = \kappa \tag{30}$$

we introduce

$$\gamma' = \frac{\gamma}{\kappa} \tag{31}$$

which leads to

$$p(r) = \frac{\gamma'}{(r+\beta)^{\alpha}} \tag{32}$$

Now we have

$$p(r) = \frac{M^{\alpha-1}}{\left[(M-1)\left(r+\beta\right)\right]^{\alpha}} \tag{33}$$

To solve (18) we seek to determine

$$\varphi_0(x,r) = \inf_r \left\{\left|p(x,r-1) - p(x,r)\right|, x \in X, r \in [1,M]\right\} \tag{34}$$

Hence we define,

$$\theta_j = d_j - d_{j+1}, \tag{35}$$

where

$$\begin{aligned} d_j &= \frac{p(r-j) - p(r-j+1)}{\gamma'} & (36)\\ &= \frac{1}{(r-j+\beta)^{\alpha}} - \frac{1}{(r-j+1+\beta)^{\alpha}} & (37) \end{aligned}$$

then for $M \geq 1$, $\alpha > 1$, we seek to establish whether $\theta_j > 0$. For notational convenience, let $\phi_j = r - j + \beta$, where $\phi_j > 0$, $\phi_{j-1} > 0$. We define

$$d_j = \frac{1}{\phi_j^{\alpha}} - \frac{1}{\phi_{j-1}^{\alpha}} \tag{38}$$

$$\widehat{d_j} = \frac{1}{\phi_j} - \frac{1}{\phi_{j-1}} \tag{39}$$

and

$$\begin{aligned} \widehat{\theta_j} &= \left(\frac{1}{\phi_j} - \frac{1}{\phi_{j-1}}\right) - \left(\frac{1}{\phi_{j+1}} - \frac{1}{\phi_j}\right) & (40)\\ &= \frac{\phi_{j-1} - \phi_j}{\phi_j \phi_{j-1}} - \frac{\phi_j - \phi_{j+1}}{\phi_j \phi_{j+1}} & (41)\\ &= \frac{\phi_{j+1}(\phi_{j-1} - \phi_j)}{\phi_j \phi_{j-1} \phi_{j+1}} - \frac{\phi_{j-1}\left(\phi_j - \phi_{j+1}\right)}{\phi_j \phi_{j-1} \phi_{j+1}} & (42) \end{aligned}$$

Hence we define,

$$\begin{aligned} \widehat{w}_j &= \phi_{j+1}(\phi_{j-1} - \phi_j) - \phi_{j-1}(\phi_j - \phi_{j+1}) & (43)\\ &= (\phi_{j+1} - \phi_{j-1}) & (44) \end{aligned}$$

as $\phi_{j-1} - \phi_j = 1$, and $\phi_j - \phi_{j+1} = 1$. Therefore,

$$\begin{aligned} \widehat{w}_j &= (r-(j+1)+\beta) - (r-(j-1)+\beta) & (45)\\ &= -2 & (46) \end{aligned}$$

Hence, $\widehat{\theta_j} < 0$. We now define

$$\overline{\theta}_j = \left(\frac{1}{\phi_j^{\alpha}} - \frac{1}{\phi_{j-1}^{\alpha}}\right) - \left(\frac{1}{\phi_{j+1}^{\alpha}} - \frac{1}{\phi_j^{\alpha}}\right) \tag{47}$$

Consider the range of possible values for α. By inspection the maximum value of α can be found as

$$\begin{aligned} \max_M(\alpha) &= \left(\frac{\log_2(3)}{\log_2(2)}\right), \ M \in [2, \ldots, \infty] & (48)\\ &= 1.585 & (49) \end{aligned}$$

and hence it is sufficient to consider the cases for $\alpha = 1$ and

$\alpha = 2$. Therefore we can define

$$\widetilde{\theta}_j = \left(\frac{1}{\phi_j^2} - \frac{1}{\phi_{j-1}^2}\right) - \left(\frac{1}{\phi_{j+1}^2} - \frac{1}{\phi_j^2}\right) \tag{50}$$

where

$$\widetilde{\theta}_j = \frac{(\phi_j+1)^2 - \phi_j^2}{\phi_j^2(\phi_j+1)^2} - \frac{\phi_j^2 - (\phi_j-1)^2}{\phi_j^2(\phi_j-1)^2} \tag{51}$$

$$= \frac{\widetilde{w}_j}{\phi_j^2(\phi_j+1)^2(\phi_j-1)^2} \tag{52}$$

Since $\phi_j^2(\phi_j+1)^2(\phi_j-1)^2 > 0$, then it is sufficient to show $\widetilde{w}_j < 0$ where

$$\widetilde{w}_j = (\phi_j-1)^2\left((\phi_j+1)^2 - \phi_j^2\right) - (\phi_j+1)^2\left(\phi_j^2 - (\phi_j-1)^2\right) \tag{53}$$

After some algebraic manipulation, it can be shown that

$$\widetilde{w}_j = -6\phi_j^2 + 2 \tag{54}$$

Solving (54), we have $\widetilde{w}_j < 0,\ \forall\, \phi_j > \frac{1}{3}(3)^{\frac{1}{2}}$. Hence, it follows that $\widetilde{\theta}_j < 0$. Finally, we consider the inequality $\widehat{\theta}_j > \widetilde{\theta}_j$. Defining $\vartheta_j = \widehat{\theta}_j - \widetilde{\theta}_j$ we have

$$\vartheta_j = \frac{-2}{\phi_j(\phi_j+1)(\phi_j-1)} - \frac{-6\phi_j^2+2}{\phi_j^2(\phi_j+1)^2(\phi_j-1)^2} \tag{55}$$

which, when solving for $\vartheta_j = 0$, reduces to

$$\phi_j^3 - 3\phi_j^2 - \phi_j + 2 = 0 \tag{56}$$

Hence, solving (56) leads to $\vartheta_j > 0, \forall\, \phi_j > 0.461$. Therefore $\widehat{\theta}_j > \widetilde{\theta}_j$ and hence it follows that $\widehat{\theta}_j > \theta_j$. Since $\widehat{\theta}_j < 0$ then $\theta_j < 0$ and finally that $d_j < d_{j+1}$ and

$$\frac{p(r-j) - p(r-j+1)}{\gamma'} < \frac{p(r-(j+1)) - p(r-j)}{\gamma'} \tag{57}$$

By induction, we have

$$\varphi_0(x,r) = p\,(M-1) - p(M) \tag{58}$$

which provides a solution for (34) and specifies the closest ranked probabilities which we now use in the following to determine an expression for N_0. Hence, using empirical probabilities we can solve (18) as

$$\Delta_0 = \widehat{p}\,(M-1) - \widehat{p}(M) \tag{59}$$

Now, since

$$\sup\left\{\widehat{p}(r) + \frac{\Delta_0}{2}\right\} = \inf\left\{\widehat{p}(r-1) - \frac{\Delta_0}{2}\right\}\Bigg|_{r=M} \tag{60}$$

then we have

$$\epsilon_r \geq \frac{\Delta_0}{4} \tag{61}$$

Hence, the value of n_0 can be calculated using the DKW inequality and the result in (16) as follows.

$$n_0 = \frac{1}{2\epsilon_r^2}\ln\left(\frac{2}{1-\zeta'}\right) \tag{62}$$

$$\leq \frac{1}{2\left(\frac{\Delta_0}{4}\right)^2}\ln\left(\frac{2}{1-\zeta'}\right) \tag{63}$$

$$\leq \frac{8}{\Delta_0^2}\ln\left(\frac{2}{1-\zeta'}\right) \tag{64}$$

Hence, from (5), the minimum number of observations required to estimate the entropy is

$$N_0 = \frac{n_0}{\widehat{p}_0(r)} \tag{65}$$

where,

$$\widehat{p}_o(r) = \widehat{p}(M) \tag{66}$$

$$= \frac{M^{\alpha-1}}{[(M-1)(M+\beta)]^{\alpha}} \tag{67}$$

Now, this gives the number of observations required to estimate the entropy within a specified degree of confidence and within specified bounds by considering the smallest probabilities used. This formula is suitable for estimating the minimum number of samples required to compute the N-gram entropy of a sequence.

D. *Remarks on Optimality*

We consider the issue of optimality of the proposed estimator based on the preceeding mathematical derivation. The question considered in this paper is how many samples are required to estimate entropy according to the specified assumptions, particularly that the sampled data is stationary, iid and generated according to a Zipf-Mandelbrot-Li law. In practice, real observed data may not be so constrained, but nevertheless these basic assumptions admit a wide range of possible systems.

Now, we consider that the estimation is in reference to probabilities assumed to be generated by a Zipf-Mandelbrot-Li system and hence optimality can only be considered in this strict sense. We may define an estimate of the number of samples for entropy as $M(\zeta')$-optimal when the estimator specifies the number of samples required to compute the entropy using a conventional plug-in method, with probabilities generated according to a Zipf-Mandelbrot-Li law with alphabet size M, such that the value of the entropy computed distinguishes between the two events with nearest probabilities, that is, $\widehat{p}(M-1)$ and $\widehat{p}(M)$ where

$$\epsilon_r = \widehat{p}\,(M-1) - \widehat{p}(M) \tag{68}$$

and the difference is estimated with a confidence limit of ζ'.

Note that this does not guarantee the entropy estimate itself will be unbiased, since this is a function of the method used to estimate the probabilities, not the estimation of the number of samples considered here. Furthermore, it must be understood that this is an optimality associated with determining the number of samples required to estimate entropy, not in the estimation of the entropy itself. Hence, given that we estimate the entropy within a confidence limit of ζ' and the specified assumptions, including event probabilities ranked according to a particular Zipfian law as described, then the number of samples are specified optimally.

E. Samples required for coarse entropy classification

Suppose we wish to detect major differences between entropies H_i, H_{i+1} due to changes in the *most frequent* symbol probabilities. What then is the minimum number of samples N_0 required? We consider this case in relation to applications of online analytic entropy calculation when the aim is not necessarily to estimate the entropy accurately, but rather where we seek to detect relative changes in the entropy characteristics of observed data due to changes in the most frequently occurring events. In this case, to find N_0 for $\sup(\epsilon_0)$, implies detecting changes due to $p(r), p(r+1), \ldots$ where $r \ll M$, e.g. $r = 1, 2$. Hence define a new alphabet size $M_c < M$, then since $p(i) > p(i+1), \forall\, i$ according to the Zipf-Mandelbrot-Li law, then

$$\sum_{i=1}^{M_c} \widehat{p}(i) \approx \sum_{i=1}^{M} \widehat{p}(i) \tag{69}$$

we may omit $\{\widehat{p}(i)\}\ \forall\, i > M_c$ (i.e. the most infrequently occurring symbols), and thus we select

$$\Delta_0 = \widehat{p}\,(M_c - 1) - \widehat{p}(M_c) \tag{70}$$

where the final number of samples may be approximated directly from (64) and (65) substituting M_c for M . Similarly, this admits other related approaches to varying Δ_0 for example, the top $q\%$ of the probabilities, e.g.

$$r = \left\lceil \frac{M}{4} \right\rceil \tag{71}$$

An example of using this approach is given in the next section.

III. Example results

1) First order entropy of English text : Consider the following example of calculating the number of samples required to determine the entropy[8]. Let the alphabet size be $M = 26$, and confidence level of the calculation be 95%, i.e. $\zeta' = 0.95$. Hence, using (64)

$$n = \frac{8}{\Delta_0^2} \ln\left(\frac{2}{1 - 0.95}\right) \tag{72}$$

where the probabilities are computed from (32) as

$$\widehat{p}(26) = \frac{\gamma'}{(26 + \beta)^{\alpha}} \tag{73}$$

with the parameters computed as:

$$\alpha = \frac{\log_2(26+1)}{\log_2(26)} \approx 1.012, \quad \beta = \frac{26}{26+1} \approx 0.963, \tag{74}$$

$$\gamma' = \frac{\kappa 26^{\alpha-1}}{(26-1)^{\alpha}} \approx 0.351 \tag{75}$$

Hence, it can be found that $\Delta_0 = 4.88 \times 10^{-4}$, and therefore $n = 1.24 \times 10^{8}$. Now, this implies the total number of observations required is then

$$\widehat{N}_0 = \frac{1.24 \times 10^{8}}{\widehat{p}(26)} = 1.06 \times 10^{9} \tag{76}$$

Thereby, this gives an estimate of the number of observation samples which may be required in order to obtain an estimate of the entropy which takes into account the smallest contributing probabilities, i.e. the most infrequent symbols. The value of approximately 1 billion samples (ie the number of letters in about 2000 novels, each 100,000 words in length) appears to be consistent with reports in the literature, and provides a useful indication of the upper bound required to compute entropy in this case.

2) Coarse entropy classification for small alphabet size : Consider the following example of calculating the number of samples required to detect the difference between entropies by finding the least number of samples required to detect *major* changes in entropy due to changes in the *most frequent* symbol probabilities. Let the alphabet size of interest be $M = 4$, and confidence level[9] of the calculation be 99%, i.e. $\zeta' = 0.99$. Hence,

$$n = \frac{8}{\Delta_0^2} \ln\left(\frac{2}{1 - 0.99}\right) \tag{77}$$

where the probabilities are computed from (28) as before, except that we now consider the minimal number of samples to detect some of the most frequently observed symbols, hence we choose $\Delta_0 = \widehat{p}\,(1) - \widehat{p}(2)$, where

$$\widehat{p}(2) = \frac{\gamma'}{(2 + \beta)^{\alpha}} \tag{78}$$

with the parameters computed as:

$$\alpha = \frac{\log_2(3+1)}{\log_2(3)} \approx 1.262, \quad \beta = \frac{3}{3+1} \approx 0.75, \tag{79}$$

$$\gamma' = \frac{\kappa^{\alpha-1}}{(3-1)^{\alpha}} \approx 1.04 \tag{80}$$

It can be found that $\Delta_0 = 0.172$ and therefore $n = 1440$. Now, this implies the total number of observations required is then

$$\widehat{N}_0 = \frac{1440}{\widehat{p}(2)} = 5626 \tag{81}$$

Hence, this gives an estimate of the number of samples required in order to distinguish two different natural sequences using the most frequently occurring symbols.

The efficacy of the method can be readily observed by

[8] It has been debated in the literature as to how closely language symbols follow Zipf's law and in addition, noted that such measurements may be subject to observational bias [28],[30]. Zipf's law has been applied to various language features, for example letters and words within human languages but can also be considered in terms of other data sets with similar forms of probabilistic structure.

[9] While normally a high confidence level would be used, eg $\zeta' = 0.95$ or $\zeta' = 0.99$, if the aim is to efficiently detect relative changes in entropy for different datasets, then a lower confidence level may be considered which would indicate the fewest samples which may be required.

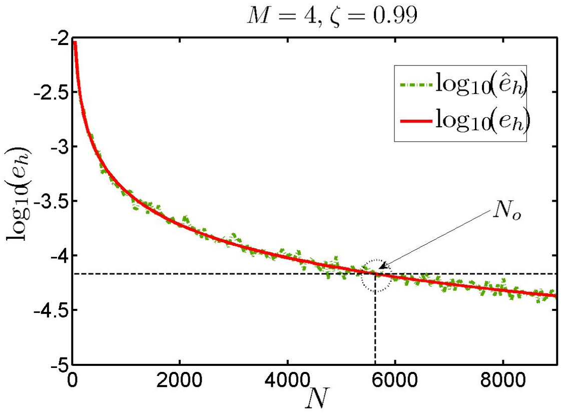

Fig. 1. Performance of the estimation method is shown here using the mse between the estimated and true entropy as a function of samples N shown on a log scale. The estimated minimal number of samples required based on coarse estimation for M=4 and a lower confidence of $\zeta' = 0.99$, is found at $N_0 = 5626$ with corresponding mse.

simulation in Fig. 1. In this case, a small alphabet size of $M = 4$ was selected which admits a set of probabilities following a Zipfian law. Hence, this set of probabilities was used to generate a precise entropy value according to (1). A set of random data was generated following the notion of an M-sided die, where we have a random variable X of a sequence $X = X_1, \ldots, X_i, \ldots$, where $X_i = x \in \mathbf{X}^M$, that is, x_i may take on one of M distinct values, $\mathbf{X}^M$ is a set from which the members of the sequence are drawn, and hence x_i is in this sense symbolic, and where each value occurs independently with empirical probability $\widehat{p}(x_i)$, $i \in [1, M]$, and in this case is estimated according to a usual histogram estimator defined in (5). Various improved estimators are possible (see for example ([13],[18])), however for our purposes we are interested to see how the classic plug-in histogram estimate for entropy behaves as a function of the number of symbol samples available and then comparing this to the theoretical result of our proposed formula. Thus, we have an actual entropy value $H_a(M)$ calculated from a set of prescribed probabilities $\{p(x_i)\}$, $i \in [1, M]$ and an empirically estimated entropy calculated as

$$H_e(M, N) = -\sum_{i=1}^{M} \widehat{p}(x_i) \log_2 (\widehat{p}(x_i)) \tag{82}$$

where N is the number of samples used to generate each empirical entropy, and the mean square error $\widehat{e}_h$ between the actual and empirical entropy is therefore defined by the ensemble mean across N_b runs as

$$\widehat{e}_h = \frac{1}{N_b} \sum_{j=1}^{N_b} (H_a(M) - H_e(M, N; j))^2 \tag{83}$$

Now, the mean square error $\widehat{e}_h$ can be approximated by a smooth function, and hence we can define (for example),

$$e_h = f_e(\widehat{e}_h) \tag{84}$$

$$f_e(x) = \theta_1 x^{\theta_2} + \theta_3 \tag{85}$$

where the parameters $\theta = [\theta_1, \theta_2, \theta_3]$ are determined by a curve fitting procedure for each N. Hence, using this simulation approach, it is possible to readily see the effectiveness of the method for determining the number of samples required to estimate entropy according to some given confidence criteria (Fig. 1).

IV. Conclusions

Shannon entropy is a well known method of measuring the information content in a sequence of probabilistic symbolic events. Entropy is calculated based on a set of distinct symbol probabilities. A problem encountered when computing entropy using conventional plug-in histogram methods is that it generally involves a large, but unknown number of samples which depends on the number of distinct symbols available.

The method presented here is quite different from the scope of other papers which develop various entropy estimation algorithms. Whereas most effort has been on developing entropy estimators with improved results in terms of bias or efficiency, here we are not proposing a new entropy estimation algorithm. Our work considers the more fundamental question, which to date and to the best of our knowledge, has not been previously addressed, that is, just how many data samples are required to be observed in order to estimate Shannon entropy?

This is a critical question, since while various methods have been proposed to reduce the number of samples which may be required to estimate entropy, it has not been clear just how many samples are actually required in a conventional model. It has been generally accepted in the literature that it is a large number, but it has been unclear just how large.

In this paper, we have proposed a method of estimating the number of samples required to estimate the Shannon Entropy for natural sequences. Accordingly, we propose a method based on a modified Zipf-Mandelbrot-Li law and the Dvoretzky-Kiefer-Wolfowitz inequality to determine the number of samples which is sufficient to probabilistically discriminate the individual symbol probabilities according to specified confidence limits. Examples have been given which show the efficacy of the proposed methodology.

The applicability of the proposed method is broad, since it is applicable to a wide range of natural sequences which have probabilistic event structure which may be described by a Zipfian law, in particular, we have chosen a Zipf-Mandelbrot-Li law. Hence the proposed method may have implications for fields of study which includes various naturally occurring phenomena such as linguistics and human languages, as well as ecological, biological, econometric and sociological systems.

Further improvements may possible by re-considering some of the assumptions, such as iid samples for the parametrization of the Zipf-Mandelbrot-Li model. In addition, since all entropy estimation algorithms must inherently rely on some method of estimating the individual probabilities, an interesting generalization of this method would be to seek to apply it to more recently introduced model-based entropy estimators. Finally, it would be of interest to apply this method to various real world applications to compare the theoretical results against experimentally obtained results.

Acknowledgement

The authors gratefully acknowledge funding support from the ARC Centre of Excellence for the Dynamics of Language and the helpful comments by reviewers on this manuscript.

References

[1] C. E. Shannon, “A mathematical theory of communication (parts I and II),” *Bell System Technical Journal*, vol. XXVII, pp. 379–423, 1948.

[2] ——, “A mathematical theory of communication (part III),” *Bell System Technical Journal*, vol. XXVII, pp. 623–656, 1948.

[3] ——, “Prediction and entropy of printed English,” *Bell System Technical Journal*, pp. 50–64, 1951.

[4] G. Barnard, “Statistical calculation of word entropies for four western languages,” *IRE Transactions Inf. Theory*, pp. 49–53, March 1955.

[5] J. Herrera and P. Pury, “Statistical keyword detection in literary corpora,” *Eur. Phys. J. B*, vol. 63, pp. 135–146, 05 2008.

[6] B. Allen, M. Kon, and Y. Bar-Yam, “A new phylogenetic diversity measure generalizing the shannon index and its application to phyllostomid bats,” *American Naturalist*, vol. 174, no. 2, pp. 236–243, 2009.

[7] C. Rao, “Diversity and dissimilarity coefficients: a unified approach,” *Theoretical Population Biology*, vol. 21, pp. 24–43, 1982.

[8] R. Lee, P. Jonathan, and P. Ziman, “Pictish symbols revealed as a written language through application of shannon entropy,” *Proc. R. Soc. A*, vol. 466, pp. 2545–2560, 2010.

[9] R. A. Khan, A. Meyer, H. Konik, and S. Bouakaz, “Facial Expression Recognition using Entropy and Brightness Features ,” in *11th Int. Conf. on Intell. Sys. Design and Applic.(ISDA)*, IEEE, Ed., Dec 2011.

[10] S. Fuhrman, M. J. Cunningham, X. Wen, G. Zweiger, J. J. Seilhamer, and R. Somogyi, “The application of Shannon entropy in the identification of putative drug targets,” *Biosystems (A6E)*, vol. 55, no. 1–3, pp. 5–14, 2000.

[11] W. Ebeling and T. Pöschel, “Entropy and long-range correlations in literary English,” *Europhysics Letters*, vol. 26, no. 4, p. 241, 1994.

[12] I. Moreno-Sánchez, F. Font-Clos, and Á. Corral, “Large-scale analysis of Zipfs law in English texts,” *PLOS ONE*, vol. 11, no. 1, 01 2016.

[13] T. Schürmann and P. Grassberger, “Entropy estimation of symbol sequences,” *Chaos*, vol. 6(3), pp. 414–427, 1996.

[14] P. Grassberger, “Finite sample corrections to entropy and dimension estimates,” *Physics Letters A*, vol. 128, no. 67, pp. 369–373, 1988.

[15] J. A. Bonachela, H. Hinrichsen, and M. A. Muñoz, “Entropy estimates of small data sets,” *Journal of Physics A: Mathematical and Theoretical*, vol. 41, no. 202001, pp. 1–9, 2008.

[16] J. Montalvão, D. Silva, and R. Attux, “Simple entropy estimator for small datasets,” *Electronics Letters*, vol. 48, pp. 1059–1061, Aug 16 2012.

[17] M. Paavola, “An efficient entropy estimation approach,” Ph.D. dissertation, University of Oulu, 2011.

[18] A. Lesne, J.-L. Blanc, and L. Pezard, “Entropy estimation of very short symbolic sequences,” *Phys. Rev. E*, vol. 79, p. 046208, Apr 2009.

[19] A. O. Schmitt and H. Herzel, “Estimating the entropy of DNA sequences,” *Journal of Theoretical Biology*, vol. 188, no. 3, pp. 369 – 377, 1997.

[20] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, “Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator,” *Ann. Math. Statist.*, vol. 27, pp. 642–669, 1956.

[21] J. L. Doob, “Heuristic approach to the Kolmogorov-Smirnov theorems,” *Ann. Math. Statist.*, vol. 20, no. 3, pp. 393–403, 09 1949.

[22] P. Massart, “The tight constant in the Dvoretzky-Kiefer-Wolfowitz Inequality,” *The Annals of Probability*, vol. 18, no. 3, pp. 1269–1283, 07 1990.

[23] R. Zielinski, “Kernel estimators and the Dvoretzky-Kiefer-Wolfowitz Inequality,” *Applicationes Mathematicae*, vol. 34, pp. 401–404, 01 2007.

[24] E. Learned-Miller and J. DeStefano, “A probabilistic upper bound on Differential Entropy,” *IEEE Trans. on Information Theory*, vol. 54, pp. 5223 – 5230, 12 2008.

[25] G. Zipf, *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: Houghton Mifflin, 1935.

[26] S. T. Piantadosi, “Zipf’s word frequency law in natural language: A critical review and future directions,” *Psychonomic Bulletin & Review*, vol. 21, no. 5, pp. 1112–1130, 2014.

[27] W. Li, “Random texts exhibit Zipf’s-law-like word frequency distribution,” *IEEE Transactions on Information Theory*, vol. 38, no. 6, pp. 1842–1845, 1992.

[28] ——, “Zipf’s law everywhere,” *Glottometrics*, vol. 5, pp. 14–21, 2002.

[29] M. A. Montemurro, “Beyond the Zipf-Mandelbrot law in quantitative linguistics,” *Physica A*, vol. 300, pp. 567–578, Nov 2001.

[30] Á. Corral, G. Boleda, and R. Ferrer-i Cancho, “Zipfs law for word frequencies: Word forms versus lemmas in long texts,” *PloS one*, vol. 10, no. 7, p. e0129031, 2015.

[31] R. Ferrer i Cancho and R. V. Solé, “The small-world of human language,” *Proceedings of the Royal Society of London B*, vol. 268, no. 1482, pp. 2261–2265, November 2001.

[32] Y.-S. Chen and F. Leimkuhler, “A relationship between Lotka’s law, Bradford’s law, and Zipf’s law,” *Journal of the American Society for Information Science.*, vol. 37, no. 5, pp. 307–314, 1986.

[33] ——, “Booth’s law of word frequency,” *Journal of the American Society for Information Science.*, vol. 41, no. 5, pp. 387–388, 1990.

[34] A. D. Booth, “A law of occurrences for words of low frequency,” *Information and Control*, vol. 10, no. 4, pp. 386–393, 1967.

[35] B. Mandelbrot, *The fractal geometry of nature*. New York: W. H. Freeman, 1983.