When do finite sample effects significantly affect entropy estimates?

T. Dudok de Wit[a]

Centre de Physique Théorique[b], Luminy, case 907, 13288 Marseille Cedex 9, France

Abstract. An expression is proposed for determining the error made by neglecting finite sample effects in entropy estimates. It is based on the Ansatz that the ranked distribution of probabilities tends to follow a Zipf scaling.

PACS. 02.50.-r Probability theory, stochastic processes, and statistics

1 Introduction

The growing interest in complexity measures and symbolic dynamics [1,2] has brought to the forefront various problems related to the estimation of entropic quantities from finite sequences [3]. Such estimates are known to suffer from a bias, which prevents quantities such as the metric entropy from being meaningfully estimated. The purpose of this letter is to provide an analytical expression for this bias, in order to test for finite sample effects in entropy estimates.

Consider the general case of a string of N symbols $\{i_1 i_2 \cdots i_N\}$, each of which belongs to a finite alphabet $\mathcal{A}$. The average informational content of substrings of length d taken from this sequence is expressed by the Shannon entropy [4]

$$H_d = - \sum_{i_1,\ldots,i_d \in \mathcal{A}} \mu\left([i_1 i_2 \cdots i_d]\right) \log \mu\left([i_1 i_2 \cdots i_d]\right) \ , \quad (1)$$

where μ is the natural invariant measure with respect to the shift. Of particular interest is the block or dynamical Shannon entropy $h_d = H_{d+1} - H_d$ from which one gets the measure-theoretic entropy of the system

$$h(\mu) = \lim_{d \to \infty} h_d \ , \quad (2)$$

a quantity that is intimately related to the Kolmogorov-Sinaï entropy in case the string represents the output of a shift dynamical system.

The main problem lies in the estimation of the empirical measure μ from a finite string of symbols. Direct box counting yields

$$\mu\left([i_1 i_2 \cdots i_d]\right) \approx \frac{\#[i_1 i_2 \cdots i_d]}{N - d + 1} \ , \quad (3)$$

where $\#[i_1 i_2 \cdots i_d]$ is the occurrence frequency of the block $i_1 i_2 \cdots i_d$ in the string. It is well known that statistical fluctuations in the sample on average lead to a systematic underestimation of the entropy. This problem becomes particularly acute as the word size increases for a given string length N. Since this deviation can easily be mistaken for the signature of a finite memory process, it is of prime importance to determine whether its origin is physical or not.

Several authors have already addressed the problem of making corrections to empirical entropy estimates [3,5–7]; their expressions are valid as long as the occurrence frequencies of the observed words are large compared to one. While this may hold for relatively short words, it breaks down for long ones, making it difficult for a small correction to be used as a safe indication for a small deviation. Our objective is to derive a more reliable (although less accurate) expression of the deviation, to be used as a warning signal against the onset of finite sample effects.

As a first guess one could require the sample to be long enough for each word to have a chance to appear. This gives $N \gg N_{\text{symb}}^d$, where N_{symb} is the cardinality of the alphabet. This criterion, however, is generally found to be too conservative because it does not take into account the grammar, *i.e.* the rules that cause some words to be forbidden or less frequent than others.

2 The Zipf-ordered distribution

To derive our expression, we first rank the words according to their frequency of occurrence: let $n_{k=1}$ denote the frequency of occurrence of the most probable word, $n_{k=2}$ of the next most probable one etc. Multiple instances of the same frequency get consecutive ranks. This monotonically decreasing distribution is called Zipf-ordered.

The Asymptotic Equipartition Property introduced by Shannon [4] states that the ensemble of words of length

[a] e-mail: `ddwit@cpt.univ-mrs.fr`
[b] Université de Provence and CNRS

d can be divided into two subsets. The first one consists of "typical words" that occur frequently and roughly have the same probability of occurrence. The other subset is made of "rare words" that belong to the tail of the distribution. According to the Shannon-Breiman-MacMillan theorem, the entropy is related to the typical words in the limit where $N \to \infty$; the contribution of rare words progressively disappears as N increases. In some sense this observation justifies the procedure to be described below.

It was noted by Pareto [8], Zipf [9] and others, and later interpreted by Mandelbrot [10] that the tail of the Zipf-ordered distribution n_k tends to follow a universal scaling law

$$n_k = \alpha k^{-\gamma}, \quad \gamma > 0, \tag{4}$$

which is found with astonishing reproducibility in economics, social sciences, physics etc. [10]. As shown in [11,12], many different systems give rise to Zipf laws, whose ubiquity is thought to be essentially a consequence of the ranking procedure.

The physical meaning of Zipf's law is still an unsettled question, although it does not seem to reflect any particular self-organization (see for example [13,14]). We just mention that a slow decay is an indication for a "rich vocabulary", in the sense that rare words occur relatively often.

The key point is that the empirical Zipf-ordered distribution has a cutoff at some finite value $k = N_{\max}$ because of the finite length of the symbol string. For the same reason, the occurrence frequencies are necessarily quantized. Our main hypothesis is that the true distribution extends beyond $N_{\max}$, up to the lexicon size $K \geq N_{\max}$, following Zipf's law with the same exponent γ. This Ansatz has already been suggested as a way to estimate entropies from long words [15].

3 Estimating the bias

Let $\hat{H}$ be the Shannon entropy computed from the empirical distribution (using Eqs. (1, 3)) and H the entropy one would obtain from a non truncated distribution, in which the frequencies are not quantized anymore and extend beyond $N_{\max}$ following Zipf's law

$$\hat{H} = -\sum_{k=1}^{N_{\max}} \frac{n_k}{\sum_{k=1}^{N_{\max}} n_k} \log \frac{n_k}{\sum_{k=1}^{N_{\max}} n_k}, \tag{5}$$

$$H = -\sum_{k=1}^{K} \frac{n_k}{\sum_{k=1}^{K} n_k} \log \frac{n_k}{\sum_{k=1}^{K} n_k}.$$

The truncation has two counteracting effects. It changes the renormalization of the occurrence frequencies and causes some of the least frequent words to be omitted.

The difference δ between the two entropy estimates

$$\delta = H - \hat{H} \tag{6}$$

is what we call the bias, to be used as a measure of the deviation resulting from finite sample effects. We shall assume that $N_{\max} \gg 1$, which is equivalent to saying that

the distribution must have a sufficiently long tail for a power law to make sense.

It is natural to define a small parameter $0 \leq \varepsilon \ll 1$, which goes to zero for a non truncated distribution

$$\varepsilon = \frac{1}{N} \sum_{k=N_{\max}+1}^{K} n_k. \tag{7}$$

Remember that $N = \sum_{k=1}^{K} n_k$ [16].

Now, assuming that Zipf's law persists for $k > N_{\max}$, we have

$$\varepsilon = \frac{1}{N} \sum_{k=N_{\max}+1}^{K} \alpha k^{-\gamma} = \frac{\alpha}{N} \left(\zeta(\gamma, N_{\max}+1) - \zeta(\gamma, K+1) \right), \tag{8}$$

where $\zeta(\gamma, m)$ is the Hurwitz or generalized Riemann zeta function. For $k > \gamma$, the following approximation holds [17]

$$\zeta(\gamma, m) = \frac{m^{1-\gamma}}{\gamma - 1} - \frac{m^{-\gamma}}{2} + \frac{m^{-\gamma-1}}{12}. \tag{9}$$

Since $K, N_{\max} \gg 1$, we may write

$$\varepsilon = \frac{\alpha}{N(\gamma - 1)} \left(N_{\max}^{1-\gamma} - K^{1-\gamma} \right). \tag{10}$$

The value of α remains to be determined. To do so, we note that the least frequent words in the Zipf-ordered distribution occur once or a few times only. One may therefore reasonably set $n_{k=N_{\max}} \approx 1$, giving $\alpha \approx N_{\max}^{\gamma}$.

The bias δ can now be expanded in powers of ε. Keeping terms of order $\mathcal{O}(\varepsilon)$ only, we have

$$\delta = -\varepsilon\hat{H} + (1 + \varepsilon) \left(\varepsilon - \sum_{k=N_{\max}+1}^{K} \frac{n_k}{N} \log \frac{n_k}{N} \right). \tag{11}$$

For the conditions stated before, the sum can be approximated by

$$-\sum_{k=N_{\max}+1}^{K} \frac{n_k}{N} \log \frac{n_k}{N} = -\varepsilon \left(\log N - \gamma \log \frac{N_{\max}}{K} \right), \tag{12}$$

finally giving the result of interest

$$\delta = \varepsilon \left(1 + \log N - \hat{H} - \gamma \log \frac{N_{\max}}{K} \right) \tag{13}$$

$$\varepsilon \approx \frac{N_{\max}}{N(\gamma - 1)} \left(1 - \left(\frac{N_{\max}}{K} \right)^{\gamma-1} \right).$$

Notice that the true entropy is always underestimated; furthermore ε is continuous at $\gamma = 1$ [18]. Most of the variation comes from the small parameter ε, whose expression reveals two different effects:

1) the ratio $N_{\max}/N$ reflects the uncertainty of the frequency estimates;

2) the scaling index γ, whose value is usually between 0.5 to 1.5, is indicative of the lacunarity of the word distribution. In the case of a shift dynamical system, γ reveals how unevenly the rare orbits fill the phase space.

For the sake of comparison, the first order approximation for finite sample effects derived in [5,6] is

$$\delta = \frac{N_{\max}}{2N}. \tag{14}$$

We conclude from equation (13) that the bias is not just related to statistical fluctuations in the empirical occurrence frequency, but is also caused by the omission of words that are asymptotically rare. If the true distribution of the ranked words were exponential or ultimately ended with an exponential tail, then our criterion would be too conservative but still reliable as such.

The following procedure is proposed for detecting the maximum word length for which entropies can be meaningfully estimated: compute Zipf-ordered distributions for increasing word-lengths d. For each length, estimate the bias δ by least-squares fitting a power law to the tail of the observed distribution. As soon as this bias exceeds a given threshold (say 10% of $\hat{H}$), then entropies computed from longer words are likely to be significantly corrupted by finite sample effects.

Equation (13) supposes that the maximum lexicon size K is known *a priori*, which is seldom the case. This is not a serious handicap, however, since the value of K has relatively little impact on the bias; a rough approximation such as $K = N_{\mathrm{symb}}^d$ may do well.

4 Two examples

To briefly illustrate the results, we now consider two examples. The first one is based on a Bernouilli process, whose entropy and Zipf-ordered distribution can be calculated analytically. The string of symbols is drawn from a two letter alphabet, one with probability λ and the other with probability $1 - \lambda$. The block entropy of this process is independent of the word length and equals

$$h = -\lambda \log \lambda - (1 - \lambda) \log(1 - \lambda) . \tag{15}$$

Figure 1 compares the true block entropy with estimates drawn from a sample of length $N = 2000$ with $\lambda = 0.15$. The departure of the empirical estimate from the true one is evident. Without knowledge of the true entropy, however, it is very difficult to tell whether the decrease of the entropy is an artifact or just the signature of a short-time memory.

The second panel displays the true and the empirical Zipf-ordered distributions as obtained for words of length $d = 9$. Zipf's law clearly holds for words whose rank exceeds about 30. After this, the scaling exponent γ is estimated, see the third panel. The decrease of this exponent with the word length d suggests that the contribution of the rare words becomes increasingly important. Finally, the bias δ, which is shown in the fourth panel, suggests
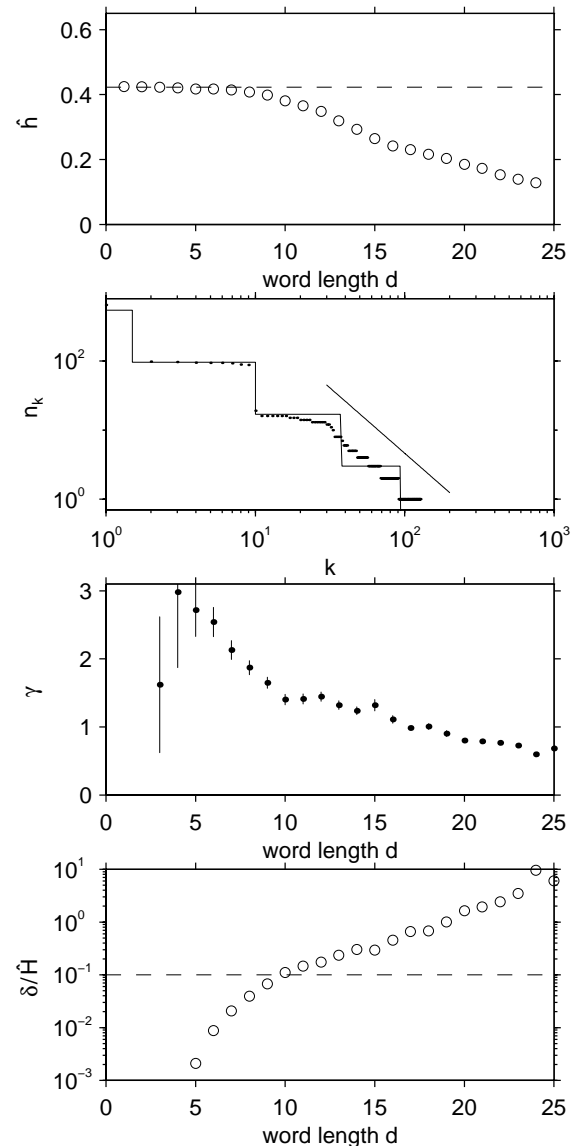

Fig. 1. Analysis of a Bernouilli sequence, with $N = 2000$ and $\lambda = 0.15$. From top to bottom: (1) the empirical block entropy and the true one (dashed), (2) the true (line) and observed (dots) Zipf-ordered distributions for words of length $d = 8$; (3) the scaling exponent γ obtained by fitting the tail of the Zipf-ordered distribution (error bars represent ± 1 standard deviation resulting from the least-squares fit), (4) the bias δ. In this case, entropies cannot be reliably estimated for word lengths beyond $d = 9$. Block entropies are normalized to $\log N_{\mathrm{symb}}$, so that the maximum possible value is 1.

that the onset of a significant bias occurs around $d = 8$; this value is indeed in agreement with the results of the first panel.

The validity of the bias estimate was tested on various examples and was found to be reliable, provided that $N_{\max} \gg 1$.

In the second example, we consider a sequence of $N = 10^4$ symbols generated by the logistic map $x_{i+1} = \lambda x_i (1 - x_i)$ in a chaotic regime with $\lambda = 3.8$.
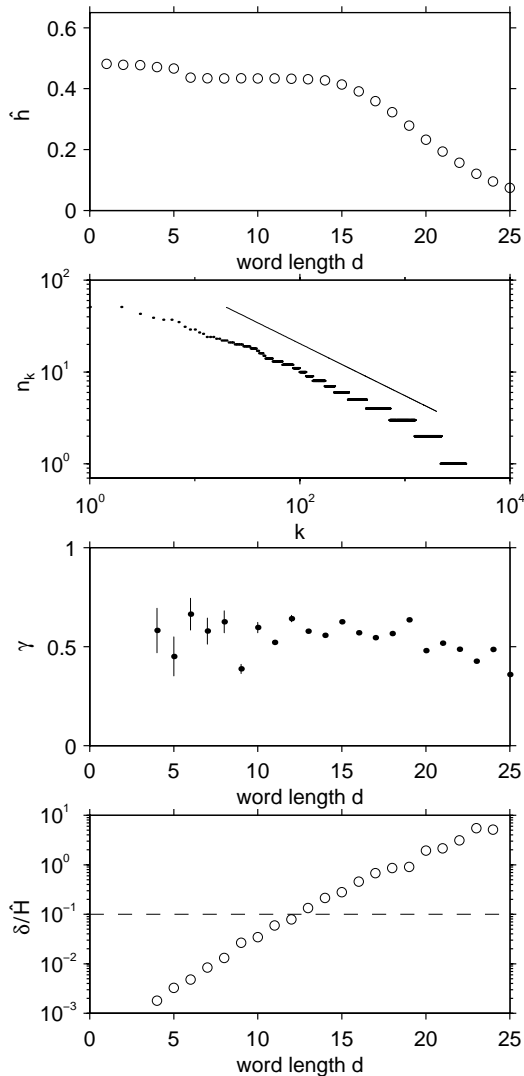
Fig. 2. Analysis of a logistic map sequence, with the same legend as the previous figure; the string length is $N = 10^4$. The second panel shows a Zipf-ordered distribution for $d = 18$. The largest word size for which the relative bias is smaller than 10%, is $d = 12$.

The (generating) partition $\mathcal{P} = \{[0, 0.5[, [0.5, 1]\}$ gives us a two-letter alphabet.

Figure 2 again shows that the block entropy decreases above a certain word length. In contrast to the previous example, the measured scaling exponent γ is small and almost constant, regardless of the word length. We believe this to be a consequence of the intricate structure of the self-similar attractor. This low value of γ already suggests that rare words should bring a significant contribution to the entropy. The bias δ finally suggests stopping at $d = 12$.

5 Conclusion

Summarizing, we have derived a simple expression (Eq. (13)) for detecting the onset of finite sample size effects in entropy estimates. It is based on the empirical evidence that rank-ordered distribution of words tend to follow Zipf's law. The criterion reveals that rare events can significantly bias the empirical entropy estimate.

References

1. C. Beck, F. Schlögl, *Thermodynamics of chaotic systems* (Cambridge University Press, Cambridge, 1993).
2. R. Badii, A. Politi, *Complexity: hierarchical structures and scaling in physics* (Cambridge University Press, Cambridge, 1997).
3. T. Schürmann, P. Grassberger, Chaos **6**, 414 (1996).
4. R.E. Blahut, *Principles and practice of information theory* (Addison Wesley, Reading, MS, 1987).
5. H. Herzel, Sys. Anal. Mod. Sim. **5**, 435 (1988).
6. P. Grassberger, Phys. Lett. A **128**, 369 (1988).
7. A.O. Schmitt, H. Herzel, W. Ebeling, Europhys. Lett. **23**, 303 (1993).
8. V. Pareto, *Cours d'économie politique* (Rouge, Lausanne, 1897).
9. G. Zipf, *Human behavior and the principle of least effort* (Addison-Wesley, Cambridge MA, 1949).
10. B. Mandelbrot, *Fractals and scaling in finance: discontinuity, concentration, risk* (Springer, New York, 1997); B. Mandelbrot, *Fractales, hasard et finance* (Flammarion, Paris, 1997).
11. R. Günther, L. Levitin, B. Schapiro, P. Wagner, Int. J. Theor. Phys. **35**, 395 (1996).
12. G. Troll, P. beim Graben, Phys. Rev. E **57**, 1347 (1998).
13. G.A. Miller, E.B. Newman, Am. J. Psychology **71**, 209 (1958).
14. W. Li, Complexity **3**, 10 (1998).
15. T. Pöschel, W. Ebeling, H. Rosé, J. Stat. Phys. **80**, 1443 (1995).
16. To be exact $N - d + 1 = \sum_{k=1}^{K} n_k$, but we use the fact that $d \ll N$.
17. J. Spanier, K.B. Oldham, *An atlas of functions* (Springer, Berlin, 1987), formula (64:9:1).
18. The term $(K/N_{\max})^{1-\gamma}$ can be large when $K \gg N_{\max}$ and $\gamma < 1$ but this divergence becomes effective long after the maximum word size has been exceeded; it is therefore not a matter of concern here.