I

\[ P_H = \sum_{\eta} \frac{1}{\eta} \eta P \]

The natural invariance measure with respect to the shift of partition

\[ P_H = \int \frac{1}{\eta} \eta P \]
a quantity that is intimately related to the Kolmogorov-Sinai entropy in case
the string represents the output of a shift dynamical system.

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

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

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, 6, 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 \( N_{\text{symb}} \) 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.

that the ensemble of words of length \( d \) 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.

2
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

$$n_k \sim k^{-\gamma}, \quad \gamma > 0,$$

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 $k = N_{\text{max}}$ 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 $N_{\text{max}}$, up to the lexicon size $K > N_{\text{max}}$, following Zipf’s law with the same exponent $\gamma$. This Ansatz has already been suggested as a way to estimate entropies from long words [15].

3. Estimating the bias

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

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

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

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 $\delta$ between the two entropy estimates

$$\delta = H - \hat{H},$$

is what we call the bias, to be used as a measure of the deviation resulting from finite sample effects. We shall assume that $N_{\text{max}} \gg 1$, 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
\begin{equation}
\varepsilon = \frac{1}{N} \sum_{k=N_{\text{max}}+1}^{K} n_k.
\end{equation}
Remember that $N = \sum_{k=1}^{K} n_k$ \cite{16}.

Now, assuming that Zipf’s law persists for $k > N_{\text{max}}$, we have
\begin{equation}
\varepsilon = \frac{1}{N} \sum_{k=N_{\text{max}}+1}^{K} \alpha k^{-\gamma} = \alpha \left( \zeta(\gamma, N_{\text{max}} + 1) - \zeta(\gamma, K + 1) \right),
\end{equation}
where $\zeta(\gamma, m)$ is the Hurwitz or generalized Riemann zeta function. For $k > \gamma$, the following approximation holds \cite{17}
\begin{equation}
\zeta(\gamma, m) = \frac{m^{1-\gamma}}{\gamma - 1} - \frac{m^{-\gamma}}{2} + \frac{m^{-\gamma-1}}{12}.
\end{equation}
Since $K, N_{\text{max}} \gg 1$, we may write
\begin{equation}
\varepsilon = \frac{\alpha}{N(\gamma - 1)} \left( N_{\text{max}}^{1-\gamma} - K^{1-\gamma} \right).
\end{equation}
The value of $\alpha$ 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 $n_{k=N_{\text{max}}+1} \approx 1$, giving $\alpha \approx N_{\text{max}}^\gamma$.

The bias $\delta$ can now be expanded in powers of $\varepsilon$. Keeping terms of order $O(\varepsilon)$ only, we have
\begin{equation}
\delta = -\varepsilon H + (1 + \varepsilon) \left( \varepsilon - \sum_{k=N_{\text{max}}+1}^{K} \frac{n_k}{N} \log \frac{n_k}{N} \right).
\end{equation}
For the conditions stated before, the sum can be approximated by
\begin{equation}
\sum_{k=N_{\text{max}}+1}^{K} \frac{n_k}{N} \log \frac{n_k}{N} = -\varepsilon \left( \log N - \gamma \log \frac{N_{\text{max}}}{K} \right),
\end{equation}
finally giving the result of interest
\begin{equation}
\delta = \varepsilon \left( 1 + \log N - H - \gamma \log \frac{N_{\text{max}}}{K} \right),
\end{equation}
\begin{equation}
\varepsilon \approx \frac{N_{\text{max}}}{N(\gamma - 1)} \left( 1 - \left( \frac{N_{\text{max}}}{K} \right)^{\gamma-1} \right).
\end{equation}
Notice that the true entropy is always underestimated; furthermore $\varepsilon$ is continuous at $\gamma = 1$ \cite{18}. Most of the variation comes from the small parameter $\varepsilon$, whose expression reveals two different effects:
1. the ratio $N_{\text{max}}/N$ reflects the uncertainty of the frequency estimates.

2. the scaling index $\gamma$, 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, $\gamma$ 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

$$\delta = \frac{N_{\text{max}}}{2N}.$$  \hspace{1cm} (14)

We conclude from eq. 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 $\delta$ 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 $H$), then entropies computed from longer words are likely to be significantly corrupted by finite sample effects.

Equation 13 supposes that the maximum lexicographical length $K$ 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 $K = N_{\text{sym}}^{d}$ may do well.

4 Two examples

To briefly illustrate the results, we now consider two examples. The first one is based on a Bernoulli 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 $\lambda$ and the other with probability $1-\lambda$. The block entropy of this process is independent of the word length and equals

$$h = -\lambda \log \lambda - (1-\lambda) \log(1-\lambda).$$  \hspace{1cm} (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 $\gamma$ 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 $\delta$, which is shown in the fourth panel, suggests that the onset of a significant bias occurs around $d \approx 8$; 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 $N_{\text{max}} \gg 1$.

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 $\lambda = 3.8$. 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 $\gamma$ 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 $\gamma$ already suggests that rare words should bring a significant contribution to the entropy. The bias $\delta$ finally suggests stopping at $d = 12$. 
Figure 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 $\gamma$ obtained by fitting the tail of the Zipf-ordered distribution (error bars represent $\pm 1$ standard deviation resulting from the least-squares fit), (4) the bias $\delta$. In this case, entropies cannot be reliably estimated for word lengths beyond $d \neq 9$. Block entropies are normalized to $\log N_{\text{sym}}$, so that the maximum possible value is 1.
Figure 2: Analysis of a logistic map sequence, with the same legend as the previous figure; the string length is $N = 10^d$. 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 $d = 12$. 
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

[8] V. Pareto, Cours d’économie politique (Rouge, Lausanne, 1897).
[16] To be exact $N - d + 1 = \sum_{k=1}^{K} n_k$, but we use the fact that $d \ll N$. 


[18] The term $(K/N_{\text{max}})^{1-\gamma}$ can be large when $K \gg N_{\text{max}}$ 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.