Zipf's law, unbounded complexity and open-ended evolution: Supplementary material*

Bernat Corominas-Murtra[1], Luís F Seoane[2,3,4] and Ricard Solé[3,4,5]

(1) Institute of Science and Technology Austria,
Am Campus 1, A-3400, Klosterneuburg, Austria

(2) Department of Physics, Massachusetts Institute of Technology,
77 Massachusetts Ave, Cambridge, MA 02139, USA

(3) ICREA-Complex Systems Lab, UPF-PRBB. Dr Aiguader 88, 08003 Barcelona, Spain

(4) Institute Evolutionary Biology, UPF-CSIC,
Pg Maritim Barceloneta 37, 08003 Barcelona

(5) Santa Fe Institute, 1399 Hyde Park Road,
87501 Santa Fe, New Mexico, USA

* Published by the *Journal of the Royal Society Interface.*

I. EVERY UNBOUNDED OEE PROCESS IN THE GENERAL SENSE CONTAINS AN UNBOUNDED OEE PROCESS IN THE STRONG SENSE

As introduced in section IIIB, a dynamical system whose description at time step t is σ_t is the result of a process which history is recorded by $\Sigma(t) \equiv \{\sigma_1, \dots, \sigma_t\}$. This is the collection of the description of our dynamical system at each and every time step until t. Note that $\Sigma(t) \subset \Sigma(t+1)$ for every t. Also, note that a given evolutionary history might contain partial evolutionary histories. Imagine, for example, that we fail to record every other instantaneous description. This is likely in empirical setups: we might be able to record a system only once a minute, or once a day.

More rigorously, take a sorted, infinite subset of the natural numbers $T \equiv \{t_1, t_2, t_3, \dots\} \subset \mathbb{N}$. Note that each t_τ is an integer and that not necessarily all integers appear in T, but that this set itself can be labeled by an index τ which runs over all the natural numbers. At each τ, the finite set $T_\tau \equiv \{t_1, \dots, t_\tau\} \subset T$ selects a subset $\Sigma'(\tau) \equiv \{\sigma_{t_1}, \dots, \sigma_{t_\tau}\}$ of the original history at time t_τ, this is: $\Sigma'(\tau) \subset \Sigma(t_\tau)$. Also, $\Sigma'(\tau) \subset \Sigma'(\tau+1)$. We say that the succession of $\Sigma'(\tau)$ for all $\tau \in \mathbb{N}$ is a partial history of the process under research. We also say that the original history (given by the succession of all $\Sigma(t)$ for all $t \in \mathbb{N}$) contains this partial history.

With these definitions it is possible to prove that every unbounded OEE process in the general sense must contain an unbounded OEE process in the strong sense (as illustrated in figure S1). Let us suppose, indeed, that our $\Sigma(t)$ obeys equations (6) and (7) of the main text. At the same time, let us also assume that, among all partial histories of this process, there is not a single one that obeys equation (9) of the main text – this is, that our unbounded OEE process in the general sense does not contain any open-ended partial history in the strong sense. This second assumption will bring us to a contradiction.

The fact that there is not any partial history obeying equation (9) of the main text means that, for whichever partial history that we choose, there is always a finite value μ such that $K(\sigma_{t_\mu}) > K(\sigma_{t_\tau})$ for all $\tau > \mu$. This implies that the description of our system reaches a maximum K^+ at some time $m \leq t_\mu$ with $K^+ \equiv K(\sigma_m)$, and that every description of the system afterwards has at most complexity K^+. The complexity of the process history at this time normalized by the number of steps is some finite number:

$$\langle K(\Sigma(m)) \rangle = \frac{K(\Sigma(m))}{m} < \infty \quad . \tag{1}$$

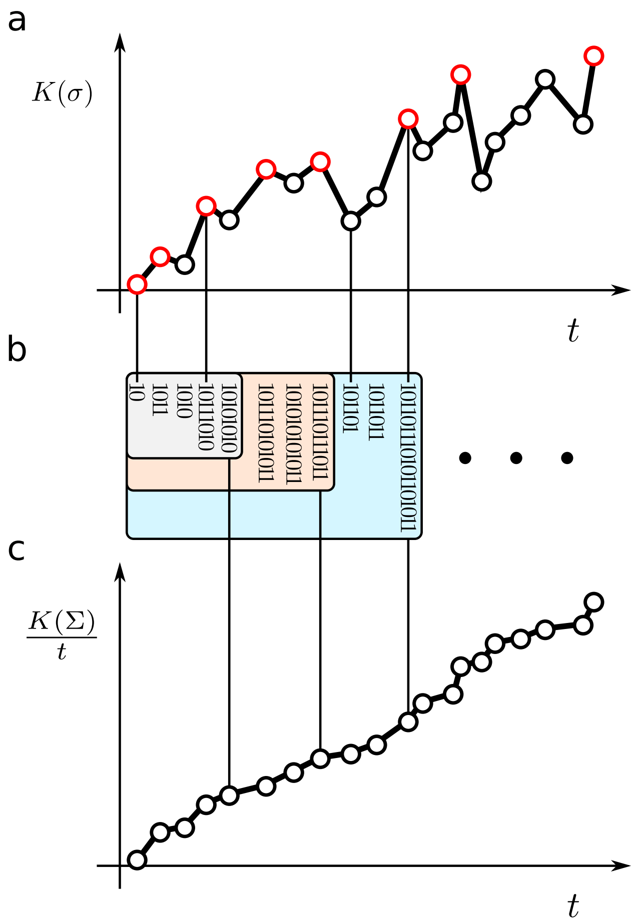

FIG. S1: **OEE in the general versus the strong sense**. a) The description of a system over time changes, not necessarily yielding systems of monotonously growing complexity. b)The evolutionary history of a system at a given time ($\Sigma(t)$) consists of the collection of all previous description of the system up to time t. c) If we deal with OEE in the general sense, the complexity (properly normalized) of the history of a system increases monotonously over time. As we prove in the text, if such a system complies with all axioms for general OEE, it must contain an OEE sub-process in the strong sense, marked with red circles in panel a).

Let us now study the case in which as much complexity as possible is added after this time step. Note that $K(\sigma_{t>m})$ is *at most* K^+. To add as much complexity as possible to the process history, we would need to append incompressible bit strings. The number of incompressible strings of length K^+ is finite, but we can still add them in a patternless fashion, also making sure correlations are not introduced with the history prior to time m. Even if we manage to do so, the complexity of the process history (normalized by the number of time steps) at any time $t > m$ is bounded by:

$$K(\Sigma(t > m)) < \frac{m \langle K(\Sigma(m)) \rangle + (t-m)K^+}{t} \quad . \tag{2}$$

For large enough times, $K(\Sigma(t > m)) \to K^+$ asymptotically. By axiom 1 (Open-Endedness), this process would be a bounded open-ended one if $K^+ > \langle K(\Sigma(m)) \rangle$. This is: bounded open ended processes in the general sense do not necessarily contain an open-ended process in the strong sense. But the upper limit K^+ to the complexity of the process history implies that our process cannot obey axiom 2 (Unboundedness), which is a contradiction because we started out by assuming that our process is unbounded and open-ended in the general sense. Hence, every unbounded open-ended process in the general sense must contain at least a partial history which is open-ended in the strong sense.

We could proceed similarly to prove that at least one partial history must exist that is unbounded in the strong sense. If not, all partial histories must have a finite $N \in \mathbb{N}$ such that there is not any t such that $K(\sigma_t) > N$. All of these bounds are finite, so there must be a maximum $K^+ = \max\{N_i\}$, where N_i is the bound of the i-th partial history of the original process (which is a countable set). We can try to build the most complex such history and come to the conclusion that, *at most*, $K(\Sigma(t)) \to K^+$ as $t \to \infty$. But, again, we departed from the hypothesis that our process is unbounded open-ended, so this is a contradiction because K^+ sets an upper bound in complexity of the process history normalized by number of time steps. Hence every unbounded open-ended process in the general sense must contain an unbounded open-ended process in the strong sense.

II. CONDITIONS FOR THE EMERGENCE OF ORDER IN THE PROBABILITY DISTRIBUTION

Throughout the text we emphasised that the probability distribution is ordered. It is therefore crucial that such ordering is maintained. Here we state the conditions under which the emerging probability distribution is ordered.

The solution that satisfies the postulates of OEE given by equations (13–15) of the main text assumes the existence of some sequence $\theta_2, ..., \theta_n, ...$ of positive real numbers by which:

$$\begin{aligned} (\forall k \leq n) \quad & p_{n+1}(k) = \theta_{n+1} p_n(k); \\ \text{if } k = n+1; \quad & p_{n+1}(k) = 1 - \theta_{n+1} \quad . \end{aligned} \tag{3}$$

This implies a successive process of rescaling of the probabilities, as long as the system grows in size. So, one has:

$$\begin{aligned} p_n(1) &= \theta_2 \cdot \theta_3 \cdot ... \cdot \theta_n = \prod_{k \leq n} \theta_k \\ p_n(2) &= (1-\theta_2) \prod_{2<k\leq n} \theta_k \\ ... &= ... \\ p_n(i) &= (1-\theta_i) \prod_{i<k\leq n} \theta_k \\ ... &= ... \\ p_n(n) &= 1-\theta_n \quad . \end{aligned}$$

We observe that we can establish a recurrence relation between probabilities:

$$\begin{aligned} p_n(1) &= \prod_{1<k\leq n} \theta_k \\ p_n(k) &= a_k p_n(k-1) \quad , \end{aligned}$$

with a_k defined as:

$$a_k \equiv \frac{(1-\theta_k)}{(1-\theta_{k-1})\theta_k} \quad .$$

If $a_k = 1$, then $(\forall k \leq n)\ p_n(k) = \frac{1}{n}$, and $H(X_n) = \log n$. It is easy to see that

$$\theta_k = \frac{k-1}{k} \Rightarrow (\forall k \leq n)\ p_n(k) = \frac{1}{n} \quad ,$$

and, consistently, $1-\theta_k = \frac{1}{k}$. Now let us suppose that the function $1-\theta_k$ is dominated by $\frac{1}{k}$, i.e., $1-\theta_k$ decays faster than $\frac{1}{k}$. Then,

$$\frac{(1-\theta_k)}{(1-\theta_{k-1})} < \frac{k-1}{k} \quad \text{and} \quad \theta_k > \frac{k-1}{k} \quad ,$$

so

$$(\forall k \leq n) \quad a_k = \frac{(1-\theta_k)}{(1-\theta_{k-1})\theta_k} < 1 \quad .$$

The immediate consequence of the above result is that:

$$p_n(1) > p_n(2) > ... > p_n(n) \quad . \tag{4}$$

We observe that if $H(X_n) < \log n$ then $\theta_k \neq \frac{k-1}{k}$. We impose that the solution taken is the one giving

$$1 > \theta_k > \frac{k-1}{k} \quad ,$$

such that equation (4) is satisfied.

III. MINIMISATION OF CONDITIONAL ENTROPY THROUGH THE K-L DIVERGENCE

In this section we will prove that the solution provided by the minimisation of the K-L divergence converges to the absolute minimum of $H(X_{n+1}|X_n)$ in an OEE statistical system. This implies that, even we cannot prove that this is the absolute solution, we can prove that it is arbitrarily close to it.

Let us have the following relation between successive probability distributions:

$$\begin{aligned} p_{n+1}(k) &= \theta_{n+1} p_n(k) \quad \forall k \leq n \\ p_{n+1}(n+1) &= 1-\theta_{n+1} \quad , \end{aligned} \tag{5}$$

which is the solution of the minimisation of the K-L divergence as shown in section IVA of the main text. This leads to an amount of noise:

$$H(X_{n+1}|X_n) = H(\theta_{n+1}) \quad , \tag{6}$$

being $H(\theta_n)$ the entropy of a Bernoulli process having parameter θ_n, i.e:

$$H(\theta_n) = -\theta_n \log \theta_n - (1-\theta_n)\log(1-\theta_n) \quad .$$

Let $\min H(X_{n+1}|H(X_n)) \leq H(\theta_n)$ be the absolute minimum of $H(X_{n+1}|X_n)$ under the conditions of OEE described in equations (13–15) of the main text. We will show that $(\forall \epsilon > 0)$ $\exists M$ for which, for any $N > M$:

$$|\min H(X_{n+1}|H(X_n)) - H(\theta_N)| < \epsilon \quad .$$

Indeed, let us suppose that our system is open-ended. This implies that $1 > \theta_n \geq (n-1)/n$ –see section II of this supplementary material. So, knowing that, by definition $H(\theta_n) = H(1-\theta_n)$ we have that for any $\epsilon' > 0$ $\exists M$ such that, for any $n > M$, $1-\theta_n < \epsilon'$. This implies that, for any $\epsilon > 0$, $\exists M$ such that, for any $n > M$:

$$H(\theta_n) < \epsilon \quad .$$

Since $H(X_{n+1}) > H(X_n)$ by the postulates of OEE, then $\min H(X_{n+1}|H(X_n)) > 0$. In addition, we have proven that $H(\theta_n) < \epsilon$. Therefore, by assuming $\min H(X_{n+1}|H(X_n)) \leq$

$H(\theta_n)$ we demonstrated that, taking the ϵ above defined: $(\forall\epsilon > 0)$ $\exists M$ for which, for any $n > M$:

$$|\min H(X_{n+1}|H(X_n)) - H(\theta_n)| < \epsilon \quad .$$

Therefore, $H(\theta_n)$ converges asymptotically to the absolute minimum.

IV. DERIVATION OF ZIPF'S LAW FROM ENTROPY CONSTRAINTS

Assume that the unboundedness condition is given as follows: there exists a unique $\mu \in (0, 1)$ such that $(\forall\epsilon > 0)(\exists N) : (\forall n > N)$:

$$\left|\frac{H(X_n)}{\log n} - \mu\right| < \epsilon \quad . \tag{7}$$

Now we want to find the asymptotic behavior of p_n, $n \to \infty$ under the above justified conditions given by equation (7) of this supplementary material and (16) of the main text. The key feature is that the following quotient:

$$(\forall k + j \leq n) \;\; f(k, k + j) = \frac{p_n(k + j)}{p_n(k)} \quad , \tag{8}$$

does not depend on n. Therefore, along the evolutionary process, as soon as

$$p_n(k), p_n(k + j) > 0 \quad ,$$

$f(k, k + j)$ remains invariant.

Now suppose that we have $p'_n \sim i^{-\gamma}$. The explicit form of its (normalized) entropy is:

$$\frac{H(X'_n)}{\log n} = \frac{1}{\log n}\left(\frac{\gamma}{Z_\gamma}\sum_{i\leq n}\frac{\log i}{i^\gamma} + \log Z_\gamma\right) \quad . \tag{9}$$

where Z_γ is the normalization constant. From the above expression, we find that, if $(\forall\delta > 0,\ n > m)(\exists N)$ such that:

$$(\forall m > N) \quad f(m, m + 1) < \left(\frac{m}{m + 1}\right)^{1+\delta} \quad ,$$

then $(\exists C < \infty \in \mathbb{R}^+)$ such that $(\forall n)(H(X_n) < C)$, leading to

$$\lim_{n\to\infty}\frac{H(X_n)}{\log n} = 0 \quad ,$$

which contradicts the assumptions of the problem, depicted by equation (7). Therefore, during the growth process,

$$f(m,m+1) > \left(\frac{m}{m+1}\right)^{(1+\delta)} \quad , \tag{10}$$

with δ arbitrarily small, provided that n can increase unboundedly. Furthermore, we observe that, if $(\forall\delta > 0,\ n > m)(\exists N)$ such that

$$(\forall m > N)\ \ f(m,m+1) > \left(\frac{m}{m+1}\right)^{(1-\delta)} \quad ,$$

then, from equation (9), one finds that:

$$\lim_{n\to\infty}\frac{H(X_n)}{\log n} = 1 \quad ,$$

again in contradiction to equation (7), except in the extreme, pathological case where $\mu = 1$, which has been ruled out by assumption. Accordingly,

$$f(m,m+1) < \left(\frac{m}{m+1}\right)^{(1-\delta)} \quad . \tag{11}$$

Combining equation (10) and (11), we have shown that the asymptotic solution is bounded by the following chain of inequalities:

$$\left(\frac{m}{m+1}\right)^{(1+\delta)} < f(m,m+1) < \left(\frac{m}{m+1}\right)^{(1-\delta)} \quad .$$

The crucial step is that, if $\mu \in (0,1)$, using the fact that equation (9) defines a continuous, smooth function in terms of γ, one can conclude that for $n \to \infty$,

$$\delta \to 0 \quad .$$

This implies, in turn, that, for $n \gg 1$:

$$f(m,m+1) \approx \frac{m}{m+1} \quad ,$$

and, from the definition of f provided in equation (8), we conclude that:

$$p_n(k) \propto \frac{1}{k} \quad ,$$

leading us to Zipf's law as the unique asymptotic solution.

V. DIVERGENCE OF THE NORMALIZATION CONSTANT

Given the expression of $p_n(1)$ obtained above:

$$p_n(1) = \prod_{2<k\leq n} (\theta_k) \quad ,$$

one can define the normalisation constant

$$C_n = \prod_{2\leq k\leq n} (\theta_k)^{-1} \quad . \tag{12}$$

We observe that we can rewrite the probability distribution p_n in the following form:

$$\frac{1}{C_n}, \frac{1-\theta_2}{C_n}, \frac{1-\theta_3}{\theta_2 C_n}, \dots$$

following the above series, it is not difficult to see that, $(\forall i)(1 < i \leq n)$:

$$p_n(i) = \frac{1-\theta_i}{C_n} \prod_{2\leq k<i} (\theta_k)^{-1} \quad .$$

Now we connected the parameters related to the increase of the entropy during the evolutionary path and the normalisation constant of the distribution. This normalisation constant will be the key of our argument. Indeed, thanks to the properties of the Riemann ζ function it is known that, if $(\exists \epsilon > 0) : (\exists m) : (\forall k > m)$

$$\left(\frac{p_n(k+1)}{p_n(k)}\right) < \left(\frac{k}{k+1}\right)^{-(1+\epsilon)} \quad ,$$

(i.e. the probability distribution p_n is *dominated* by the probability distribution $q_n(i) \propto i^{-1-\epsilon}$), then $(\exists N \in \mathbb{N}) : (N > C_\infty)$, being

$$C_\infty = \lim_{n\to\infty} C_n \quad .$$

This means that,

$$(\forall n)\; p_n(1) \geq \frac{1}{C_\infty} > 0 \quad .$$

However, in the case of p_n not being dominated by any $q_n(i) \propto i^{-(1+\epsilon)}$, things go differently, since

$$\lim_{n\to\infty} C_n = \infty \quad .$$

The direct consequence for the above consideration is that $(\forall \epsilon > 0)(\exists M)$ such that, if $n > M$, then

$$p_n(1) < \epsilon \quad .$$

The presence of an upper bound in the Shannon entropy is directly related to the divergence or convergence of C_n. If the probability distribution is not dominated by $q_n(i) \propto i^{-(1+\epsilon)}$ for any $\epsilon > 0$, then the entropy, as well as C_n, diverges. On the contrary, if p_n is dominated by $q_n(i) \propto i^{-(1+\epsilon)}$ for any $\epsilon > 0$ the entropy converges and so does C_n. In formal terms:

$$\left(\lim_{n\to\infty} C_n = \infty\right) \Leftrightarrow \left(\lim_{n\to\infty} H(X_n) = \infty\right) \quad . \tag{13}$$

Zipf's law is thus at the twilight zone between bounded and unbounded complexity. Accordingly, for an OEE system under conditions described by equations (13–15), thanks to the bound on mutual information between an arbitrary past state m and the current one n given by equation (21) of the main text, one concludes that:

$$\lim_{n\to\infty} I(X_m : X_n) \leq \lim_{n\to\infty} \frac{C_m}{C_n} H(X_m) = 0 \quad ,$$

in words, that all past information is lost.