Comment on “Universal and accessible entropy estimation using a compression algorithm”

In a recent Letter [1] a framework for estimating entropy was introduced and applied to one-dimensional and two-dimensional systems. In this Comment we show that the method is not well suited for estimating entropy in bidimensional systems presenting long-range correlations.

In Ref. [1], entropy is evaluated following this scheme: 1) discretize the considered configurations 2) store them in a 1D file 3) measure the compressed file size with a lossless compression algorithm 4) estimate the incompressibility and map it to the asymptotic entropy $S_A$. Here we test this scheme for binary variables, and we do not consider the complication produced by the coarse-graining of continuous variables.

In general, lossless compression algorithms are known to present slow entropy convergence and alternative more efficient methods are used [2]. There is a rich literature studying 1D systems but the analyses of multidimensional patterns are very few, and they are expected to present new features compared to the 1D case. Some methods are known to apply to patterns of arbitrary dimension (e.g. block entropies). In contrast, mapping multidimensional patterns to a one-dimensional sequence can be path-dependent, loses bidimensional correlations, and can even produce spurious long-range correlations [3]. The use of a locality-preserving curve, like Hilbert’s one, does not guarantee to solve these difficulties.

Systems which present long-range correlations can be used to test these effects. In the original Letter, long-range correlations are matched only in the 2D NN Ising model near the critical point. In that region, the estimated $S_A$ displays a relative error between 5% and 10%. Unfortunately, this Ising model is a particular exception where the reduction to a 1D string is known to not affect the statistics that determine entropy estimation [4]. For this reason, this system is not a good benchmark for testing entropy estimation in 2D patterns with long-range correlations. For testing the Avivery’s framework in this situation, here we consider a set of 68 different patterns obtained from built form maps of cities around the world [5]. These binary matrices of size 1000 × 1000 are particularly well suited for the test. The extensive part of their block entropies diverges [5] [6], implying the presence of very long-range correlations that make the entropy estimation particularly difficult. For these symbolic sequences we can not obtain the exact entropy values with analytical methods. Instead, we can use a reliable block-entropies method [7] which estimates the entropy by using differential entropies ($S_B$). Results are robust, as alternative block-entropies methods give equivalent outcomes [5].

Fig.1 compares the results obtained by the classical block-entropies method ($S_B$) with the ones generated by the compression algorithm ($S_A$). This second approach dramatically overestimates the entropy. The median of the percentage error is 91% and the scattering plot shows that the difference between $S_A$ and $S_B$ grows for larger $S_B$ values. The reduction of 2D patterns to a 1D string significantly destroys the involved bidimensional structures, which are particularly significant in systems with long-range correlations, and generates a substantial overestimation of the entropy.

Even if Avivery’s method is indeed elegant and computationally effective, it can not be considered accurate for general 2D systems.

E. Brigatti
Instituto de Física, Universidade Federal do Rio de Janeiro, Av. Athos da Silveira Ramos, 149, 21941-972, Rio de Janeiro, RJ, Brasil.

F.N.M. de Sousa Filho
Instituto de Computação, Universidade Federal do Rio de Janeiro, Av. Athos da Silveira Ramos, 274, 21941-916, Rio de Janeiro, RJ, Brasil.

[1] R. Avivery, M. Kornreich, and R. Beck, Universal and accessible entropy estimation using a compression algorithm, Phys. Rev. Lett. 123, 178102 (2019).
[2] Schürmann T, Grassberger P. Entropy estimation of symbol sequences. Chaos, 6, 414 (1996).

[3] Grassberger P. Toward a quantitative theory of self-generated complexity. Intl. J. Theo. Phys., 25, 907 (1986).

[4] Z Alexandrowicz. Stochastic models for the statistical description of lattice systems. J. Chem. Phys., 55 2765, (1971); S. Goldstein, R. Kuik, and A.G. Schlijper, Entropy and global Markov properties, Commun. Math. Phys., 126, 469 (1990).

[5] E Brigatti, VM Netto, FNM de Sousa Filho, and C Cacholas, Entropy and hierarchical clustering: characterizing the morphology of the urban fabric in different spatial cultures. Chaos, 31, 113138 (2021).

[6] Netto, VM, Brigatti E, Meirelles J, Ribeiro FL, Pace B, Cacholas C, Sanches P. Cities, from Information to Interaction. Entropy, 20, 834 (2018).

[7] Feldman DP, Crutchfield JP. Structural information in two-dimensional patterns: Entropy convergence and excess entropy. Phys. Rev. E, 67, 051104 (2003).