On the Dynamics of Multi-information in Cellular Automata

Gregor Chliamovitch, Bastien Chopard, and Alexandre Dupuis

Department of Computer Sciences, University of Geneva, Switzerland
Gregor.Chliamovitch@unige.ch

Abstract. After reviewing a few key quantities of information theory, we investigate in this paper the behaviour of multi-information in elementary cellular automata. It will turn out that the usual classification by Wolfram is not well supported in terms of this information measure, or, more likely, that multi-information is blind to the kind of complexity displayed by those automata.

1 Introduction

While for decades information theory did not spread much out of the frontiers of the domain it had been designed for (apart from entropy which originates from thermodynamics and is so deeply grounded into statistical physics [6]), in recent years several attempts have been made to take advantage of the whole information-theoretic toolkit in the study of statistical mechanical systems as well as in the study of complexity in the large [1]. It is nevertheless not clear thus far whether or not information theory has something fundamental to tell us about the nature of complexity at all.

Any attempt to cope with complexity has sooner or later to face one-dimensional cellular automata, since these are among the simplest complex systems one can think of. Moreover, they display a wide range of behaviours, that can be classified according to the complexity of the pattern they unfold over time. While the most famous such classification is due to Wolfram [11], it is regularly challenged and many alternatives have been proposed (see [5] and references therein).

If indeed information theory has deep insights to reveal on the nature of complexity, it is likely that automata which are very different in terms of Wolfram's notion of complexity should also exhibit different behaviours when considered from an information-theoretic viewpoint. Accordingly, our aim in this paper is to investigate the temporal behaviour of multi-information over time in all 256 elementary automata. It will turn out that while observed behaviours fall into a limited number of categories, this classification is definitely at odds with more usual ones.

We start in section 2 with an express review of information theory, putting emphasis on the concept of *multi-information* which, in our view, bears a particular relevance. We consider Wolfram's classification in section 4, but before that

J. Wąs, G.C. Sirakoulis, and S. Bandini (Eds.): ACRI 2014, LNCS 8751, pp. 87–95, 2014.

we will have to pause in order to discuss how those quite abstract information-theoretic concepts may be translated in operational terms. This is the purpose of section 3.

2 Review of Information Theory

The most fundamental building block of information theory is *entropy*. Although the term was introduced first in the context of thermodynamics, it soon received an interpretation as a measure of uncertainty contained in a probability distribution $p(X)$. More precisely, *Shannon entropy* is defined as [9]

$$H(X) := \sum_x p(x) \ln \frac{1}{p(x)}. \tag{1}$$

It provides a characterization of the uncertainty on a variable X which is unique with respect to a set of intuitive axioms [4]. In the same way, *mutual information* provides a unique caracterization of how much the knowledge of one variable X impacts the prediction on another, Y. We simply calculate the reduction of entropy brought by the knowledge of Y, i.e.

$$I(X,Y) := H(X) - H(X|Y). \tag{2}$$

where $H(X|Y)$ denotes the entropy of the conditional density $p(X|Y)$. A simple calculation shows that I can be re-written as

$$I(X,Y) = H(X) + H(Y) - H(X,Y) = \sum_{x,y} p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \tag{3}$$

and is therefore obviously symmetric. In terms of the so-called *Kullback-Leibler (KL) divergence*, which provides a measure of pseudo-distance in the space of distributions, mutual information can be seen to quantify, in a sense, how far the variables are from being independent of each other [3].

To give an example, if X and Y are both binary variables, and the joint probability is given by $p(0,0) = 0.1$, $p(0,1) = 0.2$, $p(1,0) = 0.4$ and $p(1,1) = 0.3$, the marginal probability of X is given by $p_X(0) = p(0,0) + p(0,1) = 0.3$ and $p_X(1) = p(1,0) + p(1,1) = 0.7$, while the marginal probability of Y is given by $p_Y(0) = p(0,0) + p(1,0) = 0.5$ and $p_Y(1) = p(0,1) + p(1,1) = 0.5$. The mutual information can now be calculated to be

$$\begin{aligned} I(X,Y) &= p(0,0) \ln \frac{p(0,0)}{p_X(0)p_Y(0)} + p(0,1) \ln \frac{p(0,1)}{p_X(0)p_Y(1)} \\ &\quad + p(1,0) \ln \frac{p(1,0)}{p_X(1)p_Y(0)} + p(1,1) \ln \frac{p(1,1)}{p_X(1)p_Y(1)} \\ &= 0.1 \ln \frac{0.1}{0.15} + 0.2 \ln \frac{0.2}{0.15} + 0.4 \ln \frac{0.4}{0.35} + 0.3 \ln \frac{0.3}{0.35} \\ &\simeq 0.024. \end{aligned} \tag{4}$$

Many generalizations of mutual information to more than two variables have been proposed over time (see, among others, [8]), but since any such generalization is usually crafted in order to find use in a specific domain of research, none has actually met universal consensus yet. In our view the most intuitive one is found in a quantity known as *multi-information* (and sometimes -less properly- *total correlation*). It is defined as [10]

$$M(X_1, ..., X_N) := \sum_{i=1}^{N} H(X_i) - H(X_1, ..., X_N). \tag{5}$$

When expressed directly in terms of probabilities, that is

$$M(X_1, ..., X_N) = \sum_{x_1,...,x_N} p(x_1, ..., x_N) \log \frac{p(x_1, ..., x_N)}{p(x_1)...p(x_N)}, \tag{6}$$

it becomes obvious that multi-information can be understood as the KL distance to independence between variables, and in this respect plays the same role in the multivariate case that mutual information plays in the bivariate one. This is the key quantity we will be considering in the following sections, since it treats all variables on an equal footing and does not introduce spurious distinctions between variables.

3 The Markovian Framework

An important restriction regarding the use of entropy, mutual or multi-information is that it requires the full knowledge of the probability density of the system under consideration. This knowledge of the density may be lacking sometimes, in which case the formalism breaks down, but nonetheless systems with known distribution represent a wide class on which theoretical and numerical experiments can be carried through.

Things become a bit more tedious when we want to investigate the temporal behaviour of information measures, which requires knowing how the probability density itself changes over time. Often this is done using Monte-Carlo methods, by evolving copies of the system and reconstructing the probabilities by sampling trajectories. This is for instance the approach adopted in [7], where another kind of information-theoretical measure is discussed. It has the drawback that we are then exposed to sampling errors.

An alternative way to proceed would be to determine the evolution rule itself, according to which the density evolves. Could we do that, we could so to speak follow all trajectories at once, even the least probable ones[1]. This amounts to a

[1] An illustration is provided by Wolfram's rule 40 which brings *almost* all configurations to the state where all cells take value 0, *except* for a few periodic configurations. While such configurations could well be missed by an inadequate sampling, the alternative method keeps them as long as they are not explicitly assigned a vanishing probability. See section 4 below.

description in terms of Markov chains, where the knowledge of history allows us to predict towards which states the system could evolve. We will restrict ourselves here to the case that the knowledge of a *finite* history is sufficient to predict possible futures. By extending the state space, actually all such processes can be recast in the form of *memoryless Markov chains* (or simply *Markov chains*), by what we mean that the forthcoming states can be predicted knowing the current state of the process only ; previous states do not alter the future in any way.

While this second approach seems to outperform the usual sampling in terms of accuracy, it actually suffers from its numerical cost. Assume for instance we deal with a dynamical system constituted by N agents taking binary values in $\{0, 1\}$. We have in this case 2^N possible configurations, while assuming the system is driven by a dynamics with a k-steps memory we have 2^{kN} possible relevant histories to keep into consideration, which becomes soon untractable even for small values of N and k.

Nonetheless this formalism has some advantages that cannot be given up easily. In particular it allows a more straightforward transition from numerical exploration to theoretical investigation of information measures in complex systems, and therefore seems an approach worth being promoted. Still more importantly, as we already mentioned, this approach does not require to select (arbitrarily) an initial configuration, but handles them all as long as they are not explicitly assigned probability zero[2].

In the following, we will therefore focus on Markovian processes of order $k = 1$ for the sake of tractability. In more mathematical terms, denoting by $W_{SS'}$ the probability of transition from state $S = (s_1, ..., s_N)$ (s_i denoting the i-th node) to state $S' = (s'_1, ..., s'_N)$, the formula ruling the joint density of states considered at two consecutive times is written as

$$p(S', t+1; S, t) = p(S, t)W_{SS'}. \tag{7}$$

Iterating this formula allows to write

$$p(S', t+\tau; S, t) = p(S, t)(W^\tau)_{SS'}, \tag{8}$$

and starting from this expression we can recover the probability of being in a state S' at a later time by summing on former positions, so as to get

$$p(S', t+\tau) = \sum_S p(S', t+\tau; S, t) = \sum_S p(S', t)(W^\tau)_{SS'}. \tag{9}$$

[2] Admittedly there remains some arbitrariness inasmuch as we have to select *some* initial distribution. While the transient phase will be affected by a change of initial density, the conclusions we draw regarding the equilibrium regime are nevertheless independent of the inital distribution as long as the system is ergodic. This is not always the case, but still much less restrictive than assuming that the system tends towards the same configuration whatever its initial configuration - which is certainly wrong in the case of the elementary cellular automata dealt with in this paper.

Given the full density of the system, univariate marginals appearing in the definition of multi-information are calculated by summing on irrelevant nodes. Assuming for instance we need the marginal of, say, node s_1, we have to compute

$$p(s_1, t) = \sum_{s_2,\ldots,s_N} p(S, t). \tag{10}$$

The interplay between nodes and states is a major source of confusion and requires careful attention when coming to implementation.

Since multiplying matrices of size n is a process of order $O(n^3)$, noting that $k = 1$ implies $n = 2^N$ means that the effort required by the calculation of successive probability densities grows exponentially with the number of nodes (not to mention the corresponding memory saturation).

4 Elementary Cellular Automata

In order to check whether or not Wolfram's classification can be recovered by considering the behaviour of multi-information, we scanned all 256 rules (actually only the 88 of them that are non-equivalent) and investigated how, starting from an uniform initial distribution (therefore with $M = 0$), the distribution evolves during 50 steps, which for the small systems investigated here is long enough to reach the stationary regime, and calculated the associated multi-information. Due to lack of space, only eight typical behaviours are displayed here (figure 1).[3]

We may note that while different kinds of patterns are obtained, they fall into a limited number of types, whose figure 1 provides a representative sample. These patterns are essentially characterized by 1) the length of their transient phase 2) the fact they converge either to a stable value or reach a periodic regime 3) the value of M to which they converge 4) the fact that these characteristics are independent of each other.

Rule 154 below offers a good example of a rule whose M vanishes. It is nevertheless not typical in that it stays at $M = 0$ forever and does not go through a correlated phase. Another rule displaying this behaviour is 0, but in this case this is easily understood since it jumps directly from an uncorrelated initial to a density where only the state having all nodes 0 survives. Rule 154, on the other side, displays a non-trival configurational pattern.

Rule 164 (Wolfram class II) provides an example of an automaton reaching a partly correlated phase after two iterations, followed by a decay of M towards some finite value. This behaviour is frequently encountered, with some variations regarding the intensity of the peak and the final value of M, and this for classes I and III as well as II. Rule 40 is an interesting instance of class I automaton displaying this behaviour. This rule converges to the empty configuration for most of the possible initial configurations, but in some scarcer cases it converges

[3] The results obtained for all 256 rules can be found on the webpage *http://www.cui.unige.ch/∼chopard/Sophocles*. Patterns for mutual information are displayed there as well.

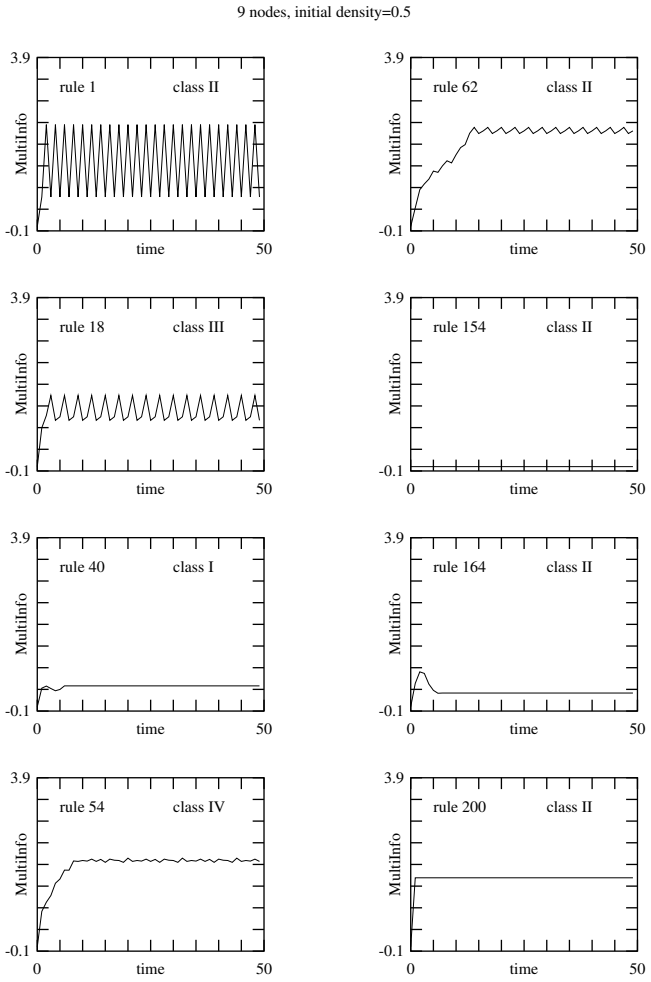

Fig. 1. A catalogue of typical behaviours displayed by multi-information in elementary cellular automata

to other stable configurations. Indeed the multi-information is seen to stabilize at some non-zero value indicating that the stable configuration is not unique when all configurations are initially taken into account.

Rule 200 (class II) shows another frequent behaviour. Here M happens to reach immediately its equilibrium value, but sometimes the transient phase can be longer. This situation is encountered in all first three classes. Note that the equilibrium value of M varies considerably ; see for instance 78 which reaches the highest M we observed.

A striking example of periodic regime is provided by rule 1, belonging to class II. Although its oscillations are unusually strong, it is a nice example of an automaton with period 2. This can be observed by looking at the configurational

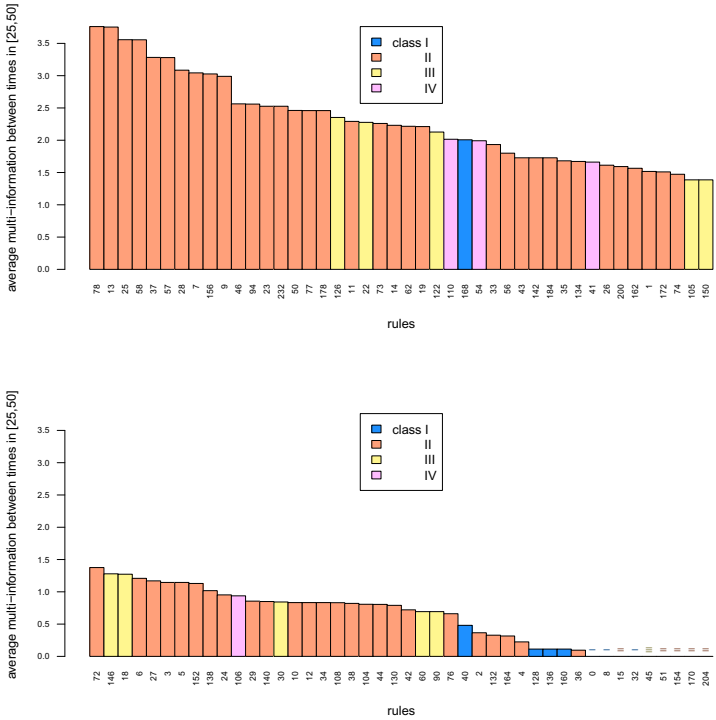

Fig. 2. All 88 elementary automata ranked according to their informational content in the stationary regime

pattern as well. A more intriguing case is 18 (class III), which oscillates with period 3. This could nevertheless not have been expected when looking at the configurational pattern, which displays a chaotic behaviour typical of class III. On the other side, 94 also exhibits oscillations of longer period, but belongs to class II. This is also the case of 62, which shows an unusually long transient phase. Interestingly this oscillatory multi-information is observed in classes II, III and IV but never in class I.

An instance of class IV is provided by rule 54. The period here is particularly long, but other rules form class IV have shorter ones so that it should not be considered as a specific feature of complex rules.

To summarize, it seems that while it is indeed possible to establish a classification of rules based on the behaviour of the multi-information, it appears hazardous to make a link with complexity in the sense of Wolfram since all kinds of patterns discussed above can be found in almost all four classes. The most we can say is that when considering both ends of the complexity spectrum, that is classes I and IV, patterns of multi-information can be safely attributed. It is difficult otherwise to discriminate between classes I and II, II and III and III and IV. Some patterns tend nevertheless to appear more (resp. less) frequently when Wolfram's complexity increases, so that it might be possible that some

distinctive features eventually emerge when considered statistically. Indeed it is shown in [2] that another complexity measure known as *Langton parameter* may be linked to a subtler information-theoretic tool.

Figure 2 provides another approach to the same conclusion, by displaying all 88 rules ranked according to the value of M they converge to (for periodic rules we indicate the average on a period). Classes are scattered and no obvious clustering emerges. Note that interestingly none of the rules converges to an M in the range $(2.5, 3)$, while on the contrary many of them converge towards $M \approx 0.9$.

Finally, we would like to emphasize that this study holds for small systems due to the constraints imposed by the Markovian formalism, and that it is not certain that our conclusions may be translated as such to large systems (even if experiments carried through different sizes suggest that while the equilibrium value of M changes with size, the qualitative behaviour remains the same). Size effects in cellular automata, which turn out to be rather subtle, are further adressed in [2].

5 Conclusion

Our goal in this paper was to confront information-theoretic methods with a well-documented instance of complex systems. It turned out that this tool appeared rather unsensitive to the kind of complexity put forward in Wolfram's classification and is, as such, hardly suitable as a basis for an alternative classification of these systems. Some clues suggest nonetheless that a more elaborated treatment might reveal some statistical regularities.

Acknowledgements. The authors acknowledge funding from the European Union Seventh Framework Programme (FP7/2007- 2013) under grant agreement number 317534 (Sophocles project). They also gratefully acknowledge the anonymous reviewers for their valuable comments.

References

[1] Ay, N., Olbrich, E., Bertschinger, N., Jost, J.: A Geometric Approach to Complexity. Chaos 21 (2011)
[2] Chliamovitch, G., Chopard, B., Velasquez, L.: (to appear)
[3] Cover, T., Thomas, J.: Elements of Information Theory. Wiley-Interscience, New York (2006)
[4] Khinchin, A.I.: Mathematical Foundations of Information Theory. Dover, New York (1957)
[5] Martinez, G.: A Note on Elementary Cellular Automata Classification. arXiv, 1306-5577 (2013)
[6] Penrose, O.: Foundations of Statistical Mechanics. Pergamon, Oxford (1969)
[7] Quax, R., Apolloni, A., Sloot, P.M.A.: The Diminishing Role of Hubs in Dynamical Processes on Complex Networks. Journal of the Royal Society Interface 10(88) (2013)

[8] Schreiber, T.: Measuring Information Transfer. Physical Review Letters 85(2), 461–464 (2000)
[9] Shannon, C.: The Mathematical Theory of Communication. Bell System Technical Journal 27, 379–439, 623–656 (1948)
[10] Watanabe, S.: Information Theoretical Analysis of Multivariate Correlation. IBM Journal 14(3), 66–82 (1960)
[11] Wolfram, S.: Statistical Mechanics of Cellular Automata. Reviews of Modern Physics 55(3), 601–644 (1983)