Complexity Measures and Physical Principles

Karoline Wiesner

School of Mathematics, University of Bristol, Bristol, U.K.

Abstract. I summarise some recent results on physics of information and discuss their impact on our understanding of complexity. A surprisingly coherent picture of measures of complexity, information theoretic functions, and thermodynamical quantities arises.

1 Introduction

Complexity is a concept thoroughly defeating any unique definition. One notion is that a complex system is an open system exchanging matter and / or signals with the environment while being robust in its structure and adapting to environmental circumstances without any internal central control unit [11]. One can take this view into a more quantitative realm by considering a complex system as an information processor. The above description can then be recast as follows: A complex system takes in and puts out information, it has an internal memory, and robust and decentralised information processing capacity. For complex systems this is a very helpful starting point as it allows quantitative statements about systems of which we do not know the Hamiltonian or any other form of energetic characterisation. The idea to put physics – which deals with matter and energy – and information – which deals with bits and memory on the same footing is not new. The most pointed statement in this context is probably Wheeler's "It from bit", implying that every physical quantity derives its function from answers to yes-no questions [20].

In the following I will sketch some recent developments in the area of physics of information and outline parallels to information-theoretic treatments of complex systems.

2 Measures of Complexity

One way of measuring the complexity of a system is to quantify the amount of correlation a system exhibits in its temporal behaviour. The temporal behaviour is captured in observations over time and can be formalised as a *stochastic process*. The most generally applicable measure of correlation in a stochastic process is the mutual information which is a function of the joint probability distribution of two random varibles[1]. We will consider the past data to be one random

[1] For definitions of Shannon entropy, conditional entropy, and mutual information, see for example [2].

A. Sanayei et al. (eds.), *ISCS 2014: Interdisciplinary Symposium on Complex Systems*,
Emergence, Complexity and Computation 14, DOI: 10.1007/978-3-319-10759-2_2

variable, $\overleftarrow{X}$, and the future data to be another random variable, $\overrightarrow{X}$. In the limit of infinite data the mutual information between these two variables has been taken as a measure of complexity ([13,4]):

$$E = I[\overleftarrow{X}; \overrightarrow{X}] \ . \tag{1}$$

E has been given different names, such as *predictive information*, excess entropy, effective measure complexity, and stored information (see [4] and refs. therein).

We can interpret E as the amount of communication between a system's past and its future. Any positive mutual information means that the system remembers some of its past and its future behaviour is influenced by that memory. In this sense, E is the system's complexity, measured in bits.

I now turn to a second measure of complexity, the *statistical complexity*, as introduced by Crutchfield and Young [5]. It is related to the *true metric complexity* initially discussed by Grassberger [8]. Given a stochastic process, the provably unique minimal (in terms of information stored) and optimal (in terms of prediction) computation-theoretic representation summarising the regularities of the process is the so-called ϵ-machine [5,13]. It consists of an output alphabet $\mathcal{X}$, a set of *states* $\mathcal{S}$ and stochastic transitions between them. For every pair of states $s, s' \in \mathcal{S}$ probabilities $\Pr(S_i = s', X_i = x | S_{i-1} = s)$ are defined for going from state s to state s' while outputting symbol $x \in \mathcal{X}$. The *statistical complexity* of a process is defined as the Shannon entropy over the stationary distribution of its ϵ-machine's states:

$$C_\mu := -\sum_s \Pr(s) \log \Pr(s) \ . \tag{2}$$

C_μ is the number of bits required to specify a particular state of the ϵ-machine. Since knowing the state means knowing everything there is to know about future observations, C_μ is the number of bits that need to be stored to optimally predict future data points from the ϵ-machine.

The statistical complexity is bounded from below by the predictive information: $E \leq C_\mu$ [13]. It is intuitive that the amount of memory of a predicting device should be at least as big as the number of bits it is predicting about a system. There have been different approaches to derive the exact difference between E and C_μ analytically. Crutchfield and co-workers explained it through the asymmetry between prediction and retrodiction [6]. They were able to show that the predictive information is equal to the mutual information between the states predicting the stochastic process forward in time and the states retrodicting the process backward in time. Wiesner et al. took a different approach to explaining the difference between the two complexity measures [19]. They considered the computational irreversibility of the ϵ-machine and showed that it accounts exactly for the difference between the two complexity measures. The extra memory kept by the ϵ-machine which is not predictive incurs an entropic cost to the computation. Computing E (Eq. 1) requires numerical approximation from finite data. Finding an analytic expression for the predictive information from the predictive states is an open problem.

Gu et al. extended the framework of ϵ-machines by allowing the states to have quantum mechanical properties [9]. Once the (classical) states $\mathcal{S}$ are known the quantum mechanical states (here in ket notation) can be easily constructed as

$$|s\rangle = \sum_{s' \in \mathcal{S}} \sum_{x \in \mathcal{X}} \sqrt{\Pr(S_i = s', X_i = x | S_{i-1} = s)} |x\rangle |s'\rangle \ , \tag{3}$$

This is related to an earlier quantum model of a stochastic process [12]. Gu et al. define the quantum complexity of the process as the von Neumann entropy over the resulting density matrix $\rho = \sum_s \Pr(s)|s\rangle\langle s|$:

$$C_q := -\mathrm{Tr}\rho \log \rho \ . \tag{4}$$

It can be easily shown that C_q is bounded above and below by the statistical complexity and the predictive information, respectively:

$$E \leq C_q \leq C_\mu . \tag{5}$$

Equality holds if and only if the upper and lower bound coincide [9,19]. Whenever C_q is strictly less than C_μ the quantum states $|s\rangle$ are non-orthogonal. This means that there exists no quantum measurement which could perfectly distinguish between them. Since the machines both predict E bits about future observations the additional bits stored in the classical machine are non-predictive. This, again, indicates that the difference between E and C_μ represents an inefficiency in the model. Explaining the gap between C_q and E remains an open problem.

3 Finding an Optimal Model under Constraints

Finding an efficient model is a classical optimisation problem under costraints: One wants to minimise the size of a model while maximising its predictive power. Any model is some form of summary of past observations. A good model retains as much *relevant* information about the past as possible and discards anything beyond that. The so-called *information bottleneck method*, introduced by Tishby, Pereira and Bialek, provides the tools for finding a minimal model under the constraint of maximal predictive power [18]. They formalise this task as finding an optimal compression of data (the "summary") under the constraint of minimal error upon decompression (equivalent to "prediction" of observations). Furthermore, the bottleneck method allows for the case where the data, X, are descriptions of another, correlated variable Y. The optimisation task they solve is to compress the data X into $\tilde{X}$ as much as possible while retaining as much information about Y as needed. Tishby et al. were able to show that the optimal compression minimises the relative entropy (also known as Kullback-Leibler divergence) between the conditional distributions $\Pr(Y|X)$ and $\Pr(Y|\tilde{X})$, $D_{KL}[\Pr(Y|X)||\Pr(Y|\tilde{X})]$.

Still et al. applied the bottleneck method to the problem of finding a predictive model for a stochastic process as discussed above [16,15], see also [14]. Here the future behaviour $\overrightarrow{X}$ of the system corresponds to the variable of interest Y. Past observations $\overleftarrow{X}$ are analogous to the data X to be compressed. The states $\mathcal{S}$ and associated transition probabilities are the compressed version $\tilde{X}$ of the data. A minimal model under the constraint of maximal predictive power can be found by solving the following optimisation problem:

$$\min_{\Pr(\mathcal{S}|\overleftarrow{X})} \left(I(\overleftarrow{X};\mathcal{S}) + \beta I(\overleftarrow{X};\overrightarrow{X}|\mathcal{S}) \right) \ , \tag{6}$$

where β is the remaining parameter (Lagrange multiplier) representing the trade-off between model complexity and predictive power. It turns out that the solution to this optimisation problem minimises the relative entropy $D_{KL}[\Pr(\overrightarrow{X}|\overleftarrow{X})||\Pr(\overrightarrow{X}|\mathcal{S})]$. In the limit of $\beta \to \infty$ equivalent to unrestricted model complexity, the states of the ϵ-machine are recovered [16].

Still et al. expanded this framework to a more general setting [17]. The view point taken is slightly different from that of optimally modeling observations of a complex system's dynamics. Rather, they consider a complex system as making a model of the world. The system remembers some of the past experiences in the world and tries to predict future experiences. Mathematically this framework is identical to the one discussed above. Still et al. define *predictive power* as the mutual information between the system's state s_t at time t and the environmental signal x_{t+1} at time $t+1$. Similarly, *instantaneous memory* is defined as the mutual information between the system's state s_t at time t and the environmental signal x_t at time t:

$$I_{\text{mem}}(t) = I(s_t;x_t) \ , \tag{7}$$

$$I_{\text{pred}} = I(s_t;x_{t+1}) \ . \tag{8}$$

Both I_{mem} and I_{pred} are instantaneous quantities, i.e. they extend over one time point into the future or the past. The instantaneous nonpredictive information is then defined as the difference between instantaneous memory and predictive power, $I_{\text{mem}} - I_{\text{pred}}$. Given these definitions, Still et al. find that the unnecessary retention of past information is fundamentally equivalent to energetic inefficiency which results in energy dissipation W_{diss}. This dissipation is an upper bound to the difference between memory and predictive information [17]:

$$I_{\text{mem}} - I_{\text{pred}} \le \beta W_{\text{diss}} \tag{9}$$

The inequality relies on a result by Jarzynski who found a mathematically exact relation between the amount of work put into a system and the resulting average free energy change which holds for systems driven arbitrarily far away from the initial equilibrium: $\langle e^{-\beta W} \rangle = e^{-\beta \Delta F}$ [10]. Here, W is the total work performed on the system, the average is taken over an ensemble of measurements, and ΔF is the total change in free energy between the initial and final state. Crooks was able to show that the assumptions of Markovian dynamics and microscopic reversibility lead to the same result [3].

4 Discussion

We have seen that the complexity of a stochastic process can be measured in various ways. While the predictive information E is the number of bits stored by the process itself, the statistical complexity C_μ and the quantum statistical complexity C_q quantify the minimum amount of memory necessary to optimally predict the process. In general all three measures differ for any given stochastic process. While the minimum amount of information required for optimal prediction is E, the actual information stored to make that prediction is usually higher due to the architecture of the storage device, i.e. the discreteness and physical nature of the states. Hence, the physical architecture of the prediction device matters.

A slight change in perspective shed new light on the physical implications of this information processing inefficiency. Instead of a complex system being predicted by a modeler as best she can, Still et al. considered a complex system which stores information about past environmental signals to predict future environmental signals. The assumption is that the system, just like the modeler before, wants to minimise the amount of information stored while predicting as much about the next environmental state as possible. Based on a direct proportionality between free energy and relative entropy they established an equivalence between the non-predictive information which the system holds and the energy dissipation during state changes of the system.

Whether and how energy dissipation caused by inefficient prediction can be measured experimentally is an open question at this point. The equivalence between complexity measures and physical principles yields lower bounds only. Heat dissipation in real complex systems might be orders of magnitude higher. The basic unit of dissipation in a logical operation is $kT \ln 2$ where k is the Boltzmann constant and T is the temperature. The heat dissipated during a logical operation in a modern processing chip, for example, is an estimated $10^{11} kT \ln 2$ at room temperature[2].

However, there exist experiments in this direction. Energy dissipation has been measured in small molecular machines performing logical operations [1]. The measured values were only a few ten times larger than the basic unit of $kT \ln 2$. Furthermore, a free energy principle has been proposed accounting for perception and learning in the brain [7]. This, again, is based on the proportionality between relative entropy and free energy which suggests that the above discussion on information and energy (dissipation) might be highly relevant in this context.

Much work is still to be done to close the gap between rigorous mathematical results in the physics of information and characterisation of complexity. I hope to have conveyed to the reader that some very promising steps have been taken.

Acknowledgements. The Author would like to thank A. Houston for very helpful comments on a draft of this article.

[2] This is an order-of-magnitude estimate based on thermal specifications of an Intel processor.

References

1. Benenson, Y., Adar, R., Paz-Elizur, T., Livneh, Z., Shapiro, E.: DNA molecule provides a computing machine with both data and fuel. Proceedings of the National Academy of Sciences 100(5), 2191–2196 (2003)
2. Cover, T.M., Thomas, J.A.: Elements of Information Theory. WileyBlackwell (2006)
3. Crooks, G.E.: Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. Physical Review E 60(3), 2721–2726 (1999)
4. Crutchfield, J.P., Feldman, D.P.: Regularities unseen, randomness observed: Levels of entropy convergence. Chaos 13(1), 25–54 (2003)
5. Crutchfield, J.P., Young, K.: Inferring statistical complexity. Physical Revue Letters 63(2), 105 (1989)
6. Crutchfield, J.P., Ellison, C.J., Mahoney, J.R.: Time's barbed arrow: Irreversibility, crypticity, and stored information. Physical Review Letters 103(9), 94101–94104 (2009)
7. Friston, K., Kilner, J., Harrison, L.: A free energy principle for the brain. Journal of Physiology-Paris 100(1-3), 70–87 (2006)
8. Grassberger, P.: Toward a quantitative theory of self-generated complexity. International Journal of Theoretical Physics 25(9), 907–938 (1986)
9. Gu, M., Wiesner, K., Rieper, E., Vedral, V.: Quantum mechanics can reduce the complexity of classical models. Nature Communications 3, 762 (2012)
10. Jarzynski, C.: Nonequilibrium equality for free energy differences. Physical Review Letters 78(14), 2690–2693 (1997)
11. Ladyman, J., Lambert, J., Wiesner, K.: What is a complex system? European Journal for Philosophy of Science 3(1), 33–67 (2013)
12. Monras, A., Beige, A., Wiesner, K.: Hidden quantum markov models and non-adaptive read-out of many-body states. Applied Mathematical and Computational Sciences 3, 93 (2011)
13. Shalizi, C.R., Crutchfield, J.P.: Computational mechanics: Pattern and prediction, structure and simplicity. Journal of Statistical Physics 104(3), 817–879 (2001)
14. Still, S.: Information bottleneck approach to predictive inference. Entropy 16(2), 968–989 (2014)
15. Still, S., Crutchfield, J.P.: Structure or noise? arXiv/0708.0654 (2007)
16. Still, S., Crutchfield, J.P., Ellison, C.J.: Optimal causal inference: Estimating stored information and approximating causal architecture. arXiv/0708.1580 (2007)
17. Still, S., Sivak, D.A., Bell, A.J., Crooks, G.E.: Thermodynamics of prediction. Physical Review Letters 109(12), 120604 (2012)
18. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. arXiv/physics/0004057 (2000)
19. Wiesner, K., Gu, M., Rieper, E., Vedral, V.: Information-theoretic lower bound on energy cost of stochastic computation. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science 468(2148), 4058–4066 (2012)
20. Zurek, W.H. (ed.): Information, physics, quantum: The search for links. Santa Fe Institute Studies in the Sciences of Complex Systems, vol. 8. Advanced Book Programme (1990)