Physical systems are often simulated using a stochastic computation where different final states result from identical initial states. Here, we derive the minimum energy cost of simulating a data sequence of a general physical system by stochastic computation. We show that the cost is proportional to the difference between two information-theoretic measures of complexity of the data—the statistical complexity and the predictive information. We derive the difference as the amount of information erased during the computation. Finally, we illustrate the physics of information by implementing the stochastic computation as a Gedanken experiment with a Szilard-type engine. The results create a new link between thermodynamics, information theory and complexity.

Keywords: stochastic processes; entropy thermodynamics; information theory; complexity measures

1. Introduction

The idea of physics as information has a long history. The concept of entropy, at the heart of information theory, originated in the theory of thermodynamics. It was Maxwell and Boltzmann who, at the beginning of the nineteenth century, recognized the intricate link between probability distributions over configurations and thermodynamics. This laid the foundation to the field of statistical mechanics. The similarity between the thermodynamic entropy and the information entropy, introduced in 1948 by Shannon, led to a whole new perspective on physical processes as storing and processing information. It also led to paradoxes such as Maxwell’s demon, which seemed to suggest that work could be generated from heat only with the use of information, which would violate the second law of thermodynamics (see Bennett 1982; Maruyama et al. 2009). The paradox is resolved when one factors in the additional energy required for the user to reset her memory (Landauer 1961; Penrose 1970; Bennett 1982).

*Author for correspondence (k.wiesner@bristol.ac.uk).
With the insight that erasing information generates entropy, Zurek (1989) found limits on the thermodynamic cost of deterministic computation using algorithmic complexity. A deterministic computation generates a unique output, given a particular input. Repeated computations yield identical results. A stochastic computation, on the other hand, yields different outputs for identical inputs. It is a useful descriptor of natural processes that are often stochastic and have different final states, given ‘identical’ initial states (within a given finite resolution or, in quantum mechanics, even infinite resolution). Here, we derive the minimum energy cost of simulating a complex dataset with a stochastic computation. We show that it is proportional to the difference between the statistical complexity (Crutchfield & Young 1989) and the predictive information of the data (Bialek et al. 2001). We derive this difference as the amount of logical irreversibility of the computation. Finally, we illustrate the physics of information by ‘implementing’ the stochastic computation with a Szilard-type engine.

It has been shown that the difference between these two measures arises from an asymmetry in the transport of information forward and backward in time (Crutchfield et al. 2009). In this paper, we give a physical explanation of this asymmetry, together with new mathematical proofs of the relevant information theory.

2. Stochastic computation and measures of complexity

Consider the following model of stochastic computation: a given computational device is in some initial state and outputs length-$N$ strings of symbols $x^N$ according to probability distribution $\Pr(X^N)$. If the distribution $\Pr(X^N)$ is statistically indistinguishable from that of an observed dataset of a physical system, then the symbol sequence $x^N$ is a simulation of that system. Where the probability distribution $\Pr(X^N)$ is a very uncompressed description, one step away from raw data of an experiment, the computational device simulating it is a very compact description, a summary of the regularities, a first step towards a ‘theory’ explaining an experiment. We call the joint probability distribution of past and future observations $\Pr(X_{-N}^{N-1})$ a stochastic process (note that beyond stationarity no further restrictions apply to the process, and it can be finite-order Markovian or non-Markovian). The provably unique minimal (in terms of information stored) and optimal (in terms of prediction) such computation-theoretic representation summarizing the regularities of a stochastic process is a so-called $\epsilon$-machine (Crutchfield & Young 1989; Shalizi & Crutchfield 2001). It consists of an output alphabet $A$, a set of causal states $S$ and stochastic transitions between them. For every pair of states $s, s' \in S$ probabilities $\Pr(S_i = s'|S_{i-1} = s, X_i = x)$ are defined for going from state $s$ to state $s'$ while outputting symbol $x \in A$. The statistical complexity of a process is defined as the Shannon entropy over the stationary distribution of its $\epsilon$-machine’s causal-states,\(^1\)

$$C_\mu := H(S).$$  \hspace{1cm} (2.1)

For definitions of Shannon entropy, conditional entropy and mutual information, see Cover & Thomas (2006).
$C_m$ is the number of bits required to specify a particular causal state in the $\epsilon$-machine. It is the number of bits that need to be stored to optimally predict future data points.

The predictive information of a dataset is given by the mutual information between the two halves; for example, the past data and the future data (Bialek et al. 2001; Shalizi & Crutchfield 2001; Crutchfield & Feldman 2003),

$$E = \lim_{N \to \infty} I[X^{-1}_N; X^{N-1}_0],$$

where $X^{-1}_N$ and $X^{N-1}_0$ are strings of random variables representing observations of a stochastic process. Predictive information is also known under the name of excess entropy, effective measure complexity and stored information (see Crutchfield & Feldman (2003) and references therein). For the following thermodynamic development of stochastic computation, we find the name predictive information most suitable. The predictive information, measured in bits, can be interpreted as the average number of bits a process stores at a given point in time and ‘transmits’ to the future. It is known that

$$C_m = E + H(S_0 | X^\infty_0)$$

and hence that $E \leq C_m$ (Shalizi & Crutchfield 2001, theorem 5). Two important properties of $\epsilon$-machines of relevance here are that the next state, given the last state and the current symbol, is uniquely determined (equation (2.4)) and that the state after the observation of a long enough sequence of symbols is uniquely determined (equation (2.5); Shalizi & Crutchfield 2001),

$$H(S_N | S_{N-1}X_N) = 0 \quad \text{(deterministic-stochastic)}$$

and

$$\lim_{N \to \infty} H(S_N | X^N_0) = 0 \quad \text{(synchronizing).}$$

Successful inference of $\epsilon$-machines ranges from dynamical systems (Crutchfield & Young 1989), spin systems (Crutchfield & Feldman 1997) and crystallographic data (Varn et al. 2002) to molecular dynamics (Li et al. 2008), atmospheric turbulence (Palmer et al. 2000), self-organization (Shalizi et al. 2004) and DNA dynamics (Kelly et al. 2012).

### 3. Thermodynamics of stochastic computation

Landauer (1961) defines an operation to be logically irreversible if the output of the operation does not uniquely define the inputs. In other words, logically irreversible operations forget information about the computational device’s preceding logical state. Landauer’s insight was that irreversible logical transformations cost energy. In the following, we discuss how Landauer’s principle and logical irreversibility apply to stochastic computation and, in particular, to the computation of a stochastic process. For a given $\epsilon$-machine, the current state and the next symbol determine the next state uniquely (equation (2.4)). The reverse, however, is not necessarily true. Given the current state and last output symbol, the previous state is not always uniquely determined. In this case, the $\epsilon$-machine is logically irreversible. Following Landauer’s definition, we define the
irreversibility of a computational step of a given \( \epsilon \)-machine as the entropy of the previous state \((S_{i-1})\), given the current state \((S_i)\) and last output symbol \((X_i)\),

\[
h_{\text{erase}} := H(S_{i-1} \mid X_i S_i).
\]  

(3.1)

For later use, we also define \( h_{\text{erase}}^N := H(S_{i-N} \mid X_{i-N+1}^i S_i) \) for \( i > N \) and \( h_{\text{erase}} := \lim_{N \to \infty} h_{\text{erase}}^N \). \( h_{\text{erase}} \) can be calculated from the \( \epsilon \)-machine directly. It quantifies the average irreversibility of a computational step of the \( \epsilon \)-machine. We now show that strict equality between \( E \) and \( C_\mu \) holds if and only if the \( \epsilon \)-machine is fully logically reversible, i.e. if and only if \( h_{\text{erase}} = 0 \).

**Theorem 3.1.** The predictive information \( E \) of a stochastic process is equal to the statistical complexity \( C_\mu \) if and only if the corresponding \( \epsilon \)-machine is logically reversible,

\[
C_\mu = E \iff h_{\text{erase}} = 0.
\]  

(3.2)

**Proof.\ '\Rightarrow\':\**

From the Markov property of the states \( S_i \) (i.e. \( H(X_1 \mid S_0^i) = H(X_1 \mid S_0) \)) it follows that \( H(S_{i-1} \mid X_0^i S_0) = H(S_{i-1} \mid X_0 S_0) \) (which is called ‘causal shielding’ in the original literature). Using this recursively and the fact that further conditioning never increases the entropy, we obtain the forward direction

\[
C_\mu = E \iff H(S_{i-1} \mid X_0^\infty) = 0 \\
\Rightarrow H(S_{i-1} \mid X_0^\infty S_0) = 0 \\
\Rightarrow H(S_{i-1} \mid X_0 S_0) = 0.
\]  

(3.3)

\('\Leftarrow\':\ Going in reverse, we have \( H(S_{N-1} \mid X_N S_N) = 0 \Rightarrow H(S_{N-1} \mid X_N^N S_N) = 0 \). Using the assumption recursively (in the general form \( H(S_{i-1} \mid X_i S_i) = 0 \)), we get \( H(S_{-1} \mid X_N^N S_N) = 0 \). In addition, we know that

\[
H(S_{-1} \mid X_0^N S_N) = H(S_{-1} \mid X_0^N) - H(S_N \mid X_0^N)
\]  

(3.4)

and hence \( H(S_{-1} \mid X_0^N) = H(S_N \mid X_0^N) \). Taking the limit \( N \to \infty \) on both sides and noting that the r.h.s. goes to zero in this limit (equation (2.5)), the claim follows.\]

Note that this result automatically implies that any \( \epsilon \)-machine with \( h_{\text{erase}} = 0 \) can be inverted in time, turning it into a retrodicter as introduced by Crutchfield et al. (2009) and thus providing an immediate and quick construction of a retrodicter for this case, saving the ‘modeller’ a second computationally costly inference procedure (for more details on computational cost, see Dana (1978)).

For future reference, we call \( H(X_i \mid S_i) := h^R \) the uncertainty of the last symbol given the current state. This reverse entropy rate is different from the reverse entropy rate referred to in dynamical systems theory (Gaspard 2004). The complementary quantity to \( h^R \) is the well-known entropy rate of a stochastic process \( h = H(X_{i+1} \mid S_i) \). We can see that the amount of information that is erased...
per computational step is upper bounded by the amount of information that is created per computational step,

\[ h_{\text{erase}} \leq h, \]  

(3.5)

by writing the joint entropy \( H(S_{i-1}X_iS_i) \) as two different sums,

\[
H(S_{i-1}X_i) + H(S_{i-1}X_i) = H(S_{i-1}X_i) + H(X_iS_i)
\]

\[
\Leftrightarrow H(X_i|S_{i-1}) + H(S_{i-1})
\]

\[
= H(S_{i-1}X_i) + H(X_iS_i) + H(S_i)
\]

\[
\Leftrightarrow h = h_{\text{erase}} + h^R,
\]

(3.6)

where, in the second line, we have used equation (2.4).

Now consider an \( \epsilon \)-machine \( \mathcal{M} \) with \( h_{\text{erase}} > 0 \). To derive the difference between \( E \) and \( C_\mu \), we construct an \( \epsilon \)-machine that outputs more than one symbol at a time as follows. For every pair of states \( s_i, s_j \in S \) of \( \epsilon \)-machine \( \mathcal{M} \), we construct the \( N^{th} \) concatenation \( \mathcal{M}^{\otimes N} \) with state transition probabilities \( \Pr(s_j|s_i,x^N) = \sum_{s_1} \Pr(s_1|x^N) \prod_{k=2}^{N} \Pr(s_k|s_{k-1}x_k) \cdot \Pr(s_1|x_1) \) upon outputting \( x^N \), where the sum runs over all state sequences \( s_1^{N-1} \) of length \( N - 1 \). Note that \( \mathcal{M} \) and \( \mathcal{M}^{\otimes N} \) have the same set of states and the same probability distribution over output strings \( \Pr(X^N) \)—so they have the same \( E \) and \( C_\mu \).

**Theorem 3.2.** The difference between the statistical complexity \( C_\mu \) and the predictive information \( E \) of a stochastic process approaches the logical irreversibility of the corresponding concatenated \( \epsilon \)-machine \( \mathcal{M}^{\otimes N} \) as \( N \to \infty \),

\[
\lim_{N \to \infty} h_{\text{erase}}^N = C_\mu - E.
\]

**Proof.** By rewriting \( H(S_{-1}X_0^NS_N) \) in two different ways, we obtain the information-theoretic equality,

\[
H(S_{-1}X_0^N) = H(S_{-1}X_0^NS_N) + H(S_N|X_0^N) - H(S_N|S_{-1}X_0^N).
\]

(3.7)

The last term is zero owing to determinism (equation (2.4)); the second term goes to zero for \( N \to \infty \). Hence, by taking the limit, we obtain

\[
H(S_{-1}X_0^N) \to H(S_{-1}X_0^NS_N) \quad \text{as} \quad N \to \infty.
\]

(3.8)

The term on the right-hand side is exactly \( h_{\text{erase}} \) of \( \mathcal{M}^{\otimes N} \). Letting \( N \to \infty \) the claim follows.

Using theorem 3.2, we can derive an information-theoretic efficiency of simulating a stochastic process and from that the minimum energy cost. Figure 1 schematically illustrates an \( \epsilon \)-machine contained in a box that, on consecutive time steps, outputs symbols visible to an outside observer. The inside of the box, and hence its entropic content, is fully characterized by two random variables: the output symbol \( X \) and the state of the machine \( S \). The computational steps are as follows. In figure 1a, the \( \epsilon \)-machine in the box is in state \( S_{i-1} \). Going to figure 1b, it generates symbol \( X_i \), according to the probability distribution \( \Pr(X_i|S_{i-1}) \), leading to an increase in entropy inside the box by \( h = H(X_i|S_{i-1}) \). Going to figure 1c the \( \epsilon \)-machine moves from state \( S_{i-1} \) to state \( S_i \). A resetting of the previous-state information causes a decrease in entropy inside the box.
by $H(S_{i-1}|S_iX_i) = h_{\text{erase}}$. Finally, in figure 1c, the symbol is ejected into the environment that decreases the entropy inside the box again, this time by $H(X_i|S_i) = h^R$. This closes one cycle of computation. Because $\epsilon$-machines are stationary and thus $H(S_i) = H(S_{i-1})$, the entropy contributions during one closed cycle must add up to zero and we obtain $h - h_{\text{erase}} - h^R = 0$, which is exactly equation (3.6).

We now modify this stochastic computation by allowing for the generation of $N + 1$ symbols at a time. The $\epsilon$-machine inside the box (figure 1) is replaced by the concatenated $\epsilon$-machine $\mathcal{M}^\otimes(N+1)$. In figure 1a, this machine starts out in state $S_{i-1}$; going to figure 1b it generates $N + 1$ symbols according to $\Pr(X_i^{i+N}|S_{i-1})$. This causes an increase in entropy inside the box by $H(X_i^{i+N}|S_{i-1})$. Going to figure 1c the concatenated $\epsilon$-machine moves to state $S_{i+N}$ erasing the previous-state information that decreases the entropy inside the box by $H(S_{i-1}|X_i^{i+N}S_{i+N})$. Ejecting the symbol sequence into the environment the entropy of the box decreases by $H(X_i^{i+N}|S_{i+N})$. Setting without loss of generality $i = 0$, we obtain for the entropy balance of one computational cycle,

$$H(X_0^N|S_{-1}) - H(S_{-1}|X_0^NS_N) - H(X_0^N|S_N) = 0. \tag{3.9}$$

The l.h.s. can be rewritten as

$$H(S_{-1}) - H(S_{-1}|X_0^NS_N) - I(S_{-1};X_0^N). \tag{3.10}$$

Letting $N \to \infty$, we obtain

$$H_{\text{erase}} + E - C_\mu = 0. \tag{3.11}$$

Hence, the entropy balance of one cycle of stochastic computation is in the limit of an infinite string of output given by equation (2.3) and theorem 3.2. We now discuss the thermodynamics of such a stochastic computation by first considering an extension of Szilard’s Gedanken experiment of a single-particle engine.
Szilard (1929) invented a Gedanken experiment of a single-particle engine to resolve Maxwell’s demon paradox that seemed to defy the second law of thermodynamics. The engine consists of a particle in a box, a measurement device locating the particle in either half of the box and a memory to store the measurement result. Szilard considered the following procedure for extracting work from the particle’s thermal motion. A user measures the particle’s position and stores this one bit of information in the memory. Subsequently, she ‘compresses’ the box to the half that contains the particle. This does not require any work. The thermal motion of the particle ‘decompresses’ the box again and hence lets the user extract work from the box without any cost. This apparent paradox was resolved by Landauer (1961), who considered storing the measurement outcome as cost. Later, Penrose (1970) and Bennett (1982) independently noticed that the source of entropy increase was not storing the measurement result but resetting2 the memory after the work extraction. Hence, the full cycle converts one bit (log 2) of information into $kT \log 2$ units of work and subsequently generating $kT \log 2$ of heat when resetting the memory.

Now, we extend Szilard’s Gedanken experiment by considering a particle whose position is governed by random variables $X_0, X_1, \ldots$ which we can indirectly measure only by recording appropriate information from correlated random variables $X_{-1}, X_{-2}, \ldots$. This is analogous to simulating a stochastic process by generating information about its future based on observations of the past. The recorded bits $X_{-1}, X_{-2}, \ldots$ can then be translated to information about $X_0, X_1, \ldots$ by the use of an appropriate simulator and henceforth converted into work. To extract the maximum possible amount of information, $E$, the theory of $\varepsilon$-machines dictates that we must record at least $C_\mu$ bits. The minimality and optimality of $\varepsilon$-machines ensures that any fewer bits would render the simulation sub-optimal. We therefore define the information-theoretic efficiency of computing a stochastic process as

$$i := \frac{E}{C_\mu} = 1 - \frac{H_{\text{erase}}}{C_\mu}. \quad (3.12)$$

The information-theoretic efficiency of a stochastic computation then determines its thermodynamic cost as $kT H_{\text{erase}}$. Note that, in the original discussions on Szilard’s engine, the information-theoretic efficiency is always one (one bit of memory, log 2, is converted into one unit of work, $kT \log 2$).

$E/C_\mu$ has been named the ‘predictive efficiency of a process as the fraction of the information it contains which actually effects [sic] the future’ (Shalizi & Moore 2003). Our results supply this concept with physicality and mathematical rigour.

4. Conclusion

We have derived the minimum energy cost of simulating a physical system as the difference between two information-theoretic complexity measures of the data. Of the two measures, the predictive information measures the amount of information stored about a process’s past transmitted to the future, and the

2Landauer’s principle is usually stated in terms of ‘erasure’. However, strictly speaking, ‘reset’ is the more appropriate term for the logically irreversible operation of resetting the memory to a default state as opposed to randomizing it—see, for example, the discussion by Ladyman et al. (2007).
statistical complexity measures the amount of information required to compute this future. Any difference between the two is given by the logical irreversibility of the computation and hence represents the minimum energy cost of physically running a simulation.

This result is complementary to the discussion of Crutchfield et al. (2009), who derive the difference between the two measures from the asymmetry of running the process in forward and reverse. Our results allow for the physical interpretation of energy cost of a stochastic computation. This reveals an intricate relation between thermodynamics, information-processing and complexity measures. They motivate the use of information-theoretic tools for studying the physics of complex systems. Furthermore, recent results that quantum simulators require less information storage than classical ones could indicate that quantum information leads to a reduced energy cost for stochastic computation (Gu et al. 2012).

K.W. acknowledges funding through EPSRC (grant no. EP/E501214/1).

References


