

Article

Information Bottleneck Approach to Predictive Inference

Susanne Still

Information and Computer Sciences Department, University of Hawaii at Mānoa, Honolulu, HI 96822, USA; E-Mail: ssill@hawaii.edu; Tel.: +1 808 956-5816; Fax: +1 808 956-3548

Received: 3 June 2013 / Accepted: 18 June 2013 / Published: 17 February 2014

Abstract: This paper synthesizes a recent line of work on automated predictive model making inspired by Rate-Distortion theory, in particular by the Information Bottleneck method. Predictive inference is interpreted as a strategy for efficient communication. The relationship to thermodynamic efficiency is discussed. The overall aim of this paper is to explain how this information theoretic approach provides an intuitive, overarching framework for predictive inference.

Keywords: predictive inference; information bottleneck method; dynamical systems; thermodynamic efficiency; far-from-equilibrium thermodynamics; computing engines

1. Introduction

“The fundamental problem of scientific progress, and a fundamental one of every day life, is that of learning from experience. Knowledge obtained in this way is partly merely description of what we have already observed, but partly consists of making inferences from past experience to predict future experience. This part may be called generalization. It is the most important part; events that are merely described and have no apparent relation to others may as well be forgotten, and in fact usually are.”

Sir Harold Jeffreys [1].

Predictive inference lies at the heart of science, as one of our fundamental cognitive tools for discovery-making. The general idea is that a good model should have predictive power, while not being overly complicated. This notion is deeply rooted in our culture, and has been made specific in a variety of different technical approaches in modern times, e.g., [1–5]. Complexity measures abound [6], as do utility functions that prescribe which aspects of the data are considered relevant. In fact,

generalization is a core issue in statistics and machine learning [3]. There, one often analyzes a “bag of data”. Time stamps are either not important or simply not available, as data may stem from experiments designed to reveal a relationship between independent and dependent variables. However, one of the most fundamental assumptions we typically make about physical reality is that of the progression of time, and a causal structure of spacetime (see e.g. [7,8]). Natural, and in particular living systems, usually are thought of as dynamical systems embedded in an environment in which they interact with other dynamical systems. Observations of dynamical systems typically yield time series data, containing the information necessary to understand the underlying physics of the observed system [9,10]. This paper thus focusses on the analysis of time series data.

Setting aside intricate philosophical arguments, let us take a simple information theoretic view that makes minimal assumptions about the process underlying the generation of the data. Motivated by the fact that any model constructed from the data will produce, in some form, a *summary* of past observations, let us try to make that summary efficient in the sense that it shall contain, to the greatest extent possible, information that can be used to predict future data without keeping irrelevant detail that would lead to unnecessary model complexity.

Formulated this way, it is clear that the Information Bottleneck (IB) [11] method provides an excellent framework for predictive inference, because it is a lossy compression scheme that finds that summary of the data which contains maximal relevant information at a fixed level of allowed detail. The data to be compressed, or summarized, are past experiences, and the summary should be useful for predicting future experiences. We can thus identify relevant information as information about future data [12–20]. The use of a model that is not overly complex is related to the goal of compression in communication. How much detail the model contains is measured by the information that is kept about the past, *i.e.*, the memory contained in the summary. Irrelevant, nonpredictive information has to be discarded.

The IB method then allows one to find, for any given amount of retained memory, that representation that has the largest predictive power by means of containing the largest amount of information about future experiences or, conversely, the most compact model with a desired level of predictive power.

Direct application of the IB method (Section 2) compresses past trajectories of a given length to predict future trajectories of a given length. While limited, this approach already has interesting applications and is related to other data analysis methods, such as to [16] the causal state partition [21] and, under restrictive assumptions [18], to slow feature analysis [22].

Going beyond the original IB framework, a dynamical learning paradigm was established in [14] (Section 4). Using this approach, feedback from the learner to the environment can be treated naturally, and connections to reinforcement learning can be made [17] (Section 4.1). The scope of the present paper is, however, limited to passive predictive inference, without feedback (Section 4.2). The ϵ -machine [21], known as an optimal predictor [23], is a limiting case of this more general approach (Section 4.3).

The dynamical approach taken in [14] allows for a thermodynamic treatment of computation and predictive inference away from thermodynamic equilibrium [24] (Section 3). Results may point towards a *physical* reason for the philosophy underlying predictive inference.

1.1. Information Theoretic Treatment of Prediction

The dynamics of complex systems can be viewed as transforming, deleting, and creating information, *i.e.* they are performing computations. A body of work has argued this, and has made use of predictive information as a general measure for temporal correlations (e.g. [20,21,25–32], and refs. therein), since it gives a fuller description than the correlation function [33].

Let the value of a signal obtained at time, t , be denoted by x_t . In general, this could be a concatenation of several measured quantities at time, t . Then the (Shannon) mutual information [34] $I[x_t; x_{t'}] := \left\langle \log \left[\frac{p(x_t, x_{t'})}{p(x_t)p(x_{t'})} \right] \right\rangle_{p(x_t, x_{t'})}$ measures how much an observation at one instant can tell us about an observation at another instant in time [25]. (Throughout the paper, information is measured in nats, for convenience, and \log denotes the natural logarithm. Brackets, $\langle \cdot \rangle_{p(\cdot)}$, denote averages, with subscripts denoting the probability distribution, p , over which the average is taken [35].) Now, let $t' = t + \Delta t$, where Δt is a finite interval, discretizing time. For simplicity of the exposition, let us rescale, such that $\Delta t = 1$. The instantaneous predictive information, $I[x_t; x_{t+1}]$, suffices to describe the complete causal connections in the observed time series, only in cases where x_t contains a complete measurement of a dynamical system, such that the dynamics are first-order (either Hamiltonian or governed by a first-order Markov process) [26]. In the general case, if x_t does not fully determine the underlying state of the observed dynamical system, then information between the extended past, $\overleftarrow{x}_t = (x_{t-\tau_P}, \dots, x_t)$, and future, $\overrightarrow{x}_t = (x_{t+\Delta t}, \dots, x_{\tau_F})$, trajectories should be considered (τ_P and τ_F parameterize the length of past and future trajectories, respectively, and here we set $\Delta t = 1$):

$$I[\overleftarrow{x}_t; \overrightarrow{x}_t] := \left\langle \log \left[\frac{p(\overleftarrow{x}_t, \overrightarrow{x}_t)}{p(\overleftarrow{x}_t)p(\overrightarrow{x}_t)} \right] \right\rangle_{p(\overleftarrow{x}_t, \overrightarrow{x}_t)} \tag{1}$$

In the limit, $\tau_F \rightarrow \infty$, the growth of predictive information as a function of τ_P reveals the complexity of the underlying physical process [29]. It has been argued that predictive information reveals the causal structure of the physical process that generated the observed signal [21,26,28,36]. This approach has been useful in dynamical systems theory and chaos theory, as well as neuroscience and machine learning ([14,16,17,20,21,23,28,29,36,37] and refs. therein). Recently, predictive information was proposed as “a universal order parameter to study phase transitions in physical systems” [38].

The information theoretic approach to predictive inference taken in the present paper is based on a simple extension of predictive information (Eq. 1) to include a (learning) system that receives the signal, x , as an input. At time, t , the state of the system, s_t , contains memory about the past trajectory of the environment, \overleftarrow{x}_t , quantified by the mutual information:

$$I[s_t; \overleftarrow{x}_t] := \left\langle \log \left[\frac{p(s_t | \overleftarrow{x}_t)}{p(s_t)} \right] \right\rangle_{p(s_t | \overleftarrow{x}_t)p(\overleftarrow{x}_t)} \tag{2}$$

If the external signal has a causal structure and temporal correlations, that is, if the predictive information, $I[\overleftarrow{x}_t; \overrightarrow{x}_t]$, is nonzero, then part of this memory may be information about the future:

$$I[s_t; \overrightarrow{x}_t] := \left\langle \log \left[\frac{p(\overrightarrow{x}_t | s_t)}{p(\overrightarrow{x}_t)} \right] \right\rangle_{p(s_t | \overrightarrow{x}_t)p(\overrightarrow{x}_t)} \tag{3}$$

If the probabilistic map that determines the state s_t from the input data can be chosen freely, then there is the potential for predictive filtering. The model then contains predictions encoded in the distributions, $p(\vec{x}_t|s_t)$, and the predictive bits retained (Equation 3) quantify the predictive power of the model.

This paper discusses different formalizations of learning systems. Direct application of the IB method (Section 2) results in the construction of a map from past trajectories to system states, $p(s_t|\overleftarrow{x}_t)$. Alternatively, the learning system can be viewed as a dynamical system itself (Section 4), coupled to the environment. Then, a recursive information bottleneck algorithm (RIB) can be used to find optimally predictive dynamics (Section 4.2, 4.3). When the physical reality of the system is taken into account, energetic considerations factor in (Section 3). Finally, there can be feedback from the learning system to the environment, endowing the system with some level of control (Section 4.1). This approach describes interactive learning. It goes beyond passive predictive inference (and therefore beyond the scope of the present paper) and extends the IB framework to a feedback situation.

The material in this paper is a synthesis of results that have been presented at conferences, published elsewhere [14–17,24], and some new material. Section 2 makes use of material from [15,16], Section 3 uses material from [24], and Section 4 uses material from [14,17].

2. Direct Use of Information Bottleneck for Predictive Inference

Compression of past trajectories results in a model that contains a map from \overleftarrow{x}_t to the summary variable, s_t , which we can think of as denoting the state of a learning machine. Another function maps from s_t to future trajectories. Both of these functions can, in general, be probabilistic maps. A model therefore contains the probability distributions $p(s_t|\overleftarrow{x}_t)$ and $p(\vec{x}_t|s_t)$. In theory, past and future trajectories could be infinite.

Let us measure the model's complexity by the amount of information that the state contains about the data it summarizes, that is, the mutual information [34], $I[s_t; \overleftarrow{x}_t] = H[s_t] - H[s_t|\overleftarrow{x}_t]$ [35]; or in other words, the model's *memory*. For deterministic maps, we have $H[s_t|\overleftarrow{x}_t] = 0$, and the memory thus reduces to the entropy, $H[s_t]$, which reflects, for predictive systems, the statistical complexity [28]. However, in the general case of probabilistic maps, $I[s_t; \overleftarrow{x}_t]$ is a more adequate measure of model complexity than $H[s_t]$. To see this, consider a model with a large number of states, n , but where each data point is mapped with equal probability to each of the n states [15]: $p(s_t|\overleftarrow{x}_t) = 1/n$. The probability of states is then $p(s_t) = \langle p(s_t|\overleftarrow{x}_t) \rangle_{p(\overleftarrow{x}_t)} = 1/n$, independent of the exact form of $p(\overleftarrow{x}_t)$. The entropy of this model is large, $H[s_t] = \log(n)$. This is misleading, because the model is actually not complex at all—it would have the same capacity if all n states were replaced by one effective state. The mutual information captures this, because $H[s_t|\overleftarrow{x}_t] = \log(n)$, and hence, $I[s_t; \overleftarrow{x}_t] = 0$.

The predictive power of the model is given by the information that the model captures about future experiences, \vec{x}_t . One then looks for an assignment of pasts to states that results in a model with maximal predictive power at fixed memory. Formally, this is done by solving the constrained optimization problem [11,16]:

$$\max_{p(s_t|\overleftarrow{x}_t)} \left(I[s_t; \vec{x}_t] - \lambda I[s_t; \overleftarrow{x}_t] \right) \quad (4)$$

where λ is a Lagrange multiplier controlling the trade-off between model complexity and predictive power [11,15,16]. Importantly, when the past is known, then the probability of the future does not depend

on knowledge of the state, *i.e.* $p(\vec{x}_t|\overleftarrow{x}_t, s_t) = p(\vec{x}_t|\overleftarrow{x}_t)$. This optimization problem is thus identical to the information bottleneck problem: past trajectories, \overleftarrow{x}_t , are compressed, such that information about the future, \vec{x}_t , is kept, *i.e.*, predictive information is relevant information. The optimization is, of course, equivalent to Shannon’s rate-distortion theory [34], with the the relative entropy, or Kullback-Leibler divergence [39], $D_{KL}[p(\vec{x}_t|\overleftarrow{x}_t)||p(\vec{x}_t|s_t)] = \left\langle \log \left[\frac{p(\vec{x}_t|\overleftarrow{x}_t)}{p(\vec{x}_t|s_t)} \right] \right\rangle_{p(\vec{x}_t|\overleftarrow{x}_t)}$, used as a distortion function [15] measuring the predictability loss one encounters in replacing the trajectory, \overleftarrow{x}_t , by the summary, s_t . For each value of λ , this optimization results in an optimal probabilistic assignment of past trajectories to model states, *i.e.*, IB finds a *family* of optimal models [11], parameterized by λ . Each obeys the self consistent equations [11,16]:

$$p(s_t|\overleftarrow{x}_t) = \frac{p(s_t)}{Z(\overleftarrow{x}_t)} \exp \left(-\frac{1}{\lambda} D_{KL}[p(\vec{x}_t|\overleftarrow{x}_t)||p(\vec{x}_t|s_t)] \right) \tag{5}$$

$$p(\vec{x}_t|s_t) = \sum_{\overleftarrow{x}_t} p(\vec{x}_t, \overleftarrow{x}_t) \frac{p(s_t|\overleftarrow{x}_t)}{p(s_t)} \tag{6}$$

$$p(s_t) = \sum_{\overleftarrow{x}_t} p(s_t|\overleftarrow{x}_t)p(\overleftarrow{x}_t) \tag{7}$$

where $Z(\overleftarrow{x}_t)$ ensures normalization:

$$Z(\overleftarrow{x}_t) = \sum_{s_t} p(s_t) \exp \left(-\frac{1}{\lambda} D_{KL}[p(\vec{x}_t|\overleftarrow{x}_t)||p(\vec{x}_t|s_t)] \right) \tag{8}$$

The exponential distribution in Equation (5) can be compared to a Gibbs-Boltzmann distribution. With this analogy, the Lagrange parameter, λ , has been identified as a “temperature”-like parameter (not to be confused with physical temperature, *cmp.* Sec. 3.4). The analogy leads to the intuition that in the limit of large λ , fluctuations prevent any structure from being resolved. As λ is lowered, more and more structure is recovered as solutions pass through a series of phase transitions [11].

Note that the model’s prediction is represented by the distribution, $p(\vec{x}_t|s_t)$. The model is constructed from the input, $p(\vec{x}_t, \overleftarrow{x}_t)$. In practice, this is usually not known, but has to be estimated from the given data, *e.g.* via normalized frequencies [16], or model assumptions have to be made (*e.g.* [18]). However, once the model, Equations (5-8), has been constructed, then any trajectory \overleftarrow{x}_t can be mapped onto a state s_t via the probabilistic map $p(s_t|\overleftarrow{x}_t)$. From knowledge of the state, an estimate of the future can be generated using the distribution $p(\vec{x}_t|s_t)$.

The probability of the state sequence $(\dots, s_{t-1}, s_t, s_{t+1}, \dots)$ given the input data (past trajectories) is proportional to (Equations 5-8) $\exp \left[-\sum_t \left(\frac{1}{\lambda} D_{KL}[p(\vec{x}_t|\overleftarrow{x}_t)||p(\vec{x}_t|s_t)] - \log[p(s_t)] \right) \right]$. The first term in the sum results in the most likely state sequence having minimal predictability loss. This term becomes increasingly relevant as λ decreases. The second term, $\log[p(s_t)]$, corresponds to an extra penalty, favoring simpler models over more complex ones. This explains why, out of all possible, equally predictive representations [21,23,28,36], the minimal one is chosen [16].

To study the causal compressibility [15] of a signal, $x(t)$, one can then plot the predictive power of the best possible model *vs.* its memory, that is, both quantities are evaluated at the solution to the optimization at different values of the Lagrange multiplier to obtain an information curve [11] in analogy

to a rate-distortion curve. Processes of qualitatively different causal structure can be thus identified [15]. Numerically, the curve can be traced out using the information bottleneck algorithm [11], which is closely related to the Blahut-Arimoto algorithm [40,41].

2.1. Asymptotic Behavior

The full power of this method is revealed by studying the solution in the regime in which the trade-off parameter, $\lambda \rightarrow 0$, so that the emphasis is on predictive power. When infinitely long pasts are used to predict infinite futures, then the states computed in this limit are the *causal states* constituting minimal sufficient statistics [21,23,28,36]. This result holds independent of the exact form of the distribution that generated the input data. A detailed proof can be found in [16]. Here, we give only an intuitive derivation. Note that similar intuition has previously been pointed out in [42].

Using results from [43], it is easy to see that in the limit $\lambda \rightarrow 0$, $p(s_t|\overleftarrow{x}_t)$ will tend towards zero for all states, except those that minimize $D_{KL}[p(\overrightarrow{x}_t|\overleftarrow{x}_t)||p(\overrightarrow{x}_t|s_t)]$. Assuming that there is no restriction on the state space of the learning machine, one can always ensure that there exists one state, s^t , such that $D_{KL}[p(\overrightarrow{x}_t|\overleftarrow{x}_t)||p(\overrightarrow{x}_t|s^t)] = 0$. By this argument, the conditional probability, $p(s_t|\overleftarrow{x}_t)$, equals one if $s_t = s^t$ and zero otherwise. This constitutes a deterministic assignment of pasts to states. A method known as *deterministic annealing* [44,45] can be used to numerically find the solution for the $\lambda \rightarrow 0$ limit.

The distributions specified by Equation (5) assign past trajectories to model states by means of distributional clustering [11,46]. In the context of time series data, this means that two pasts, \overleftarrow{x}_t and \overleftarrow{x}'_t , that have similar conditional future distributions, $p(\overrightarrow{x}_t|\overleftarrow{x}_t)$ and $p(\overrightarrow{x}_t|\overleftarrow{x}'_t)$, will likely end up in the same *cluster*, denoted by s_t . This results in a partitioning of the space of all past trajectories. In general, this partition is what is often called *soft*, or *fuzzy*, because the assignments are probabilistic. The resulting partition can only be *hard* when the assignments become deterministic.

The hard partition discovered in the $\lambda \rightarrow 0$ limit can alternatively be described by an equivalence relation, yielding the very definition of the causal state partition [21]: two past trajectories, \overleftarrow{x}_t and \overleftarrow{x}'_t , are equivalent for purposes of prediction, if and only if $p(\overrightarrow{x}_t|\overleftarrow{x}_t) = p(\overrightarrow{x}_t|\overleftarrow{x}'_t)$. They are then mapped onto the same *causal state*. This equivalence relation has many desirable properties, most notably, the causal states are unique and minimal sufficient statistics [21,23,28,36,37]. The causal state partition is a probabilistic bisimulation [47], and is also fundamentally related to observable operator models [48], see e.g. [49].

This result shows that IB has the capacity for predictive inference when used on time series data, because it discovers minimal sufficient statistics. This fact has motivated the use of the name *optimal causal inference* (OCI) [16]. For finite lengths, τ_P and τ_F , an equivalence relation can be defined in analogy to the causal state partition, and OCI recovers the corresponding partition, in the limit $\lambda \rightarrow 0$. Other algorithms exist for constructing the causal state partition [31,50]. One advantage of the IB approach is that it allows for a principled relaxation of the complexity constraint by adjustment of λ .

Here, the words causality and causal inference do not refer to what has been coined causal inference in statistics (see e.g. [51] and references therein), an approach that involves the logic of counterfactuals. In statistics it often makes sense that abstracting away from given data to as-yet unseen data does not

necessarily rely on the data being ordered in time. But the notion of causality cannot easily be separated from the concept of time in any physically meaningful way. Therefore, temporal causal structure has to be taken into account to understand the physics underlying natural computation, particularly if one assumes that nature computes by means of its dynamics (see Sections 1.1 and 3.4).

2.2. Advantages and Disadvantages

This approach has the great advantage that no assumptions have to be made about the distribution underlying the generation of the time series. OCI finds the best model in terms of predictive power at any desired level of complexity.

While conceptually elegant, in practice $p(\overleftarrow{x}_t, \overrightarrow{x}_t)$ has to be estimated from the data. Finite sample effects set a natural lower bound on λ , due to an upward bias in the estimated information content (see [43] and references therein). Therefore, the number of causal states that can be used to describe the data without over-fitting is limited. The bias correction method of [43] has been used in the predictive inference context [16]. In general, it works well only when the number of data is significantly larger than the number of bins that are used to estimate probabilities (by normalized counts).

Since \overleftarrow{x}_t and \overrightarrow{x}_t are potentially infinitely long sequences, finite sample effects could easily be overwhelming. The first simple step to address this problem is to make τ_P and τ_F finite and not too large. This restriction, however, comes at the disadvantage of restricting the predictive power of the model and its ability to discover structure. It also introduces a new problem: the discovered model might depend on how we chose to distribute the total length, $\tau_P + \tau_F$, to the parameters, τ_P and τ_F , respectively. The recursive IB we discuss in Section 4.2 addresses this problem.

2.3. Linear and Gaussian Models

An alternative is to use a model for the process that generated the data. For linear models with Gaussian noise, the explicit form of the IB solution can be calculated analytically [52]. This is known as Gaussian information bottleneck (GIB). Applied to time series, for $\tau_P = \tau_F = 1$, one assumes that past, x_t , and future, x_{t+1} , are jointly multivariate Gaussian variables [18]. GIB furthermore assumes that there is a *linear* transformation, the matrix, M , mapping the input data to the model variable, here, $s_t = Mx_t + \xi$, in a noisy fashion, where $\xi \simeq \mathcal{N}(0, \Sigma)$ is normally distributed. In this approach, GIB is used to find a reduced description of the underlying model. The optimization is now over the linear transformation, M , and the bottleneck problem becomes [18]:

$$\max_M (I[s_t, x_{t+1}] - \lambda I[s_t; x_t]) \quad (9)$$

The matrix that solves this optimization problem contains, as λ decreases, an increasing number of (scaled) eigenvectors of $[I - (\Sigma_{x_t; x_{t+1}} \Sigma_{x_t}^{-1})^2]$ [18,52], where Σ_{x_t} denotes the covariance matrix of the inputs, $\Sigma_{x_t; x_{t+1}}$ characterizes temporal correlations, and I is the identity matrix. The method is related [18] to slow feature analysis [22], and is able to deconvolve and filter composite signals [18].

When analyzing nonlinear complex systems in which the Gaussian assumption is violated, the practitioner has to consider carefully whether the method can still be applied.

3. Thermodynamic Foundations

Predictive inference is an efficient way of processing information, implemented by living organisms in various ways. Prediction can be useful for different stages of information processing, ranging from genetic networks, to vision, to motor behavior, to higher cognitive function [20]. All information processing has to happen on physical devices, natural or synthetic, and many authors have argued that, in general, all information is physical (e.g., [53]). Ultimately, one would like to know if there are physical reasons for the emergence of predictive inference.

Whenever many small (computing) units are tightly packed together, as is the case in living systems, heat generation due to dissipation generally poses a problem. Thermodynamic efficiency is thus a relevant consideration, and is also becoming increasingly relevant for the design of modern artificial systems, as the size of components shrinks and their packing density increases. This section asks about the relationship between thermodynamic efficiency and information processing efficiency (in the sense implemented by predictive inference). Could energetic efficiency be an underlying motivation for predictive inference?

Memory and predictive power are the relevant diagnostics for how well a system is implementing predictive inference: on the one hand, a model should have large predictive power, on the other hand, it should not be overly complex, *i.e.*, we do not wish to retain memory beyond what is useful for prediction. The nonpredictive part of the memory thus measures the inefficiency of the model, in terms of prediction, by quantifying how much irrelevant information is retained.

It turns out that there is a simple relationship between dissipation and instantaneous nonpredictive information [24]. Before this is explained in Section 3.4, a brief overview of some relevant context is given in Sections 3.1 to 3.3.

Let us imagine that predictive inference is implemented by means of a physical device, and, as before, let us denote the state of this computing system, at time, t , by s_t . The input time series, $\mathbf{X} = x_0, x_1, \dots, x_\tau$, is fed into the system by means of a change in some external parameter(s). For simplicity, we shall denote these environmental variables also by x_t (assuming that a one-to-one mapping can be constructed between the input time series and the external parameters driving the computing system). Changes in the external signal then cause changes in the system's state.

In the previous section, the dependency of states on data was given abstractly by the probabilistic map, $p(s_t | \overleftarrow{x}_t)$, from *all* past experiences up to time, t , onto states s_t . However, specifying this map in an explicit way in practice would require a buffer for the entire history, in order to determine the state s_t directly from the currently observed trajectory.

An alternative strategy is to use the system's state space as memory by making the state-update depend not only on the input data, but also on the system's previous state. The dynamics of the device can be characterized by conditionally Markovian transition probabilities, $p(s_t | s_{t-1}, x_t)$. These dynamics result in an implicit model of the input time series, and also determine the physics of the computing device, as the incoming data drive the computing machine via changes in control parameters. During this process, work is done. The thermodynamic inefficiency of this process can be characterized by the amount of work that is lost to dissipation. Since this process may drive the computing system arbitrarily far from thermodynamic equilibrium, equilibrium thermodynamics is no longer an adequate description.

3.1. Driven Systems Far from Thermodynamic Equilibrium

During the last two decades, significant progress has been made in understanding driven systems far from equilibrium [54,55], most notably, Jarzynski’s work relation [56],

$$\Delta F = -\beta^{-1} \log \langle e^{-\beta W} \rangle, \tag{10}$$

associating the work, W , done on a thermodynamic system to the resulting change in free energy of the system, ΔF . The brackets, $\langle \cdot \rangle$, denote the statistical average. It is assumed that the system is started in thermodynamic equilibrium, then is driven arbitrarily far from equilibrium, due to changes in external parameters that follow a known protocol (the experiment). After execution of this protocol, the system is finally allowed to relax back to thermodynamic equilibrium. Throughout, the system is connected to a heat bath at inverse temperature $\beta = 1/k_B T$, where k_B is Boltzmann’s constant, and T is the temperature. The work relation holds for systems driven arbitrarily far from thermodynamic equilibrium, thus going beyond the near-equilibrium predictions of linear response theory [57].

The fact that the work done on the system, on average, has to be larger than the work that can be derived from the system, $\langle W \rangle \geq \Delta F$, follows from Equation (10) [56] via Jensen’s inequality. The r.h.s. of Equation (10) can be expanded into a sum of cumulants of W . Assuming a Gaussian distribution, only the first two survive and a fluctuation-dissipation relation follows [56]: $\frac{\beta}{2} (\langle W^2 \rangle - \langle W \rangle^2) = \langle W \rangle - \Delta F$.

While the original derivation was done using Hamiltonian dynamics, the work relation can also be derived for stochastic systems governed by conditionally Markovian dynamics [58]. State-to-state transitions are given by $p(s_t|s_{t-1}, x_t)$, taking the system through a sequence of states $\mathbf{S} = s_0, \dots, s_\tau$ in response to the input, or protocol, $\mathbf{X} = x_0, \dots, x_\tau$ which is assumed to be given. The system is in thermodynamic equilibrium at $t = 0$, so that the Boltzmann distribution, $p_{\text{eq}}(s|x) := e^{-\beta(E(s,x)-F[x])}$ describes the initial distribution, $p(s_0|x_0) = p_{\text{eq}}(s_0|x_0)$ (subscript “eq” denotes thermodynamic equilibrium). The equilibrium free energy is $F[x] = \langle E(s, x) \rangle_{p_{\text{eq}}(s|x)} - k_B T H[p_{\text{eq}}(s|x)]$, where $k_B H[p_{\text{eq}}(s|x)]$ is the thermodynamic entropy, given by [59] the Shannon entropy of the equilibrium distribution, $H[p_{\text{eq}}(s|x)] := -\langle \log[p_{\text{eq}}(s|x)] \rangle_{p_{\text{eq}}(s|x)}$, times the Boltzmann constant k_B .

The total work done, W , can be split into incremental changes in energy [58,60], as can the total heat, Q , exchanged with the bath (heat flowing into the system is positive by convention):

$$W = \sum_{t=0}^{\tau-1} W[x_t \rightarrow x_{t+1}; s_t] := \sum_{t=0}^{\tau-1} (E(s_t, x_{t+1}) - E(s_t, x_t)) \tag{11}$$

$$Q = \sum_{t=0}^{\tau-1} Q[s_t \rightarrow s_{t+1}; x_{t+1}] := \sum_{t=0}^{\tau-1} (E(s_{t+1}, x_{t+1}) - E(s_t, x_{t+1})) \tag{12}$$

Energy changes during these work- and relaxation-steps sum up to the total change in energy, $W + Q = \Delta E = E(s_\tau, x_\tau) - E(s_0, x_0)$ (first law of thermodynamics).

Now, assume that after completion of the protocol at time, τ , the system can relax back to thermodynamic equilibrium. Then the amount of work dissipated during this process is that part of the total work, W , that did not contribute to increasing the equilibrium free energy of the system, *i.e.*, $W - (F[x_\tau] - F[x_0]) = W - \Delta F$.

Consider the protocol run in reverse time, $\bar{\mathbf{X}} = x_\tau \dots, x_0$, and ask for the probability, $p_R(\bar{\mathbf{S}}|\bar{\mathbf{X}})$, of finding the exact reverse-time path through state space, $\bar{\mathbf{S}} = s_\tau \dots, s_0$. The ratio between forward time

probability, $p_F(\mathbf{S}|\mathbf{X})$, and reverse time probability, $p_R(\bar{\mathbf{S}}|\bar{\mathbf{X}})$, depends exponentially on the work done in excess of the equilibrium free energy change [58,60,61]:

$$\frac{p_F(\mathbf{S}|\mathbf{X})}{p_R(\bar{\mathbf{S}}|\bar{\mathbf{X}})} = e^{\beta(W-\Delta F)} \tag{13}$$

The work relation, Equation (10), then follows immediately from normalization of probability [60]:

$$1 = \int ds_0 \cdots \int ds_\tau p_R(\bar{\mathbf{S}}|\bar{\mathbf{X}}) = \left\langle \frac{p_R(\bar{\mathbf{S}}|\bar{\mathbf{X}})}{p_F(\mathbf{S}|\mathbf{X})} \right\rangle_{p_F(\mathbf{S}|\mathbf{X})} = e^{\Delta F} \langle e^{-\beta W} \rangle_{p_F(\mathbf{S}|\mathbf{X})},$$

$$e^{-\Delta F} = \langle e^{-\beta W} \rangle_{p_F(\mathbf{S}|\mathbf{X})}.$$

3.2. Nonequilibrium Free Energy and Dissipation

If the system does not instantaneously relax back to thermodynamic equilibrium after being driven away from equilibrium, then detailed knowledge of the system’s state would allow for the extraction of additional free energy, beyond the free energy of the corresponding equilibrium system. This additional free energy (at time t) is proportional [26] to the relative entropy, or *Kullback-Leibler divergence*, between the out-of-equilibrium distribution, p_t , and the corresponding Boltzmann distribution, $p_{\text{eq}}^{x_t} := p_{\text{eq}}(s_t|x_t)$,

$$F_{\text{add}}[p_t, x_t] := k_B T D_{\text{KL}}[p_t || p_{\text{eq}}^{x_t}]. \tag{14}$$

The total nonequilibrium free energy [62–64] is then given by the equilibrium free energy $F[x_t]$ plus this additional free energy:

$$F_{\text{neq}}[p_t, x_t] := F[x_t] + F_{\text{add}}[p_t, x_t] = \langle E(s_t, x_t) \rangle_{p_t} - k_B T H[p_t], \tag{15}$$

with $H[p_t]$ denoting the Shannon entropy associated with the distribution p_t . The second equality follows directly from inserting the Boltzmann distribution into Equation (14).

The additional free energy may be difficult to harness, as doing so requires knowledge about the system that may not be available. However, it could theoretically be used by a clever device. Dissipation, *i.e.*, the work that is *irretrievably* lost, is therefore work done on the system in excess of the system’s *nonequilibrium* free energy change ΔF_{neq} , *i.e.*,

$$\langle W_{\text{diss}} \rangle = \langle W \rangle - \langle \Delta F_{\text{neq}} \rangle. \tag{16}$$

Brackets denote the statistical average. Since we are interested in how much information a system can carry about an arbitrary environment, we have to allow the external driving signal (= protocol) to be stochastic and, as anything else would be too limiting a restriction. We are therefore interested in quantities, averaged not only over $P(\mathbf{S}|\mathbf{X})$, but also over $\mathbf{P}(\mathbf{X})$, and therefore we have to take the average over the joint distribution $P(\mathbf{S}, \mathbf{X})$. Note, however, that an argument analogous to the above can be made when the protocol is assumed to be fixed (see e.g., [65] and references therein).

Average dissipation (Equation 16) is related to the average work done in excess of the equilibrium free energy change, $\langle W_{\text{ex}} \rangle := \langle W \rangle - \Delta F$, by the change in additional free energy due to being out of equilibrium ΔF_{add} , *i.e.*, the change in relative entropy (see Equation (14)):

$$\langle W_{\text{diss}} \rangle = \langle W_{\text{ex}} \rangle - \Delta F_{\text{add}}. \tag{17}$$

Using the fact that dissipation is non-negative, we see that $\langle W_{\text{ex}} \rangle \geq \Delta F_{\text{add}}$. Extra work, in the amount of ΔF_{add} can be extracted from a system that was started in thermodynamic equilibrium ($F_{\text{add}}[p_0, x_0] = 0$) and driven out of equilibrium by the protocol \mathbf{X} , as for such a system $\Delta F_{\text{add}} = F_{\text{add}}[p_\tau, x_\tau] \geq 0$ due to the non-negativity of relative entropy.

This observation has been used to motivate the claim that additional work could be extracted using a “feedback” protocol, *i.e.*, an experimental protocol that is adjusted in response to knowledge of the system, available via a measurement (see e.g. [65,66] and references therein). The notion of “feedback” in this context (e.g. [66]) is more restrictive than the feedback referred to in Sec. 4.1, and in most of the robotics and signal processing literature, where typically *both*, the system and the environment (here: the protocol) evolve in time and influence each other.

3.3. Landauer’s Principle

Imagine building a computing machine that is composed of many small devices (microscale, or even nanoscale). If every part of the machine dissipates heat, it may become a challenge to keep the machine from overheating. Synthetic devices face this problem, as do biological computing systems. It is thus relevant to know what the physical limits on heat generation are, and how they can be achieved.

Landauer argued [67] that the heat generated when one bit of information is erased from a device has to be at least $k_B T \ln(2)$. Take, for example, a simple model for a bistable system: a particle in a double well potential. Assume that one could measure which well the particle is in, but that one would not attempt any other measurements on the device. Then, from the point of view of the potential observer, *i.e.*, the user of the device, this device can store one bit of information. The particle can be either in the left or in the right well.

Now assume that at the beginning of an “erasure” protocol, the observer does not know, but could measure, where the particle is. Proceed with the erasure of this one bit of information by deforming the potential, such that the particle is forced into, say, the left well. At the end of the protocol, the user knows where the particle is. Hence, no further information can be obtained from the device. Therefore, one bit of information has been deleted. This is Szilard’s engine [68] run in reverse [69]. Landauer assumed that both at the beginning and at the end of the protocol the device would be in thermodynamic equilibrium.

He argued that the information erased, \mathcal{I}_e^L , is then directly related to the difference in system entropy, $\Delta H := H[p_{\text{eq}}(s_\tau|x_\tau)] - H[p_{\text{eq}}(s_0|x_0)]$ (given the protocol and assuming that the protocol starts and ends in equilibrium): $\mathcal{I}_e^L = -\Delta H / \log(2)$, where the factor $\log(2)$ converts from nats to bits.

The change in equilibrium free energy, $\Delta F = F[x_\tau] - F[x_0]$, can be written as an average change in energy and an entropic change: $\Delta F = \langle \Delta E \rangle - k_B T \Delta H$. Combining this with the first law of thermodynamics, $W + Q = \Delta E$, we get $\langle W \rangle - \Delta F = -\langle Q \rangle - k_B T \log(2) \mathcal{I}_e^L$. This quantity has to be non-negative, according to the second law of thermodynamics, and therefore we have $-\langle Q \rangle \geq k_B T \log(2) \mathcal{I}_e^L$. So, if one bit of information is erased, $\mathcal{I}_e^L = 1$, then the generated heat is at least $k_B T \log(2)$.

Landauer’s argument is a restatement of the second law of thermodynamics, based on the fact that dissipation is the total change in entropy, which splits up into the change in the environment, given by the heat generated, plus the change in the system’s entropy, which, in turn, he argued, is the negative

information erased during a protocol. Landauer’s principle can be generalized to the case where one does not assume that the system ends in thermodynamic equilibrium [24,65], and to stochastic driving [24]. These generalizations equally are restatements of the (“generalized” [65]) second law.

3.4. Thermodynamics of Prediction

While the treatment of systems driven far from thermodynamic equilibrium usually assumes that a driving protocol is given by some prescribed experimental protocol (e.g. [56,58,65,70]), we obviously cannot make this restricting assumption in the context of adaptation and learning. Instead, the focus has to be on systems embedded in *stochastic* environments.

Let us therefore assume that the protocol is not known, but rather reflects one instantiation of a stochastic environment, which can be described by some distribution, $P(\mathbf{X})$, underlying the generation of the data, x_0, \dots, x_T . The system’s dynamics could be adapted to the environment. This may have occurred via some natural process, for example, evolution, or by means of engineering a synthetic device in some optimal fashion. A system that is adapted to a certain type of stochastic environment would not be optimized with respect to one particular realization of the environment, but rather with respect to the average. Therefore, the following treatment focusses on average quantities, averaged not only over $P(\mathbf{S}|\mathbf{X})$, but also over all possible realizations of the environment (or protocol), $P(\mathbf{X})$.

When the environment is changed from x_t to x_{t+1} , work is being done (see Eq (11)). How much of this work can be extracted from the system? We assume that at each point in time only the current state of system and environment can be measured, and that there is no additional, extraneous lookup table (or any additional memory of any kind). The best estimate of the current environmental signal given the system’s state is then $p(x_t|s_t)$, and the best estimate of the next environmental signal is $p(x_{t+1}|s_t)$. Conversely, the best estimate of the system’s state given the environment is $p(s_t|x_t)$ before the change in the environment, and $p(s_t|x_{t+1})$ after. The equilibrium free energy change associated with this change is simply $\Delta F = F[x_{t+1}] - F[x_t]$, but the associated nonequilibrium free energy change is $\Delta F_{\text{neq}}[x_t \rightarrow x_{t+1}; s_t] := F_{\text{neq}}[p(s_t|x_{t+1})] - F_{\text{neq}}[p(s_t|x_t)]$ [71]. The average instantaneous work lost is then given by $\langle W_{\text{diss}}[x_t \rightarrow x_{t+1}; s_t] \rangle := \langle W[x_t \rightarrow x_{t+1}; s_t] \rangle_{p(s_t, x_t, x_{t+1})} - \langle \Delta F_{\text{neq}}[x_t \rightarrow x_{t+1}; s_t] \rangle_{p(x_t, x_{t+1})}$. This dissipated work is proportional [24] to instantaneous nonpredictive information, $I[s_t, x_t] - I[s_t, x_{t+1}]$:

$$\beta \langle W_{\text{diss}}[x_t \rightarrow x_{t+1}; s_t] \rangle = \beta \langle E(s_t, x_{t+1}) \rangle_{p(s_t, x_{t+1})} - \langle E(s_t, x_t) \rangle_{p(s_t, x_t)} - \beta \left(\langle F_{\text{neq}}[p(s_t|x_{t+1})] \rangle_{p(x_{t+1})} - \langle F_{\text{neq}}[p(s_t|x_t)] \rangle_{p(x_t)} \right) \tag{18}$$

$$= H[s_t|x_{t+1}] - H[s_t|x_t] \tag{19}$$

$$= I[s_t, x_t] - I[s_t, x_{t+1}]. \tag{20}$$

Line (19) follows directly from averaging Equation (15) over the external signal, whereby the conditional entropy of the system conditioned on the external control parameter can be expressed in terms of that part of the average energy that is not available as nonequilibrium free energy:

$$H[s_t|x_t] = \beta \langle \langle E(s_t, x_t) \rangle_{p(s_t|x_t)} - F_{\text{neq}}[p(s_t|x_t)] \rangle_{p(x_t)}. \tag{21}$$

The last equality (20) follows immediately from the fact that mutual information measures a reduction in entropy [39]: $I[s, x] = H[s] - H[s|x]$.

Importantly, the instantaneous nonpredictive information allows us to judge the model that is implicit in the system’s dynamics by the criteria governing predictive inference, namely that a good model should have large predictive power while not being overly complicated. Some part of the instantaneous memory, $I[s_t, x_t]$, retained in the system’s state at time t , is predictive, $I[s_t, x_{t+1}]$. But the rest, $I[s_t, x_t] - I[s_t, x_{t+1}]$, represents a measure for how much nonpredictive “clutter” [20], or “nostalgia” [24], is kept at any instant in time. In that way it characterizes the ineffectiveness, or inefficiency, of the predictive inference that the system implements implicitly via its dynamics. It reaches zero when none of the instantaneous memory gets wasted on nonpredictive information, *i.e.* $I[s_t, x_t] = I[s_t, x_{t+1}]$. This could be achieved trivially by keeping no memory at all, $I[s_t, x_t] = 0$, and having no predictive power. However, a system that fulfills a function and operates at a finite rate usually has to be correlated to some degree with the environmental driving, and thus will have nonzero instantaneous memory. In many cases $I[s_t, x_{t+1}] \leq I[s_t, x_t]$, but in some cases, it is possible to have more instantaneous predictive power than instantaneous memory, *i.e.* $I[s_t, x_{t+1}] > I[s_t, x_t]$, because information from the more distant past that may increase predictive power can be carried by the system dynamics.

The proportionality between instantaneous nonpredictive information and dissipation, Equation (20), carries over to the quantum regime [72], where this framework has proven useful to give a new interpretation to quantum discord, identified as “the thermodynamic inefficiency of the most energetically efficient classical approximation of a quantum memory” [72].

Total instantaneous nonpredictive information (summed over the entire protocol) $I_{\text{nonpred}} := \sum_{t=0}^{\tau-1} (I[s_t, x_t] - I[s_t, x_{t+1}])$, provides a lower bound on work lost on average, assuming that the system is in thermodynamic equilibrium at the beginning of the protocol [24]:

$$\beta \langle W_{\text{ex}} \rangle \geq I_{\text{nonpred}}. \tag{22}$$

To see this, let us compute the contributions to dissipation during relaxation steps, and add them to the dissipation during work steps (given by Equation 20). Note, that during a relaxation step, no work is done. Furthermore, the environmental variable does not change, and neither does the free energy of the corresponding equilibrium distribution. Therefore, dissipation during relaxation steps is solely given by the negative change in additional free energy, which is proportional to the difference in Kullback-Leibler divergence from equilibrium,

$$-\Delta F_{\text{add}}[s_{t-1} \rightarrow s_t; x_t] = k_B T (D_{\text{KL}}[p(s_{t-1}|x_t)||p_{\text{eq}}^{x_t}] - D_{\text{KL}}[p(s_t|x_t)||p_{\text{eq}}^{x_t}]) . \tag{23}$$

Since the relative entropy between the actual distribution and the corresponding equilibrium distribution is a Lyapunov function [73], we obtain, using Equation (20), the following bound on the quantity

$$\begin{aligned} \left\langle W - (F_{\text{neq}}[p(s_\tau|x_\tau)] - F_{\text{neq}}[p(s_0|x_0)]) \right\rangle &= \left\langle \sum_{t=0}^{\tau-1} (W_{\text{diss}}[x_t \rightarrow x_{t+1}; s_t] - \Delta F_{\text{add}}[s_t \rightarrow s_{t+1}; x_{t+1}]) \right\rangle \tag{24} \\ &\geq k_B T I_{\text{nonpred}} . \tag{25} \end{aligned}$$

The average dissipation on the l.h.s., in turn, sets a lower bound to the average work done in excess of equilibrium free energy change, $\langle W_{\text{ex}} \rangle$: remember that, if the system is in thermodynamic equilibrium at time $t = 0$, we have $F_{\text{add}}[p(s_0|x_0)] = 0$, and the l.h.s of (24) is then $\langle W \rangle - \Delta F - F_{\text{add}}[p(s_\tau|x_\tau)] \leq \langle W_{\text{ex}} \rangle$, due to the non-negativity of relative entropy. Altogether, we arrive at (22).

Inequality (22) leads to a refinement of Landauer's argument. We use, as before, the fact that $\beta W_{\text{ex}} = -\beta \langle Q \rangle - \mathcal{I}_e$, (where $\mathcal{I}_e = \log(2) \mathcal{I}_e^L$ is Landauer's erasure in nats, for convenience), to arrive at the conclusion that the heat leaving the system is lower bound by

$$-\beta \langle Q \rangle \geq \mathcal{I}_e + I_{\text{nonpred}}. \quad (26)$$

Landauer's bound is augmented by the total instantaneous nonpredictive information.

These insights are applicable to systems that can be driven arbitrarily far from thermodynamic equilibrium, spanning a wide range, including artificial computing devices, as well as biomolecular machines. The direct connection between nonpredictive information and dissipation hints towards a possible underlying physical reason for the emergence of predictive inference: naturally occurring computation may be implementing predictive inference because this information processing strategy may also allow for the efficient use of energy. This may be relevant on the small scales on which the machinery of life operates, where $k_B T$ is not negligible. Some bio-machines approach 100% thermodynamic efficiency, such as the F_1 -ATPase [74], a molecule crucial for the energy metabolism of cells. When driven in a natural fashion, the stall torque is near the maximal values possible, given the free energy liberated by ATP hydrolysis and the size of the rotation. Similar efficiencies have been observed also in other bio-molecular motors, e.g., [75]. Such optimal behavior may imply that even these micro-machines might implement predictive inference implicitly via their dynamics.

4. Recursive Schemes

The physical view of information processing laid out above motivates that one could compress a time series by constructing a *dynamical* rule that determines the conditional state-updates, such that the resulting model has maximal predictive power at fixed memory [14].

4.1. Interactive Learning

In the most general case, learning machines are able to interact with, and change, their environment, as animals do when they learn. The learning system then no longer learns passively. Instead, its actions feed back to the environment so that the time evolution of the environment depends on the actions taken. Thereby the future that the learning system encounters depends, to some degree, on its own actions. This type of learning is *interactive* learning.

Dynamical rules have to be found not only for the internal state of the learning machine, but also for the actions that the machine can take. If one postulates that the predictive power of the resulting behavior should be maximized at fixed coding cost, then an information theoretic approach reveals that optimal action policies *must* balance control and exploration [14]. The approach provides a generalization of previously existing concepts, such as the causal state partition, to interactive learning. It furthermore allows for an intuitive treatment of curiosity-driven reinforcement learning [17], to which many alternative and some related approaches exist (see discussion and references in [17]).

4.2. Recursive Information Bottleneck Method (RIB)

For the present discussion of predictive inference, the treatment shall remain restricted to representations extracted from the data that do not influence future experiences directly (*i.e.* passive learning). This case is contained in the more general treatment developed in [14] and can be retrieved directly by deleting the action variables throughout [14]. By doing so, one obtains a *recursive* version of the information bottleneck method.

The learning machine updates its state from s_{t-1} to s_t , after it receives a new input, x_t . The update dynamics are given by $p(s_t|s_{t-1}, x_t)$. They characterize the model, together with the predictions encoded in $p(\vec{x}_t|s_t)$, and the state distribution, $p(s_t)$. The coding rate associated with the dynamics, $I[s_t; \{s_{t-1}, x_t\}]$, measures the complexity of the model. Dynamics are then constructed, such that predictive power is maximized, under a constraint on the coding cost, or memory [14]:

$$\max_{p(s_t|s_{t-1}, x_t)} \left(I[s_t, \vec{x}_t] - \lambda I[s_t; \{s_{t-1}, x_t\}] \right). \tag{27}$$

This optimization problem is solved by dynamics that obey the following equations (Equations (14)–(17) in [14], with actions taken out, and history = $\{x_t, s_{t-1}\}$):

$$p(s_t|x_t, s_{t-1}) = \frac{p(s_t)}{Z(x_t, s_{t-1})} \exp \left(-\frac{1}{\lambda} D_{KL}[p(\vec{x}_t|x_t, s_{t-1})||p(\vec{x}_t|s_t)] \right) \tag{28}$$

with

$$p(\vec{x}_t|s_t) = \frac{1}{p(s_t)} \sum_{x_t, s_{t-1}} p(\vec{x}_t|x_t, s_{t-1})p(s_t|x_t, s_{t-1})p(x_t, s_{t-1}), \tag{29}$$

$$p(s_t) = \sum_{x_t, s_{t-1}} p(s_t|x_t, s_{t-1})p(x_t, s_{t-1}), \tag{30}$$

$$Z(x_t, s_{t-1}) = \sum_{s_t} p(s_t) \exp \left(-\frac{1}{\lambda} D_{KL}[p(\vec{x}_t|x_t, s_{t-1})||p(\vec{x}_t|s_t)] \right). \tag{31}$$

These dynamics group new incoming data, x_t , together with the current state, s_{t-1} , according to their similarity in terms of conditional future distributions, thereby creating, in each step, an incrementally more predictive model.

Since the state, s_t , is computed from the variables, s_{t-1} and x_t , knowing it does not add information about the future, due to the data processing inequality [39]. This is equivalent to saying that knowledge of the new state does not change the distribution over futures, when the old state and the current input are given, because the new state is obtained as a function of only these two variables and hence, no external information enters: $p(\vec{x}_t|s_t, s_{t-1}, x_t) = p(\vec{x}_t|s_{t-1}, x_t)$. A similar assumption is made in the original Information Bottleneck method, where the compression of the input data does not change the distribution over the relevant quantity, assuming the data are given. This assumption is crucial and warrants the name “recursive information bottleneck (RIB)” for this new, recursive compression scheme. Note that in the general case, however, when the learner’s actions can change its future input, this assumption is no longer valid, and the rate-distortion framework has to be extended beyond its original scope, leaving the recursive information bottleneck method as a special case of this more general scenario [14]. Learning

with feedback as in [14] can easily be extended from the “dynamical” learning discussed in this Section to the “static” learning treated by the original IB method (see Sec. 2).

The RIB method’s algorithmic procedure [14] starts by initializing the machine states to a uniform distribution that is uncorrelated with the observations, and as a consequence, the first iteration of RIB simply runs the IB algorithm (as explained in Sec. 2) with pasts of length one, and futures of length τ_F , where the initial input distribution, $p(\vec{x}_t|x_t)$, is acquired from the input data. Then, a sequence of L states is produced, using the optimal assignments which have been obtained by the initial compression: $p_0(s_t|x_t)$. The duration, L , controls how long the learning machine tests its model of the environment, before re-evaluating it. After L steps, which are used to produce new states and to acquire the associated new input statistics, $p_j(\vec{x}_t|x_t, s_{t-1})$, RIB then solves Equations (28)–(31) and produces new optimal state assignments, $p_j(s_t|x_t, s_{t-1})$, which are, in turn, used to produce the next L states. This procedure is repeated iteratively. (The index j labels iteration number.)

Each step, j , has a trade-off parameter, λ_j , associated with it. Sampling errors can lead to over-fitting when this parameter is too small [43]. This becomes more pronounced, the smaller L is. L can be as short as $L = 1$, its minimum, when the algorithm operates in an “online”-fashion. One can then decrease λ_j as a function of j , since sampling errors decrease as more data are accumulated over time. If a finite time series is given, then the maximum value of L is limited by the length of the time series. In that case, when L is set to its maximum, the algorithm makes several passes over the entire data batch. For theoretical guidance regarding λ_j , results from [43] can be used.

A possible advantage of this scheme is that the state space of the learning machine will increase only as much as necessary for prediction. Instead of having to estimate $p(\overleftarrow{x}_t, \vec{x}_t)$ (as for OCI), one iteratively re-estimates $p(\vec{x}_t|x_t, s_{t-1})$, potentially easing the estimation problem significantly.

However, depending on the complexity of the input data, the model quality may rely on the use of long futures. If infinitely long futures are required to learn the best possible model, then any practical procedure has to be suboptimal, as it has to deal with finite data. But it is interesting to study this asymptotic regime ($\tau_F \rightarrow \infty$) theoretically, in order to understand the general capacity of the method.

4.3. Asymptotic Behavior of Recursive Information Bottleneck

In the limit $\tau_F \rightarrow \infty$, and taking the limit $\lambda_j \rightarrow 0 \forall j$, RIB finds the causal state partition (and thus minimal sufficient statistics), together with deterministic state transitions, yielding as a special case in this limit the “ ϵ -machine”, which is the unique maximally predictive and deterministic Hidden Markov Model of a given time series [21,23,28,36].

To see how this works, consider pasts of length $\tau_P = t$, whereby $\overleftarrow{x} = (x_0, \dots, x_t)$, and define the function, $f(\overleftarrow{x}_t)$, which creates a partition of the space of all past trajectories, such that all \overleftarrow{x}_t which give the same the conditional future distribution, $p(\vec{x}_t|\overleftarrow{x}_t)$, are mapped onto the same value, $f(\overleftarrow{x}_t)$, i.e. $p(\vec{x}_t|\overleftarrow{x}_t) = p(\vec{x}_t|f(\overleftarrow{x}_t))$. This is one (intuitive) way of defining the causal state partition. We have to show that the states of RIB recovers $f(\overleftarrow{x}_t)$, $\forall t$. This can be done by mathematical induction (we present here an intuitive argument and leave tedious technical details for later).

Since the RIB procedure is initialized by OCI run on pasts of length one, we have after the initial optimization, in the limit $\lambda_0 \rightarrow 0$, the assignments $p_0(s_t|\overleftarrow{x}_0) = \begin{cases} 1 & \text{if } s_t = s^0 \\ 0 & \text{otherwise} \end{cases}$, where the optimal solution is $s^0 = f(\overleftarrow{x}_0)$ [16]. This is the basis step for the proof by induction. The inductive hypothesis is that $s^t = f(\overleftarrow{x}_t)$, and we have to show that if this is true, then $s^{t+1} = f(\overleftarrow{x}_{t+1})$.

Since we let $\lambda_j \rightarrow 0, \forall j$, we know [16,43] that the optimal solution is the one that minimizes the relative entropy in the exponent of Equation 28. So, we have to evaluate $D_{\text{KL}}[p(\overrightarrow{x}_{t+1}|x_{t+1}, s_t = s^t)||p(\overrightarrow{x}_{t+1}|s_{t+1})]$. We can write $p(\overrightarrow{x}_{t+1}|x_{t+1}, s^t) = p(\overrightarrow{x}_{t+1}, x_{t+1}|s^t)/p(x_{t+1}|s^t)$. Recall that \overrightarrow{x}_{t+1} contains all data from x_{t+2} onwards. Therefore $p(\overrightarrow{x}_{t+1}, x_{t+1}|s^t) = p(\overrightarrow{x}_t|s^t)$. Using the inductive hypothesis and the definition of f , we then have $p(\overrightarrow{x}_{t+1}|x_{t+1}, s^t) = p(\overrightarrow{x}_t|\overleftarrow{x}_t)/p(x_{t+1}|\overleftarrow{x}_t)$. Finally, we pull out x_{t+1} in the numerator to get $p(\overrightarrow{x}_{t+1}|x_{t+1}, s^t) = p(\overrightarrow{x}_{t+1}|x_{t+1}, \overleftarrow{x}_t) = p(\overrightarrow{x}_{t+1}|\overleftarrow{x}_{t+1})$. That means that the optimal solution is given by the states s_{t+1} that minimize $D_{\text{KL}}[p(\overrightarrow{x}_{t+1}|\overleftarrow{x}_{t+1})||p(\overrightarrow{x}_{t+1}|s_{t+1})]$. Those are [16,43] $s_t^{t+1} = f(\overleftarrow{x}_{t+1})$. Thus, as $t \rightarrow \infty$, the causal state partition is discovered. The resulting HMM is related to the ϵ -machine, because the corresponding state transition probabilities are deterministic as $\lambda \rightarrow 0$. Define $D(s_{t+1}) := D_{\text{KL}}[p(\overrightarrow{x}_{t+1}|x_{t+1}, s^t)||p(\overrightarrow{x}_{t+1}|s_{t+1})] - D_{\text{KL}}[p(\overrightarrow{x}_{t+1}|x_{t+1}, s^t)||p(\overrightarrow{x}_{t+1}|s^{t+1})] \geq 0$, which is non-negative by definition of the optimal s^{t+1} . Now consider the probability that the system goes from the optimal state s^t to optimal state s^{t+1}

$$p(s^{t+1}|x_{t+1}, s^t) = \left(1 + \sum_{s_{t+1} \neq s^{t+1}} \frac{p(s_{t+1})}{p(s^{t+1})} e^{-\frac{1}{\lambda} D(s_{t+1})} \right)^{-1} \xrightarrow{\lambda \rightarrow 0} 1 \tag{32}$$

which tends to one as $\lambda \rightarrow 0$, assuming that $p(s^{t+1})$ is non-zero.

This result holds in general, regardless of the underlying distribution that generated the data. It is important, because it shows that RIB has the capacity to discover the minimal unique and sufficient predictive model, a representation that can be regarded as the best possible predictive model that can be constructed from observations of a stochastic process alone [23,36,42]. Conveniently, we do not have to evaluate or compare infinite past trajectories when using this recursive method. Furthermore, while the ϵ -machine is defined for infinite future trajectories, it is obvious from the above treatment that when the notion of causal state partition is extended to finite futures, the argument above still applies for finite values for τ_F .

4.4. Modeling Linear Dynamical Systems

Assume that the transfer function that characterizes a dynamical system generating the observed data is linear with additive Gaussian noise. Then, using the IB to find a reduced representation of the dynamical system becomes an Eigenvalue problem. The reduced system dynamics can be computed analytically, given the assumed underlying linear system dynamics. This approach was coined *past-future* IB (PFIB) [19]. The Eigenvectors that determine the reduced system are the same as those computed by canonical correlation analysis (CCA) [19,76].

5. Conclusions

Predictive inference can be interpreted as a strategy for effective and efficient communication: past experiences are compressed into a representation that is maximally informative about future experiences. The information bottleneck (IB) framework can thus be applied, either in a direct way, or in its recursive form (RIB). Both methods find, asymptotically, the causal state partition, *i.e.*, minimal sufficient statistics. RIB additionally recovers, asymptotically, the ϵ -machine, which is a maximally predictive and minimally complex deterministic HMM, believed to be the best predictive description of a stochastic process that can be extracted from the data alone.

While the main appeal of the IB framework is its generality, and that no assumptions have to be made about the distribution that generated the data, linear and Gaussian model assumptions do result in signal processing methods that are related to known methods, such as canonical correlation analysis.

Beyond philosophical motivations, the information theoretic approach to predictive inference laid out here can also be motivated from thermodynamic considerations. Prediction and energetic efficiency are tightly coupled, because instantaneous nonpredictive information is fundamentally related to dissipation. Implemented on a physical system, predictive inference may thus constitute a strategy for using energy efficiently by minimizing dissipation.

In summary, predictive inference may have advantages, not only in the abstract world of thoughts, where it enables efficient communication, but also in a concrete thermodynamic sense. The information bottleneck framework offers an intuitive approach to an overarching theory.

Acknowledgments

I thank my collaborators: countless inspiring conversations with William Bialek initiated and crucially formed this line of work, which then grew from joint work with Doina Precup, later from work with James P. Crutchfield and Chris Ellison, and finally from a collaboration with Anthony J. Bell, Gavin E. Crooks and David A. Sivak. Discussions with Léon Bottou, Michael R. DeWeese, Imre Kondor, Matteo Marsili, Ilya Nemenmann, Barak Pearlmutter, Robert S. Shaw and Chris Wiggins were indispensable. I am furthermore grateful for conversations with Gal Chechik, Arne Grimsmo, Juergen Jost, Wolfgang Löhr, Fritz Sommer, and Naftali Tishby. I thank Sarah Marzen and Lisa Miller for reading the manuscript and providing valuable feedback.

References

1. Jeffreys, H. *Theory of Probability*, 3rd ed.; Oxford University Press: Oxford, UK, 1998; First version published in 1939.
2. Geisser, S. *Predictive Inference: An introduction*; CRC Press: New York, NY, USA, 1993. Volume 55.
3. Vapnik, V. *Statistical Learning Theory*; John Wiley and Sons, New York, NY, USA, 1998.
4. Chaitin, G.J. *Algorithmic Information Theory*; Cambridge University Press, Cambridge, UK, 2004. Volume 1.

5. Kolmogorov, A.N. On tables of random numbers. *Sankhyā: Indian J. Stat. Series A* **1963**, *25*, 369–376.
6. Ladyman, J.L.; Lambert, J.; Wiesner, K. What is a complex system? *Euro. J. Phil. Sci.* **2013**, *3*, 33–67.
7. Straumann, N. General relativity and relativistic astrophysics. *Research supported by the Swiss National Science Foundation*; Springer-Verlag: Berlin, Germany, 1984; p.468.
8. Misner, C.W.; Thorne, K.S.; Wheeler, J.A. *Gravitation*; Macmillan: Landon, UK, 1973.
9. Packard, N.; Crutchfield, J.; Farmer, J.; Shaw, R. Geometry from a Time Series. *Phy. Rev. Lett.* **1980**, *45*, 712.
10. Eckmann, J.P.; Ruelle, D. Ergodic theory of chaos and strange attractors. *Rev. Modern phy.* **1985**, *57*, 617.
11. Tishby, N.; Pereira, F.; Bialek, W. The information bottleneck method. In Proceedings of the 37th Annual Allerton Conference, Monticello, IL, USA, 1999; pp. 363–377.
12. Bialek, W. Thinking about the brain. *Physics of Bio-molecules and Cells, Ecole dte de physique theorique Les Houches Session LXXV*; Springer-Verlag: Berlin, Germany, 2001; pp. 485–577.
13. Bialek, W.; Nemenman, I.; Tishby, N. Predictability, Complexity and Learning. *Neural Computat.* **2001**, *13*, 2409–2463.
14. Still, S. Information-theoretic approach to interactive learning. **2007**. arXiv: 0709.1948; Revised version: EPL (2009) 85, 28005.
15. Still, S.; Crutchfield, J.P. Structure or Noise? *arXiv* **2007**, available at arXiv:0708.0654.
16. Still, S.; Crutchfield, J.P.; Ellison, C. Optimal Causal Inference. *arXiv* **2007**, arXiv: 0708.1580; Revised version: CHAOS (2010), 20, Special Issue on Intrinsic and Designed Computation: Information Processing in Dynamical Systems, 037111.
17. Still, S.; Precup, D. An information theoretic approach to curiosity-driven reinforcement learning. *Theor. Biosci.* **2012**, *131*, 139–148. (Original version submitted to NIPS 2006).
18. Creutzig, F.; Sprekeler, H. Predictive Coding and the Slowness Principle: an Information-Theoretic Approach. *Neural Computat.* **2008**, *20*, 1026–1041.
19. Creutzig, F.; Globerson, A.; Tishby, N. The Past-Future Information Bottleneck of Dynamical Systems. *Phys. Rev. E* **2009**, *79*, 041925.
20. Bialek, W. *Biophysics: Searching for Principles*; Princeton University Press: Princeton, NJ, USA, 2012.
21. Crutchfield, J.P.; Young, K. Inferring Statistical Complexity. *Phys. Rev. Lett.* **1989**, *63*, 105–108.
22. Wiskott, L.; Sejnowski, T. Slow feature analysis: Unsupervised learning of invariances. *Neur. Comput.* **2002**, *14*, 715–770.
23. Shalizi, C.R.; Crutchfield, J.P. Computational Mechanics: Pattern and Prediction, Structure and Simplicity. *J. Stat. Phys.* **2001**, *104*, 817–879.
24. Still, S.; Sivak, D.A.; Bell, A.J.; Crooks, G.E. Thermodynamics of Prediction. *Phys. Rev. Lett.* **2012**, *109*, 120604.
25. Shaw, R.S. Strange attractors, chaotic behavior, and information flow. *Z. Naturforsch* **1981**, *36*, 80–112.

26. Shaw, R. *The Dripping Faucet as a Model Chaotic System*; Aerial Press: Santa Cruz, CA, USA, 1984.
27. Grassberger, P. Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.* **1986**, *25*, 907–938.
28. Crutchfield, J.P. The Calculi of Emergence: Computation, Dynamics, and Induction. *Physica D* **1994**, *75*, 11–54.
29. Nemenman, I. Information theory and learning: A physical approach. *arXiv preprint physics/0009032* **2000**.
30. Bialek, W.; Nemenman, I.; Tishby, N. Predictability, Complexity and Learning. *Neural Comput.* **2001**, *13*, 2409–2463.
31. Shalizi, C.R.; Crutchfield, J.P. Information Bottlenecks, Causal States, and Statistical Relevance Bases: How to Represent Relevant Information in Memoryless Transduction. *Adv. Complex Sys.* **2002**, *5*, 91–95.
32. Crutchfield, J.P. Between order and chaos. *Nat. Phys.* **2011**, *8*, 17–24.
33. Li, W. Mutual information functions versus correlation functions. *J. Stat. Phys.* **1990**, *60*, 823–837.
34. Shannon, C.E. A mathematical theory of communication. *Bell. Syst. Tech. J.* **1948**, *27*, 379–423, 623–656.
35. Notation employed in this paper relies on context to distinguish random variables from realizations thereof (conventionally, random variables would be denoted by capital letters). We write out the explicit dependency of entropy on the distribution only in places where it is relevant to put emphasis on the distribution, using the notation $H[p(x|y)] := -\langle \log[p(x|y)] \rangle_{p(x|y)}$, where $p(x|y)$ is a conditional distribution over x , given y . In all other places we use the shortcut $H[x] := -\langle \log[p(x)] \rangle_{p(x)}$ for entropy, $H[x|y] := -\langle \log[p(x|y)] \rangle_{p(x,y)}$ for conditional entropy, and $I[x, y] := \left\langle \log \left[\frac{p(x,y)}{p(x)p(y)} \right] \right\rangle_{p(x,y)}$ for mutual information.
36. Crutchfield, J.P.; Shalizi, C.R. Thermodynamic Depth of Causal States: Objective Complexity via Minimal Representations. *Phys. Rev. E* **1999**, *59*, 275–283.
37. Shalizi, C.R.; Crutchfield, J.P. Information Bottlenecks, Causal States, and Statistical Relevance Bases: How to Represent Relevant Information in Memoryless Transduction. *Adv. Complex Sys.* **2002**, *5*, 1–5.
38. Tchernookov, M.; Nemenman, I. Predictive information in a nonequilibrium critical model. *arXiv*, **2012**, arXiv:1212.3896.
39. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: Hoboken, NJ, USA, 2006.
40. Arimoto, S. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Info. Theor.* **1972**, *18*, 14–20.
41. Blahut, R.E. Computation of channel capacity and rate-distortion functions. *IEEE Trans. Info. Theor.* **1972**, *18*, 460–473.
42. Shalizi, C.R. Causal architecture, complexity and self-organization in the time series and cellular automata. PhD thesis, University of Wisconsin–Madison, Madison, USA, 2001.

43. Still, S.; Bialek, W. How many clusters? An information theoretic perspective. *Neural Computat.* **2004**, *16*, 2483–2506.
44. Rose, K.; Gurewitz, E.; Fox, G.C. Statistical Mechanics and Phase Transitions in Clustering. *Phys. Rev. Lett.* **1990**, *65*, 945–948.
45. Rose, K. Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems. *Proc. IEEE* **1998**, *86*, 2210–2239.
46. Pereira, F.; Tishby, N.; Lee, L. Distributional Clustering of English Words. 30th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Columbus, Ohio, 1993; pp. 183–190. available at xxx.lanl.gov/pdf/cmp-lg/9408011.
47. Milner, R. An Algebraic notion of simulation between programs. Proceeding of International Joint Conference on Artificial Intelligence, London, UK, September, 1971; pp. 481 – 489.
48. Jaeger, H. Observable operator models for discrete stochastic time series. *Neural Comput.* **2000**, *12*, 1371–1398.
49. Löhr, W. Models of Discrete-Time Stochastic Processes and Associated Complexity Measures. PhD thesis, Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany, 2010.
50. Shalizi, C.R.; Klinkner, K.; Crutchfield, J.P. An Algorithm for Pattern Discovery in Time Series. *arXiv* **2002**, arXiv:cs/0210025.
51. Pearl, J. Causal inference in statistics: An overview. *Stat. Surv.* **2009**, *3*, 96–146.
52. Chechnik, G.; Globerson, A.; Tishby, N.; Weiss, Y. Information Bottleneck for Gaussian variables. *J. Mach. Learn. Res.* **2005**, *6*, 165–188.
53. Plenio, M.B.; Vitelli, V. The physics of forgetting: Landauer's erasure principle and information theory. *Contemp. Phys.* **2001**, *42*, 25–60.
54. Jarzynski, C. Nonequilibrium work relations: foundations and applications. *Eur. Phys. J. B* **2008**, *64*, 331–340.
55. Jarzynski, C. Equalities and Inequalities: Irreversibility and the Second Law of Thermodynamics at the Nanoscale. *Annu. Rev. Condens. Matter Phys.* **2011**, *2*, 329–351.
56. Jarzynski, C. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* **1997**, *78*, 2690.
57. Chandler, D. *Introduction to Modern Statistical Mechanics*; Oxford University Press, Oxford, UK, 1987.
58. Crooks, G.E. Nonequilibrium Measurements of Free Energy Differences for Microscopically Reversible Markovian Systems. *J. Stat. Phys.* **1998**, *90*, 1481.
59. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630.
60. Crooks, G.E. Excursions in Statistical Dynamics. PhD thesis, University of California, Berkeley, USA, 1999.
61. Crooks, G.E. Entropy production fluctuation theorem and the nonequilibrium work relation for free-energy differences. *Phys. Rev. E* **1999**, *60*, 2721.
62. Gaveau, B.; Schulman, L.S. A general framework for non-equilibrium phenomena: The master equation and its formal consequences. *Phys. Lett. A* **1997**, *229*, 347–353.

63. Qian, H. Relative Entropy: Free Energy Associated with Equilibrium Fluctuations and Nonequilibrium Deviations. *Phys. Rev. E* **2001**, *63*, 042103.
64. Crooks, G.E. Beyond Boltzmann-Gibbs statistics: Maximum entropy hyperensembles out-of-equilibrium. *Phys. Rev. E* **2007**, *75*, 041119.
65. Esposito, M.; VandenBroeck, C. Second law and Landauer principle far from equilibrium. *EPL (Europhys. Lett.)* **2011**, *95*, 40004.
66. Sagawa, T.; Ueda, M. Fluctuation Theorem with Information Exchange: Role of Correlations in Stochastic Thermodynamics. *Phys. Rev. Lett.* **2012**, *109*, 180602.
67. Landauer, R. Irreversibility and heat generation in the computing process. *IBM J. Res. Develop.* **1961**, *5*, 183–191.
68. Szilard, L. Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen. *Zeitschrift für Physik* **1929**, *53*, 840–856.
69. Magnasco, M.O. Szilard's heat engine. *EPL (Europhys. Lett.)* **1996**, *33*, 583.
70. Seifert, U. Stochastic thermodynamics: principles and perspectives. *EPJ B* **2008**, *64*, 423–431.
71. The argument x_t is dropped, because it is already spelled out explicitly in p_t , *i.e.*, we replace $F_{\text{add}}[p(s_t|x_t), x_t]$ by the shorthand $F_{\text{add}}[p(s_t|x_t)]$.
72. Grimsmo, A.L. Quantum correlations in predictive processes. *Phys. Rev. A* **2013**, *87*, 060302. arXiv:1302.5552.
73. Schlögl, F. On stability of steady states. *Zeitschrift für Physik* **1971**, *243*, 303–310.
74. Kinosita, K.; Yasuda, R.; Noji, H.; Adachi, K. A rotary molecular motor that can work at near 100% efficiency. *Philos. T. Roy. Soc. B* **2000**, *355*, 473–489.
75. Cappello, G.; Pierobon, P.; Symonds, C.; Busoni, L.; Gebhardt, J.C.M.; Rief, M.; Prost, J. Myosin V stepping mechanism. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 15328–15333.
76. Hotelling, H. Relations between two sets of variates. *Biometrika* **1936**, *28*, 321–377.