Information Theory of Decisions and Actions

Naftali Tishby and Daniel Polani

Abstract

The perception-action cycle is often defined as "the circular flow of *information* between an organism and its environment in the course of a sensory guided sequence of actions towards a goal" (Fuster 2001, 2006). The question we address in this paper is in what sense this "flow of information" can be described by Shannon's measures of information introduced in his mathematical theory of communication. We provide an affirmative answer to this question using an intriguing analogy between Shannon's classical model of communication and the Perception-Action-Cycle. In particular, decision and action sequences turn out to be directly analogous to codes in communication, and their complexity — the minimal number of (binary) decisions required for reaching a goal — directly bounded by information measures, as in communication. This analogy allows us to extend the standard Reinforcement Learning framework. The latter considers the future expected reward in the course of a behaviour sequence towards a goal (value-to-go). Here, we additionally incorporate a measure of information associated with this sequence: the cumulated information processing cost or bandwidth required to specify the future decision and action sequence (information-to-go).

Using a graphical model, we derive a recursive Bellman optimality equation for information measures, in analogy to Reinforcement Learning; from this, we obtain new algorithms for calculating the optimal trade-off between the value-to-go and the required information-to-go, unifying the ideas behind the Bellman and the Blahut-Arimoto iterations. This trade-off between value-to-go and information-to-go provides a complete analogy with the compression-distortion trade-off in source coding. The present new formulation connects seemingly unrelated optimization problems. The algorithm is demonstrated on grid world examples.

Naftali Tishby
Hebrew University Jerusalem, e-mail: tishby@cs.huji.ac.il

Daniel Polani
University of Hertfordshire e-mail: d.polani@herts.ac.uk

1 Introduction

To better understand intelligent behaviour in organisms and to develop such behaviour for artificial agents, the principles of perception, of intelligent information processing, as well as of actuation undergo significant study. Perception, information processing and actuation per se are often considered as individual, separate input-output processes. Much effort is devoted to understand and study each of these processes individually. Conceptually, the treatment of such input-output models is straightforward, even if its details are complex.

Compared to that, combining perception, information processing and actuation together introduces a feedback cycle that considerably changes the "rule of the game". Since actuation changes the world, perception ceases to be passive and will, in future states, generally depend on actions selected earlier by the organism. In other words, the organism controls to some extent not only which states it wishes to visit, but consequently also which sensoric inputs it will experience in the future.

The "cycle" view has intricate consequences and creates additional complexities. It is, however, conceptually more satisfying. Furthermore, it can help identifying biases, incentives and constraints for the self-organized formation of intelligent processing in living organisms — it is no surprise that the *embodied intelligence* perspective is adopted by many AI researchers (Pfeifer and Bongard 2007) which is intimately related to the perception-action cycle perspective. It has been seen as a path towards understanding where biological intelligence may have risen from in evolution and how intelligent dynamics may be coaxed out of AI systems.

A challenge for the quantitative treatment of the perception-action cycle is that there are many ways of modeling it which are difficult to compare. Much depends on the choice of architecture, the selected representation and other aspects of the concrete model. To alleviate this unsatisfactory situation, recent work has begun to study the perception-action cycle in the context of a (Shannonian) information-theoretic treatment (Bialek et al. 2001; Touchette and Lloyd 2000, 2004; Klyubin et al. 2004, 2007), reviving early efforts by Ashby (1956).

The information-theoretic picture is universal, general, conceptually transparent and can be post hoc imbued with the specific constraints of particular models. On the informational level, scenarios with differing computational models can be directly compared with each other. At the same time, the informational treatment allows one to incorporate limits in the information processing capacity that are fundamental properties of a particular agent-environment system.

This is especially attractive in that it seems to apply to biologically relevant scenarios to some extent; details of this view are increasingly discussed (Taylor et al. 2007; Polani 2009). Under this perspective, in essence, one considers e.g. the informational cost of handling a given task. Vice versa, one can ask how well one can actually handle a task if constraints on the computational power are imposed (here in the form of limited informational bandwidth).

On the other side, there is the established framework of *Markovian Decision Problems* (MDPs) which is used to study how to find *policies* (i.e. agent strategies) that perform a task well, where the quality of the performance is measured via some

cumulative reward value which depends on the policy of the agent. The MDP framework is concerned with describing the task and with solving the problem of finding the optimal policy. It is not, however, concerned with the actual processing cost that is involved with carrying out the given (possibly optimal) policies. Thus, an optimal policy for the MDP may be found which maximizes the reward achieved by an agent, but which does not heed possible computational costs or limits imposed on an agent — this is in contrast to simple organisms which cannot afford a large informational bandwidth or minimal robots for which a suboptimal, but informationally cheap performance would be sufficient.

It is therefore the goal of the present paper to marry the MDP formalism with an information-theoretic treatment of the processing cost required by the agent (and the environment) to attain a given level of performance (in terms of rewards). To combine these disparate frameworks, we need to introduce notions from both information theory as well as from the theory of MDPs.

To limit the complexity of the present exposition, this paper will concentrate only on modelling action rewards; we will not address here its symmetric counterpart, namely (non-informational) costs of sensoric data acquisition. Furthermore, we will skim only the surface of the quite intricate relation of the present formalism with the framework of predictive information (e.g. in Sec. 5.3.2).

2 Rationale

An important remark about the present paper is that its core intention is the development of a general and expressive framework, and not a particularly efficient algorithm. In adopting the principled approach of information theory, we are here not concerned with producing an algorithm which would compare performance-wise with competitive state-of-the-art MDP learners (such as, e.g. Engel et al. 2003; Jung and Polani 2007); and neither are we concerned with proposing or investigating particular flavours of the perception-action architecture. Instead, the idea behind the application of information theory to the perception-action cycle is to open the path towards a new way of looking at the perception-action cycle with the hope that this will lead to new concepts, new insights and, possibly, new types of questions.

In particular, because of its universality, and because the framework of information theory has deep ramifications into many fields, including physics and statistical learning theory, it makes different architectures, models as well as scenarios comparable under a common language. This allows information theory to be applied across various and quite disparate domains of interest (such as e.g. robotics, language, or speech); in addition, it opens up a natural approach to bridge the gap between the study of artificial and of biological systems. Information theory gives rise to a rich and diverse set of theorems and results, far beyond its original application to communication theory. These results include the formulation of fundamental bounds for computational effort and/or power. Furthermore, information-optimal solutions ex-

hibit a significant array of desirable properties, among other being least biased or committed, or being maximally stable (see Sec. 8.2).

The particular slant that the present paper takes is essentially to consider the issue of optimal control in the guise of MDPs and expressing it in the language of information. This is not a mere equivalent re-expression of the control task. Rather, it adds a rich structural and conceptual layer to it.

In recent work, the classical task of optimal control has found an elegant reformulation in probabilistic and information-theoretic language: the control task is formulated as a probabilistic perturbation of a system by an agent, and the Bellman equation governing the optimal control solution can be expressed instead as a Kullback-Leibler divergence minimization problem (Todorov 2009); by identifying optimal control problems with Bayesian inference problems (Strens 2000), the efficient methods for graphical model inference become available for the computation of the optimal policies (Kappen et al. 2009). This can be generalized to consider directly the desired distribution of outcomes of the agent's action (Friston et al. 2006; Friston 2009).

In fact, this interpretation invites an even more general picture of the control problem: instead of specifying the reward structure externally, one can consider the intrinsic informational dynamics of an agent interacting with its environment. On the one hand, the study of the information flows in such a system gives important insights into the operation of the agent-environment system (Ay and Wennekers 2003; Lungarella and Sporns 2006; Ay and Polani 2008). On the other hand, one can obtain natural, intrinsically driven ("self-motivated", "reward-less") agent behaviours by optimizing information flows (Lungarella and Sporns 2005; Sporns and Lungarella 2006), predictive information (Ay et al. 2008) or the agent-external channel capacity of the perception-action cycle ("empowerment", Klyubin et al. 2005a, 2008). Because of its cyclical character, the interplay between agent-external and agent-internal information processing has been likened to an informational "Carnot Cycle" (Fry 2008), whose optimal thermal efficiency would find an informational analog in the concept of *Dual Matching*, which is essentially the *joint source-channel coding* proposed in (Gastpar et al. 2003).

Here, we will revisit the earlier mentioned MDP problem, and again we will employ an informational view. However, here, unlike in (Todorov 2009; Kappen et al. 2009), the informational picture will not be used to implement a Bayesian inference mechanism that realizes a solution to the Bellman equation — quite the opposite: we will, in fact, stick with the classical decision-theoretic Bellman approach to MDP. We will, however, combine this approach with a perspective which elevates the information used in the agent's decision process to the "first class object" of our discourse. Adopting this philosophy, we will see how the ramifications of the information approach will project into a wide array of disparate issues, ranging from the analogy between the perception-action cycle and Shannon's communication channel to the relation between the Dual Matching condition mentioned earlier ("Carnot optimality") and the perfectly adapted environment in Secs. 5.3.2 and 7.2.

3 Notation

3.1 Probabilistic Quantities

We will employ uppercase characters for random variables $X, Y, Z \ldots$, lowercase characters for concrete values $x, y, z \ldots$ that they assume and curved characters $\mathscr{X}, \mathscr{Y}, \mathscr{Z}$ for their respective domains. For simplicity we will assume that the domains are finite.

The probability that a random variable X assumes a value $x \in \mathscr{X}$ is denoted by $\Pr(X = x)$. However, to avoid unwieldy expressions, we will, by abuse of notation, write $p(x)$ with x being the lowercase of the random variable X in question. Occasionally, we will need to associate two different distributions with the same random variable domain $\mathscr{X}$; these cases will be clearly indicated, and the corresponding distributions will be denoted by the notation $p(x), \hat{p}(x), q(x) \ldots$ and similar. Where a random variable is subscripted, such as in X_t, a different index t will denote different variables.

3.2 Entropy and Information

We review some of the basic concepts of Shannon's information theory and notational conventions relevant for this paper. For a more complete discussion see e.g. (Cover and Thomas 1991).

Define the *entropy* $H(X)$ of a random variable X as

$$H(X) := -\sum_{x \in \mathscr{X}} p(x) \log p(x) \tag{1}$$

where the result is expressed in the unit of *bits* if the logarithm is taken with respect to base two; we assume the identification $0 \log 0 = 0$. Furthermore, in the following we will drop the summation domain $\mathscr{X}$ when the domain is obvious from the context. Also, we will write $H[p(x)]$ instead of $H(X)$ if we intend to emphasize the distribution p of X.

The entropy is a measure of uncertainty about the outcome of the random variable X before it has been measured or seen, and is a natural choice for this (Shannon 1949). The entropy is always nonnegative. It vanishes for a deterministic X (i.e. if X is completely determined) and it is easy to see (e.g. using the Jensen inequality) that $H[p(x)]$ is a convex functional over the simplex of probability distributions which attains its maximum $\log |\mathscr{X}|$ for the uniform distribution, reflecting the state of maximal uncertainty.

If a second random variable Y is given which is jointly distributed with X according to the distribution $p(x,y)$, then one can define the joint entropy $H(X,Y)$ trivially as the entropy of the joint random variable (X,Y). Furthermore, one can now define the *conditional entropy* as

$$H(X|Y) \stackrel{\Delta}{=} \sum_y p(y) H(X|Y=y) := -\sum_y p(y) \sum_x p(x|y) \log p(x|y) \,.$$

The conditional entropy measures the remaining uncertainty about X if Y is known. If X is fully determined on knowing Y, it vanishes. With the conditional entropy, one obtains the basic additivity property, or *chain rule* for the entropy,

$$H(X,Y) = H(Y) + H(X|Y) = H(X) + H(Y|X) \,. \tag{2}$$

This additivity of the entropy for (conditionally) independent variables can, in fact, be taken as the defining property of Shannon's entropy, as it uniquely determines it under mild technical assumptions.

Instead of the uncertainty that remains in a variable X once a jointly distributed variable Y is known, one can ask the converse question: how much uncertainty in X is resolved if Y is observed, or stated differently, how much *information* does Y convey about X. This gives rise to the highly important notion of *mutual information* between X and Y, which is expressed in the following equivalent ways,

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) \,. \tag{3}$$

It is non-negative and symmetric, and vanishes if and only if X and Y are independent. The mutual information turns out to play a major role in Shannon's source and channel coding theorems and other important ramifications of information theory.

A closely related, technically convenient and theoretically important quantity is the *relative entropy*, or *Kullback-Leibler divergence*: assume two distributions over the same domain $\mathscr{X}$, $p(x)$ and $q(x)$, where p is absolutely continuous with respect to q (i.e. $q(x) = 0 \Rightarrow p(x) = 0$). Then define the relative entropy of p and q as:

$$D_{\mathrm{KL}}[p||q] \stackrel{\Delta}{=} \sum_x p(x) \log \frac{p(x)}{q(x)} \,, \tag{4}$$

with the convention $0 \log \frac{0}{0} \stackrel{\Delta}{=} 0$. The relative entropy is a measure how much "compression" (or prediction, both in bits) could be gained if instead of an hypothesized distribution q of X, a concrete distribution p is utilized. It is the mean code length difference if $q(x)$ is assumed for the prior distribution of X but $p(x)$ is the actual distribution.

We mention several important properties of the relative entropy needed for the rest of the paper (for details see e.g. Cover and Thomas 1991). First, one has $D_{\mathrm{KL}}[p||q] \geq 0$ with equality if and only if $p = q$ everywhere (almost everywhere in the case of continuous domains $\mathscr{X}$). In other words, one can not do better than actually assuming the "correct" q. Second, the relative entropy can become infinite if for an outcome x that can occur with nonzero probability $p(x)$ one assumes a probability $q(x) = 0$. Third, the mutual information between two variables X and Y can be expressed as

$$I(X;Y) = D_{\mathrm{KL}}[p(x,y)||p(x)p(y)] \tag{5}$$

where we write $p(x)p(y)$ for the product of the marginals by abuse of notation. Basically, this interprets mutual information as how many bits about Y can be extracted from X if X and Y are not independent, but jointly distributed. If they are indeed independent, that value vanishes.

Furthermore, we would like to mention the following property of the relative entropy:

Proposition 1 *$D_{KL}[p||q]$, is convex in the pair (p,q).*

The first important corollary form this is the strong concavity of the entropy functional, i.e. that there is a unique distribution with maximum entropy in any (compact) convex subset of the simplex. Since the mutual information can be written as

$$I(X;Y) = D_{\mathrm{KL}}[p(x,y)||p(x)p(y)] = \mathbb{E}_y D_{\mathrm{KL}}[p(x|y)||p(x)]\,,$$

one can assert that $I(X;Y)$ is a concave function of $p(x)$ for a fixed $p(y|x)$, and a convex function of $p(y|x)$ given $p(x)$. This is relevant because it guarantees the unique solution of the two fundamental optimization problems of information theory. The first is $\min_{p(y|x)} I(X;Y)$, given $p(x)$ and subject to other convex constraints, i.e., the source coding or Rate-Distortion function. The second is $\max_{p(x)} I(X;Y)$, given $p(y|x)$ and possibly other concave constraints, i.e., the channel capacity problem.

4 Markov Decision Processes

4.1 MDP: Definition

The second major building block for the treatment of perception-action cycles in the present paper is the framework of Markov Decision Processes (MDPs). It is a basic model for the interaction of an organism (or an artificial agent) with a stochastic environment. We note that, as discussed in Sec. 8.2, the Markovicity of the model is not a limitation; while our exposition of the formalism employed below will indeed assume access to the full state of the system for the purposes of computation, it will fully retain the ability to model the agent's subjective view.

In the MDP definition, we follow the notation from (Sutton and Barto 1998). Given a state set $\mathscr{S}$, and for each state $s \in \mathscr{S}$ an action set $\mathscr{A}(s)$, an MDP is specified by the tuple $(\mathbf{P}^{s'}_{s,a}, \mathbf{R}^{s'}_{s,a})$, defined for all $s, s' \in \mathscr{S}$ and $a \in \mathscr{A}(s)$ where $\mathbf{P}^{s'}_{s,a}$ defines the probability that performing an action a in a state s will move the agent to state s' (hence

$$\sum_{s'} \mathbf{P}^{s'}_{s,a} = 1 \tag{6}$$

holds) and $\mathbf{R}^{s'}_{s,a}$ is the expected reward for this particular transition (note that $\mathbf{R}^{s'}_{s,a}$ depends not only on starting state s and action a, but also on the actually achieved final state s').

4.2 The Value Function of an MDP and its Optimization

The MDP defined in Sec. 4.1 defines a transition structure plus a reward. A *policy* specifies an explicit probability $\pi(a|s)$ to select action $a \in \mathscr{A}(s)$ if the agent in a state $s \in \mathscr{S}$. The policy π has the character of a conditional distribution, i.e. $\sum_{a \in \mathscr{A}(s)} \pi(a|s) = 1$.

Given such a policy π, one can now consider the cumulated reward for an MDP, starting at time t at state s_t, and selecting actions according to $\pi(a|s_t)$. This action gives a reward r_t and the agent will find itself in a new state s_{t+1}. Iterating this forever, one obtains the total reward for the agent following policy π in the MDP by cumulating the rewards over the future time steps:

$$\mathfrak{R}_t = \sum_{t'=t}^{\infty} r_{t'}\,, \tag{7}$$

where $\mathfrak{R}_t$ is the total reward accumulated[1] into the future starting at time step t.

Due to the Markovian nature of the model, this cumulated reward will depend only on the starting state s_t and, of course, the policy π of the agent, but not time, so that the resulting future expected cumulative reward value can be written $V^\pi(s)$ where we dropped the index t from the state s.

Usually in the Reinforcement Learning literature a distinction is made between *episodic* and *non-episodic* MDPs. Strictly spoken, Eq. (7) applies to the non-episodic case. For the episodic case, the sum in Eq. (7) is not continued to infinity, but stops at reaching so-called *goal* states. However, to unify the treatment, we will stick with Eq. (7) and rather suitably adapt $(\mathbf{P}^{s'}_{s,a}, \mathbf{R}^{s'}_{s,a})$.

A central result from the theory of Dynamic Programming and Reinforcement Learning is the so-called Bellman recursion. Even while the function V^π would be, in principle, computed by averaging over all possible paths generated through the policy π, it can be expressed through the *Bellman Equation* which is a recursive equation for V^π:

$$V^\pi(s) = \sum_{a \in \mathscr{A}(s)} \pi(a|s) \cdot \sum_{s' \in \mathscr{S}(s)} \mathbf{P}^{s'}_{s,a} \cdot \left[\mathbf{R}^{s'}_{s,a} + V^\pi(s')\right] \tag{8}$$

where the following assumptions hold:

1. a sums over all actions $\mathscr{A}(s)$ possible in s;
2. s' sums over all successor states $\mathscr{S}(s)$ of s;
3. for an episodic MDP, $V^\pi(s)$ is defined to be 0 if s is a goal state.

In the Reinforcement Learning literature, it is also common to consider the value function V being expanded per-action into the so-called Q function as to reflect

[1] To simplify the derivations, we will always assume convergence of the rewards and not make use of the usual MDP discount factor; in particular, we assume either episodic tasks or nonepisodic (continuing) tasks for which the reward converges. For further discussion, see also the remark in Sec. 8.2 on *soft policies*.

which action a in a state s achieves which reward:

$$Q^{\pi}(s,a) = \sum_{s' \in \mathscr{S}(s)} \mathbf{P}^{s'}_{s,a} \cdot \left[\mathbf{R}^{s'}_{s,a} + V^{\pi}(s') \right] . \tag{9}$$

In turn, $V^{\pi}(s)$ can be obtained from $Q^{\pi}(s,a)$ by averaging out the actions a with respect to the policy π.

To streamline and simplify the notation in the following, we assume without loss of generality that, in principle, all actions $\mathscr{A} \stackrel{\Delta}{=} \bigcup_{s \in \mathscr{S}} \mathscr{A}(s)$ are available at each state s and all states $\mathscr{S}$ could be potential successor states. For this, we will modify the MDP in a suitable way: actions $a \in \mathscr{A}$ that are not in the currently legal action set $\mathscr{A}(s)$ will be excluded from the policy, either by imposing the constraint $\pi(a|s) \stackrel{!}{=} 0$ or, equivalently, by setting $\mathbf{R}^{s'}_{s,a} \stackrel{!}{=} -\infty$. Similarly, transition probabilities $\mathbf{P}^{s'}_{s,a}$ into non-successor states s' are assumed to be 0. Furthermore, for an episodic task, a goal state s is assumed to be absorbing (i.e. $\mathbf{P}^{s'}_{s,a} = \delta_{s,s'}$ where the latter is the Kronecker delta) with transition reward $\mathbf{R}^{s'}_{s,a} = 0$.

The importance of the Bellman Equation (8) is that it provides a fixed-point equation for V^{π}. The value function which fulfils the Bellman Equation is unique and provides the cumulated reward for an agent starting at a given state and following a given policy. Where a value function V which is not a fixed point is plugged into the right side of Eq. (8), the equation provides a contractive map which converges towards the fixed point of the equation and thus computes the value of V^{π} for a given policy π. This procedure is called *value iteration*.

Finally, for an optimization of the policy — the main task of Reinforcement Learning — one then greedifies the policy, obtaining π', inserts it back into the Bellman Equation, recomputes $V^{\pi'}$ and continues the double iteration until convergence (Sutton and Barto 1998). This two-level iteration forms a standard approach for MDP optimization. In the current paper, we will develop an analogous iteration scheme which, however, will also cater for the informational cost that acting out an MDP policy entails. How to do this will be the topic of the coming sections.

5 Coupling Information with Decisions and Actions

The treatment of an MDP, fully specified by the tuple $(\mathbf{P}^{s'}_{s,a}, \mathbf{R}^{s'}_{s,a})$, is usually considered complete on finding an optimal policy. However, recent arguments concerning biological plausibility indicate that any hypothesized attempt to seek optimal behaviour by an organism needs to be balanced by the considerable cost of information processing which increasingly emerges as a central resource for organisms (Laughlin 2001; Brenner et al. 2000; Taylor et al. 2007; Polani 2009). This view has recently led to a number of investigations studying the optimization of informational quantities to model agent behaviour (Klyubin et al. 2005a; Prokopenko et al. 2006; Ay et al. 2008). The latter work creates a connection between *homeokinetic*

dynamics and the formalism of predictive information (Der et al. 1999; Bialek et al. 2001).

In view of the biological ramifications, the consideration of an *explicit* reward becomes particularly relevant. The question then becomes to find the optimal rewards that an organism can accumulate under given constraints on its informational bandwidth; or, more generally, the best possible trade-off between how much reward the organism can accumulate vs. how much informational bandwidth it needs for that purpose.

In this context, it is useful to contemplate the notion of *relevant information*. The concept of relevant information stems from the information bottleneck formalism (Tishby et al. 1999). To do so, one interprets an agents' actions as the relevance indicator variables of the bottleneck formalism. This provides an informational treatment of single-decision sequential (Markovian) decision problems (Polani et al. 2001, 2006), quantifying how much informational processing power is needed to realize a policy that achieves a particular value/reward level. Since we determine the relevance of information by the value it allows the agent to achieve, we speak in this context also of *valuable information*[2]. A related, but less biologically and resource-motivated approach is found in (Saerens et al. 2009). The topic of the present paper is a significant generalization of these considerations to the full-fledged perception-action cycle, and their treatment in the context of a generalized Bellman-type recursion.

5.1 Information and the Perception-Action Cycle

To apply the formalism of information theory, the quantities involved are best represented as random variables. Specifically in the context of an MDP, or more specifically of an agent acting in an MDP, one can make use of the formalism of (Causal) Bayesian Networks (Pearl 2000; Klyubin et al. 2004, 2007). Here, we introduce for completeness the general Bayesian Perception-Action Cycle formalism, before specializing it to the particular case studied in the present paper.

5.1.1 Causal Bayesian Networks

We briefly recapitulate the definition of *(Causal) Bayesian Networks*, also called *probabilistic graphical models* (see Pearl 2000). Causal Bayesian Networks provide a way to describe compactly and transparently the joint distribution of a collection of random variables.

[2] Valuable information is not to be confused with the *value of information* introduced in (Howard 1966). Serving similar purposes, it is conceptually different, as it measures the value difference attainable in a decision knowing vs. not knowing the outcome of a given random variable. Stated informally, it could be seen as "non information-theoretic conjugate" of valuable information.

A Bayesian network G over a set of random variables $\mathbf{X} \equiv \{X_1, \ldots, X_n\}$ is a Directed Acyclic Graph (DAG) G in which each vertex is annotated by (or identified by) names of the random variables X_i. For each such variable X_i, we denote by $\mathsf{Pa}[X_i]$ the set of parents of X_i in G and $\mathsf{Pa}[x_i]$ their values. We say that a distribution $p(\mathbf{x})$ is *consistent* with G, written $p \sim G$, if it can be factored in the form:

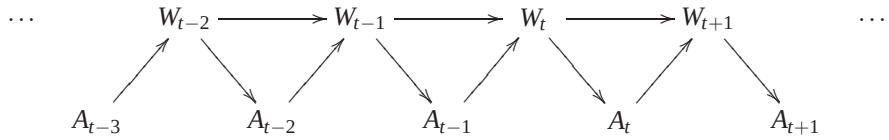
$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid \mathsf{Pa}[x_i]) .$$

(To simplify notation, we adopt the canonical convention that a conditional distribution conditioned against an empty set of variables is simply the unconditional distribution).

5.1.2 Bayesian Network for a Reactive Agent

To see how Bayesian Networks can be applied to model an agent operating in a world, consider first a minimal model of a reactive agent. The agent carries out actions A after observing the state W of the world to which it could, in principle, have full access (this assumption will be revisited later). Such an action, in turn, transforms the old state of the world into a new world state.

In different states, typically different actions will be taken, and if one wishes to model the behaviour of the agent through time, one needs to unroll the cycle over time. This leads to the Causal Bayesian Network which is shown in Eq. (10); here the random variables are indexed by the different time steps $\ldots t-3, t-2, \ldots, t, t+1 \ldots$. Each arrow indicates a conditional dependency (which can also be interpreted as causal in the sense of Pearl). For instance, at time t, the state of the world is W_t which defines the (possibly probabilistic) choice of action A_t immediately (no memory); this action, together with the current state of the world determines the next state of the world W_{t+1}, etc.

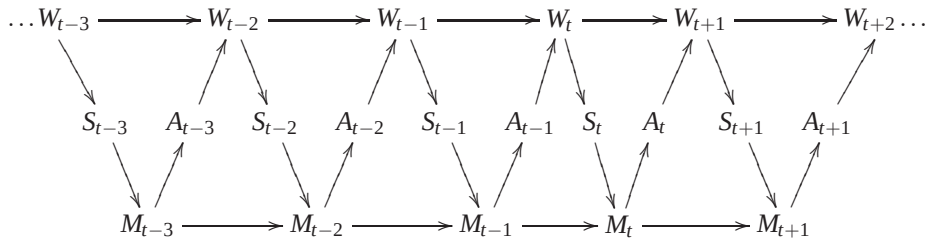


(10)

Note that in this model there is a priori no limitation on the arrow from W_t to A_t, i.e. on $p(a_t|w_t)$, and the agent could theoretically have full access to the state W_t.

5.1.3 Bayesian Network for a General Agent

After the simplest of the cases, consider a significantly more general case, the perception-action cycle of an agent with sensors and memory, as in Eq. (11).



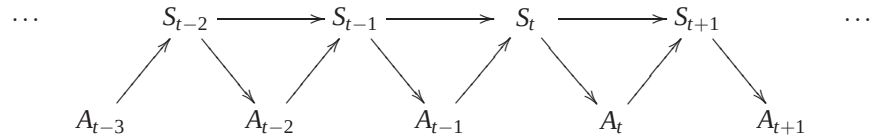
(11)

Here, W_t represents again the state of the world at time t, A_t the actions; additionally, one has S_t as the sensor and a memory variable M_t. The sensor variable S_t models the limited access of the agent to the environment, the memory variable allows the agent to construct an internal model of the external state that can depend on earlier sensoric observations.

A few notes concerning this model:

1. it is formally completely symmetric with respect to an exchange of agent and environment — the state of the environment corresponds to the memory of the agent; the interface between environment and agent is formed by the sensors and actuators (the only practical difference is that the arrows inside the agent are amenable to adaptation and change, while the environmental arrows are slow, difficult and/or expensive to modify, but this is not reflected in the model skeleton).
2. from the agent's point of view, the sensor variable essentially transforms the problem into a POMDP; however, in the Bayesian Network formalism, there is nothing special about the limited view of the agent as opposed to an all-knowing observer. Where necessary, the eagle's eye perspective of an all-knowing observer can be assumed. In particular in conjunction with informational optimization, this allows one to select the level of informational transparency in the network between various nodes, in particular to study various forms and degrees of access to world information that the agent could enjoy, such as e.g. in sensor evolution scenarios (Klyubin et al. 2005b);
3. in the following, we will consider only agents which, in principle, could have full access to the world state. In other words, we will collapse W and S into the same variable (we postpone the discussion of the full system to a future study);
4. we note furthermore that modeling a memory implies that one wishes to separately consider the informational bandwidth of the environment and that of the agent; else the environment itself could be considered as part of the agent's memory. In this paper, we are indeed interested in the information processed in the complete agent-environment system and not the individual components. Here, we will therefore consider a reactive agent without a memory. However, this is not a limitation in principle, as the approach presented is expected to carry over naturally to models of agents with memory.

With these comments, for the current paper, we finally end up at the following diagram which will form the basis for the rest of the paper:



(12)

which is essentially diagram (10), but where environmental state and sensor state have been collapsed into the same random variable S.

5.2 Actions as Coding

To obtain an intuition how actions can be interpreted in the language of coding, represent a given MDP as a graph[3] with transition edges between a state s and its successor state s'. In this graph, the edges are labeled by the actions a available in the given states (as well as the transition probabilities and rewards). Multiple edges can connect a given state s and its successor state s' as long as they differ by their action label, in other words differing actions may lead to the same transition.

From this picture, it is easy to see that any MDP can be modified without loss of generality in such a way that the actions have the same cardinality in every state. In the simplest case, we can always make this cardinality 2 and all the actions/decisions binary. To see this, consider the standard example of a grid-world, or a maze. As explained above, describe such a world by a (directed) graph with vertices of various degrees. Replace now every junction (which could have an arbitrary number of possible action choices) by a roundabout. In a roundabout, every decision becomes binary: continue on the roundabout or take the turn. Thereby, the decisions/actions become binary (left/right), without changing anything else in the problem.

More generally, this can be done by transforming any complex decision into a binary decision tree. Thus, long decision/action sequences are encoded as long bit sequences. It is important to note, though, that each such bit can have entirely different semantic context, depending on its location on the graph.

Assume now for this section that our MDP is made of only binary decisions and/or actions, denoted by $a_j \in \{0,1\}$, and treat decision sequences as finite binary strings $a_1 a_2 \ldots a_k \equiv \mathbf{a}$. Denote furthermore by l the standard string length function operating on these strings: $l(\mathbf{a}) \equiv l(a_1 a_2 \ldots a_k) \stackrel{\Delta}{=} k$.

We can now draw simple direct analogies between binary decision sequences and binary codes in communication. We assume that there exist finite sequences of actions (at least one) which deterministically connect any state s with any other state s' (the MDP is simply connected, and deterministic). Denote by $\mathbf{a}_{s,s'}$ such a (binary)

[3] This is a transition graph and should not be confused with the Bayesian Network Graph.

action sequence, and by $l(\mathbf{a}_{s,s'})$, its length. The following properties either follow immediately or are easy to verify:

- Unique decoding: a binary decision sequence uniquely determine the end-state s', given the starting state s.
- Concatenation: for any three states s_1, s_2, s_3, the concatenation of $\mathbf{a}_{s_1,s_2} \circ \mathbf{a}_{s_2,s_3}$ is a sequence that connects s_1 to s_3.
- Minimal length: for any two states s, s' we can define $l_{s\to s'} \stackrel{\Delta}{=} \min_{\mathbf{a}_{s,s'}} l(\mathbf{a}_{s,s'})$. We will call this minimal length between s and s' the *decision complexity* of the path $s \to s'$ in the MDP.

Now, in general, while decision/action sequences have similarities to codes as descriptions of trajectories on a graph, there are also important differences. First, in general it is not always the case that the goal of an MDP is reaching a particular target state (or a set of target states), as in the case of the maze. Sometimes the optimal solution for an MDP is distributed and more diffused, or it involves an extended behaviour of potentially infinite duration as is required in optimal control, wealth accumulation, or survival. In particular, the required action trajectory may turn out not to be deterministic, or not finite, or neither. Moreover, if the MDP is noisy (i.e. $\mathbf{P}^{s'}_{s,a}$ is not deterministic), a given action sequence does not necessarily determine a unique trajectory of states, but rather a distribution of state sequences.

This poses no problem though, since, as in standard information theory, above notion of decision complexity can be smoothly generalized to still have meaning in the probabilistic or non-finite action sequence: for this purpose one defines it to be the *expected* number of binary decisions, or bits of information, required by the agent in the future to achieve a certain expected cumulated reward value $V^\pi(s_t)$ if one starts out at a particular state s_t at a given time t in the MDP. The decision complexity is essentially a generalization of the notion of relevant information (mentioned earlier in Sec. 5) from individual actions to complete action sequences starting at the current time and extending into the future.

Furthermore, one can extend the information processed by the agent alone to the information processed by the whole system, encompassing both agent and environment. Hereby, one moves from decision complexity to *process information*. This quantity is computed for the whole agent-environment system, beginning with the current state (and agent action) and extended towards the open-ended future of the system. Strictly spoken, we only consider the process information towards the future, not the past. In analogy to the term "cost-to-go", we will speak of *process information-to-go*, or simply *information-to-go* $\mathfrak{I}^\pi(s_t, a_t)$ which is computed specifying a given starting state s_t and an initial action a_t and accumulating information-to-go into the open-ended future.

The information-to-go is the information needed (or missing) to specify the future states and actions relative to some prior knowledge about the future. One component of this information is the due to uncertainty in the decisions; the other is the information that is "given" or "processed" by the environment in the state transitions. For discrete states and actions we can think of it as the future state-action entropy, conditioned on the current state-action pair. More generally, one would use

the Kullback-Leibler divergence D_{KL} of the actual distribution relative to the prior distribution. This is actually implements an "informational regret" which generalizes the simple conditional entropy (and is better-behaved in the case of continuous states and actions). We define the notion of information-to-go formally in Sec. 6.2, Eq. (20).

Let us emphasize again that the philosophy of the present approach is intimately related to the concept of relevant information (Tishby et al. 1999; Polani et al. 2001, 2006) which quantifies the minimal informational cost for individual decisions an agent needs to take to achieve a given future expected cumulated reward (i.e. value). A difference between relevant information and the present study is that here we consider the information processed by whole system, not just the agent. However, by far the most important distinction is that we generalize this concept to include the information to be processed by the system not just for one time step, but over the *whole* period of the run and thus through multiple cycles of the perception-action cycle projected into the future.

The intimate relation between the theories of information and of coding would suggest an interpretation of the formalism in terms of coding. However, statements in coding theory are typically of asymptotic nature. Now, in general, a run of an agent through an MDP does not need to be of infinite length, not even on average: finite-length runs are perfectly possible. To reconcile this conflict, consider the following two interpretations which we expect to be able to recover the asymptotics required for a coding-theoretic interpretation:

1. runs are restarted after completion, infinitely often, thereby extending the MDP into an infinite future;
2. assume that the total available information processing power of the agent is pooled and shared by a large number of separate decision processes; a particular decision process will utilize a certain amount of information processing bandwidth at a given time, and the information-theoretic formalism then describes the usage averaged over all processes (weighted with their probability of occurring). This second view introduces an ensemble interpretation of the formalism.

We believe that both, the infinite length run and the ensemble interpretation allow a connection of the present formalism to coding theory and provide tight bounds in the asymptotic case. This question will be pursued further in the future, but is outside the remit of the present paper.

5.3 Information-To-Go

To prepare the ground for the later technical developments, in the present section, we first give a high-level outline of the coming discussions.

5.3.1 A Bellman Picture

The information-to-go, $\Im^{\pi}(s_t, a_t)$, is an information-theoretic quantity that is associated with every state-action pair, in analogy to the $Q(s_t, a_t)$ function which can be considered as the *value- (or reward-)to-go* from Reinforcement Learning. The information-to-go quantifies how many bits one needs on average to specify the future state-action sequence in an MDP (or its informational regret) relative to a prior.

In Sec. 6 we shall suggest a principled way of generating $\Im^{\pi}$-optimal soft-max policies. Strikingly, these solutions obey a Bellman-type recursion equation, analogous to Reinforcement Learning; they involve a locally accumulated quantity which can be interpreted as *(local) information gain*, the information provided by the environment associated with each state transition and action, in analogy to the local reward in MDPs.

The local information gain will be shown to be composed of two natural terms. The first is the information measuring the environmental response to an action, i.e. the information processed in the transition as the system moves to a new state; it is determined by the MDP probabilities $\mathbf{P}^{s'}_{s,a}$. The second term is the information *required* by the agent to select the valuable actions and is determined by the policy $\pi(a|s)$.

The combination of the Value and Information Bellman equations gives a new Bellman-like equation for the linear (Lagrangian) combination of the two, the *free energy* of the MDP, denoted $F^{\pi}(s_t, a_t, \beta)$, which reflects the optimal balance between the information-to-go and value-to-go achieved. For given transition probabilities, $\mathbf{P}^{s'}_{s,a}$, and given policy $\pi(a|s)$, we can calculate the free energy for every state/action pair by solving the Bellman equation, for any given trade-off between information-to-go and value-to-go. This essentially implements a novel variant of the rate-distortion formalism which applies to the value-information trade-off of MDP sequences.

5.3.2 Perfectly Adapted Environments

This formulation gives rise to a second new insight, namely the characterization of the *perfectly adapted environment* by further minimizing the free energy with respect to the MDP probabilities, $\mathbf{P}^{s'}_{s,a}$. The MDP transition probabilities that minimize the free energy are shown to be exponential in the reward, $\mathbf{R}^{s'}_{s,a}$. In that particular case all the information about the future is valuable and the optimal policy turns out to be also the one that minimizes statistical surprises.

Perfectly adapted environments form another family of scenarios (besides the classical model of Kelly gambling, Kelly 1956) with the property that the maximization of information about the future is equivalent to the maximization of the value of the expected reward. In general, this is not the case: rather, the current state of the system will provide valuable (relevant) as well as non-valuable information about the future. Non-valuable (irrelevant) information about the future is the information that can be safely ignored (in a bottleneck sense) without affecting the future

expected reward. It is the valuable (relevant) information only which affects the future reward. In the special cases of both Kelly gambling as well as our case of the perfectly adapted environment, however, all information is valuable — maximizing the future information is equivalent to maximizing the expected future reward. When the bandwidth for future information is limited, this leads to a suboptimal trade-off with respect to achievable future expected reward.

The interest in studying perfectly adapted environments stems from the fact that they provide the key for linking predictive information with task-relevant information. It has been hypothesized that living organisms maximize the predictive information in their sensorimotor cycle and this hypothesis allows to derive a number of universal properties of the perception-action cycle (Bialek et al. 2001; Ay et al. 2008), and, if this hypothesis has merit, this would imply an interpretation of organismic behaviour in terms of (Kelly) gambling on the outcome of actions. On the other hand, actual organismic rewards could in principle be structured in such a way that much of the predictive information available in the system would turn out to be irrelevant to select the optimal action; the distinction between relevant and irrelevant information provides a characterization of an organism's niche in its information ecology.

However, under the assumption that information acquisition and processing is costly for an organism (Laughlin 2001; Polani 2009), one would indeed expect a selection pressure for the formation of sensors that capture just the value-relevant component of the predictive information, but no more. A step further goes the hypothesis that, over evolutionary times, selective pressure would even end up realigning the reward and the informational structures towards perfectly adapted environments. Although here we are concentrating only on their theoretical implications, it should be mentioned that all these hypotheses imply quantitative and ultimately experimentally testable predictions.

5.3.3 Predictive Information

We assume in Eq. (12) that the agent has full sensoric access to the state of the world. This is a special case of the more general case where the information-to-go is the information that the agent at the current time has on the future of the system and which is extracted from past observations. One implication of this assumption is that the future information of the organism is bounded by the *predictive information* (defined e.g. in Shalizi and Crutchfield 2002; Bialek et al. 2001) of the environment (Bialek et al. 2007). In Fig. 3 from (Bialek et al. 2007), information about the past, the future, adaptive value and resources are put in relation to each other, and the predictive information (i.e. the information which the past of the system carries about the future) corresponds to its 3rd quadrant. Note that this is not the full predictive information, but only the valuable (relevant) component of the predictive information, in the sense that it identifies the information necessary to achieve a given value in a given reward structure. Its supremum is the total valuable information that the environment carries about the future. The organism cannot have more future valuable

information than is present in the environment, and will, usually, have less since it is bounded by metabolic, memory or computational resources. As was shown by (Bialek et al. 2001), for stationary environments the predictive information grows sub-linearly with the future horizon window (it is sub-extensive). On the other hand, for stationary environments and ergodic MDPs the information-to-go grows linearly with the horizon (future window size), for large enough windows.

5.3.4 Symmetry

We wish to attract attention to a further observation: we commented already in Sec. 5.1.3 that the Bayesian Network is symmetric with respect to environment and agent — but, in addition, the Bayesian Network is also structurally symmetric with respect to an interchange of past and future whereby the role of sensing and acting is switched. This structural symmetry is reflected in the essential interchangeability of the past and future axes in the 3rd quadrant of Fig. 3 in (Bialek et al. 2007). To complete the symmetry, we would need to additionally introduce a sensoric cost in analogy to the actuatoric reward which is already implemented in the present paper. Of course, the symmetry is only structural; in the framework, past and future are of course asymmetric, since we compress the past and predict the future — i.e. we minimize the information about the past and maximize the information about the future (and see e.g. also Ellison et al. 2009). Likewise, the symmetry between environment and agent is only structural, but the flexibility and the characteristic dynamics will in general differ strongly between the environment and the agent.

5.4 The Balance of Information

The relevance of the informational treatment of the perception-action cycle arises to some degree from the fact that information, while not a conserved quantity, observes a number of consistency and bookkeeping laws. Other such laws are incarnated as informational lower or upper bounds. Always implicit to such considerations is of course the fact that, in the Shannon view, the source data can — in principle — be coded, transmitted through the channel in question and then suitably decoded to achieve the given bounds.

We note that the multi-staged treatment of communication channels separating source and channel coding as espoused in the classical presentation by Shannon is not necessarily the most biologically relevant scenario. For instance, it was noticed by Berger (2003) that biology might be using non-separable information in the sense of using joint source-channel coding (Csiszár and Körner 1986). This view is of particular interest due to the discovery of the general existence of optimally matched channels (Gastpar et al. 2003). The simplicity and directness they afford, suggests that such channels may have relevance in biology. Specifically, with these advantages, it is conceivable that biological perception-action cycles would profit from

co-evolving all their components towards optimally matched channels; this particularly since biological channels are likely to have had sufficient time and degrees of freedom to evolve optimally matched channels.

If this hypothesis is valid, biological channels and perception-action cycles will not just strive to be informationally optimal, but also fulfil the additional constraints imposed by the optimally matched channel condition. In a metaphorical way, this hypothesis corresponds to an "impedance match" or a balance criterion for information flowing between the organism and the environment in the cycle.

The optimal match hypothesis, for one, contributes to the plausibility of the informational treatment for the understanding of biological information processing; in addition, it provides a foundation for predictive statements, both quantitative and structural. It is beyond the scope of the present paper to dwell on these ramifications in detail. However, it should be kept in mind that these are an important factor behind the relevance of informational bookkeeping principles for the studies of the perception-action cycle.

With these preliminaries in place, the present section will now review some elementary information-theoretical bookkeeping principles. The reader is already acquainted with them is invited to only lightly skim this section for reference.

5.4.1 The Data Processing Inequality and Chain Rules for Information

Consider a simple linear *Markov Chain*, a special case of a Bayesian Network, consisting of three random variables: $U \rightarrow X \rightarrow Y$. Then, the *Data Processing Theorem* states that Y can not contain more information about U than X, formally

$$I(X;U) \geq I(U;Y)\,.$$

In other words, Y can at most reflect the amount of information about U that it acquires from X, but no more than that. While information cannot grow, it can be lost in such a linear chain. However, to reacquire lost information, it would need to feed in from another source. The insight gained by the data processing inequality can furthermore be refined by not just quantifying, but actually identifying the information from a source variable that can be extracted downstream in a Markov Chain. One method to do so is, for instance, the Information Bottleneck (Tishby et al. 1999).

As a more general case, consider general finite sequences of random variables, $X^n \equiv (X_1, X_2, ..., X_n)$. From Eq. (2) it is easy to see that one has

$$H(X_1, X_2, ..., X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + ... + H(X_n|X^{n-1})\,. \quad (13)$$

In the case of finite — say m-th order — Markov chains, this simplifies drastically. Here, a variable X_k is screened by the m previous variables $X_{k-m}, \ldots, X_{k-1}$ from any preceding variable, that is, a conditional entropy simplifies according to

$$H(X_k|X^{k-1}) \equiv H(X_k|X_1, \ldots, X_{k-1}) = H(X_k|X_{k-m}, \ldots, X_{k-1})$$

(without loss of generality assume $k > m$, or else pad by empty random variables X_i for $i \leq 0$).

Similarly to Eq. (13) one has for the mutual information with any additional variable Y the relation

$$I(X_1, X_2, ..., X_n; Y) = I(X_1; Y) + I(X_2; Y|X_1) + I(X_3; Y|X_2, X_1) + ... + I(Y; X_n|X^{n-1}), \tag{14}$$

where the *conditional mutual information* is naturally defined as $I(X;Y|Z) \equiv H(X|Z) - H(X|Y,Z)$. The conditional mutual information can be interpreted as the information shared by X and Y once Z is known.

We have seen in Eq. (5), mutual information can be expressed in terms of the Kullback-Leibler divergence. Thus, the chain rule of information is, in fact, a special case of the chain rule of the Kullback-Leibler divergence (see also Cover and Thomas 1991):

$$\begin{aligned} D_{\mathrm{KL}}[p(x_1,\ldots,x_n)||q(x_1,\ldots,x_n)] = D_{\mathrm{KL}}[p(x_1)||q(x_1)] + D_{\mathrm{KL}}[p(x_2|x_1)||q(x_2|x_1)] + \ldots \\ + D_{\mathrm{KL}}[p(x_n|x^{n-1})||q(x_n|x^{n-1})] \end{aligned} \tag{15}$$

with the conditional Kullback-Leibler divergence defined as

$$D_{\mathrm{KL}}[p(y|x)||q(y|x)] \stackrel{\Delta}{=} \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)}. \tag{16}$$

5.4.2 Multi-Information and Information in Directed Acyclic Graphs

A multivariate generalization of the mutual information is the multi-information. It is defined as

$$\mathbf{I}[p(X)] = \mathbf{I}(X_1, X_2, ..., X_n) = D_{\mathrm{KL}}[p(x_1, x_2, ..., x_n)||p(x_1)p(x_2)...p(x_n)]. \tag{17}$$

There are various interpretations for the multi-information. The most immediate one derives from the Kullback-Leibler representation used above: in this view, the multi-information measures by how much more one could compress the random variable $(X_1, X_2, \ldots, X_n)$ if one treated it as a joint random variable as opposed to a collection of independent random variables $X_1, X_2, \ldots, X_n$. In other words, this is a measure for the overall dependency of these variables that could be "squeezed out" by joint compression. The multi-information has proven useful in a variety of fields, such as the analysis of graphical models (see e.g. Slonim et al. 2006).

Proposition 2 *Let* $\mathbf{X} = \{X_1, \ldots, X_n\} \sim p(\mathbf{x})$*, and let G be a Bayesian network structure over* $\mathbf{X}$ *such that* $p \sim G$*. Then*

$$\mathbf{I}[p(\mathbf{x})] \equiv \mathbf{I}(\mathbf{X}) = \sum_i I(X_i; \mathsf{Pa}[X_i]).$$

That is, the total multi-information is the sum of "local" mutual information terms between each variable and its parents (related additivity criteria can be formulated for other informational quantities, see also Ay and Wennekers 2003; Wennekers and Ay 2005).

An important property of the multi-information is that it is the cross-entropy between a model multivariate distribution and the "most agnostic", completely independent prior over the variables; therefore, it can be used to obtain finite sample generalization bounds using the PAC-Bayesian framework (McAllester 1999; Seldin and Tishby 2009).

6 Bellman Recursion for Sequential Information Processing

The language and the formalisms needed to formulate the central result of the present paper are now in place. Recall that we were interested in considering an MDP not just in terms of maximized rewards but also in terms of information-to-go.

We consider complete decision sequences and compute their corresponding information-to-go during the whole course of a sequence (as opposed to the information processed in each single decision, as in Polani et al. 2001, 2006). We will combine the Bayesian Network formalism with the MDP picture to derive trade-offs between the reward achieved in the MDP and the informational effort or cost required to achieve this reward. More precisely, unlike in the conventional picture of MDP where one essentially seeks to maximize the reward no matter what the cost of the decision process, we will put an informational constraint on the cost of the decision process and ask what the best reward is which can be achieved under this processing constraint.

It turns out that the resulting formalism resembles closely the Bellman recursion which is used to solve regular MDP problems, but it applies instead to informational quantities. This is in particular interesting since informational costs are not extensive as MDP rewards are (Bialek et al. 2001).

Before we proceed to introduce the algorithm, note that the reward is only associated with the agent's choice of actions and the ensuing transitions. Thus, only that information about the future is relevant here which affects the rewards. In turn, the component of entropy of the future which is not going to affect the reward can be ignored. Basically, this is a "rate-distortion" version of the concept of statistical sufficiency: we are going to ignore the variability of the world which does not affect the reward.

6.1 Introductory Remarks

Consider now the stochastic process of state-action pairs

$$S_t, A_t, S_{t+1}A_{t+1}, S_{t+2}A_{t+2}, \ldots, S_{t+n}A_{t+n}, \ldots$$

where the state-action pairs derive from an MDP whose Bayesian Network corresponds to Eq. (12), beginning with the current state S_t and action A_t. The setup is similar to (Klyubin et al. 2004; Still 2009).

We reiterate the argument from Sec. 5.1.3, point 2 and emphasize once more that for our purposes, it is no limitation to assume that the agent has potentially unlimited access to the world state and we therefore can exclusively consider MDPs instead of POMDPs. The information/value trade-off will simply find the best possible pattern of utilization of information.

For a finite informational constraint, this still implicitly defines a POMDP, however one that is not defined by a particular "sensor" (i.e. partial observation) structure, but rather by the quantitative limits of the informational bandwidth. The formalism achieves the best value for a given informational bandwidth in the sense that no other transformation of the MDP into a POMDP utilizing the same information processing bandwidth will exceed the optimal trade-off solution with respect to value. In the following, we will thus consider the system from an eagle's eye perspective where for the purposes of the computation we have — in principle — access to all states of the system, even if the agent itself (due to its information bandwidth constraints) may not.

To impose an explicitly designed POMDP structure (e.g. incorporating physical, engineering or other constraints), one could resort to the extended model from Eq. (11) instead. The latter incorporates sensors (i.e. explicit limitations to what the agent can access from the environment) as well as memory. Considering the latter turns the perception-action cycle into a full formal analogy of Shannon's communication channel. In this case, however, one typically needs to include also the agent memory into the picture. For such a scenario, preliminary results indicate that informational optimality criteria have the potential to characterize general properties of information-processing architectures in a principled way (van Dijk et al. 2009).

Concludingly, the formalism introduced here is not limited to reactive agents and in future work we will extend it to memory-equipped agents. Here, however, we limit ourselves to reactive agents, as these already represent an important subset of the systems of interest and provide a transparent demonstration of the central ideas of our approach.

6.2 Decision Complexity

Assume that, at time t, the current state and action are given: $S_t = s_t, A_t = a_t$. The distribution of successor states and actions in the following time steps $t+1, t+2, \ldots$ is given by $p(s_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}, \ldots | s_t, a_t)$. We assume now a fixed prior on the distribution of successive states and actions: $\hat{p}(s_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}, \ldots)$.

Define now the process complexity as the Kullback-Leibler divergence between the actual distribution of states and actions after t and the one assumed in the prior:

$$\Im^{\pi}(s_t,a_t) \stackrel{\Delta}{=} \mathbb{E}_{p(s_{t+1},a_{t+1},s_{t+2},a_{t+2},\ldots|s_t,a_t)} \log \frac{p(s_{t+1},a_{t+1},s_{t+2},a_{t+2},\ldots|s_t,a_t)}{\hat{p}(s_{t+1},a_{t+1},s_{t+2},a_{t+2},\ldots)}. \quad (18)$$

$\Im^{\pi}(s_t,a_t)$ measures the informational regret of a particular sequence relative a prior probability for the sequence. The prior encodes all information known about the process which can range from a state of complete ignorance up to a full model of the process (in which case $\Im^{\pi}(s_t,a_t)$ would vanish).

However, we want to consider priors which are simpler than the full MDP model. Of particular interest are those where the components of process $S_{t+1}A_{t+1}, S_{t+2}A_{t+2}, \ldots$ are independent, i.e. where the prior has the form

$$\hat{p}(s_{t+1},a_{t+1},s_{t+2},a_{t+2},\ldots) = \hat{p}(s_{t+1})\hat{\pi}(a_{t+1})\hat{p}(s_{t+2})\hat{\pi}(a_{t+2})\ldots$$

where all "hatted" distributions are the individual priors on the respective random variables (we denote the priors for the actions by $\hat{\pi}$ instead of $\hat{p}$ for reasons that will become clearer below, see e.g. Eq. (19)). With such a choice of the prior, $\Im^{\pi}(s_t,a_t)$ becomes a measure for the interaction between the different steps in the decision cascade[4].

Selecting the priors $\hat{p}(s_{t+1}), \hat{\pi}(a_{t+1}), \hat{p}(s_{t+2}), \hat{\pi}(a_{t+2}), \ldots$ beforehand and independently from the MDP corresponds to the most agnostic assumption. Another specialization is the *stationarity* assumption that the random variables $S_{t+1}, S_{t+2}, \ldots$ and $A_{t+1}A_{t+2}, \ldots$ are i.i.d. and share the same state distributions $\hat{p}(s_{t+1}), \hat{p}(s_{t+2}), \ldots$ and action distributions $\hat{\pi}(a_{t+1}), \hat{\pi}(a_{t+2}), \ldots$.

For our purposes, it is useful to mention the criterium of *consistency*. Consistency can be total or partial. *Total consistency* means that $\hat{p}(s_{t+1}), \hat{\pi}(a_{t+1}), \hat{p}(s_{t+2}), \hat{\pi}(a_{t+2}), \ldots$ result from the marginalization of the *total* original distribution which itself is consistent with the Bayesian Network Eq. (12). In the case of total consistency, $\Im^{\pi}(s_t,a_t)$ becomes the multi-information between the state/action variables $S_{t+1}A_{t+1}, S_{t+2}A_{t+2}, \ldots$ throughout the sequence.

On the other hand, *partial consistency* means that only parts of the relations in the Bayesian Network are respected in forming the factorization.

In the present paper, we will use close to minimal assumptions: we assume stationarity with partial consistency, where the state distributions $\hat{p}(s_{t+1}), \hat{p}(s_{t+2}), \ldots$ are the same for all times, and the action distributions are consistent with them via the policy π which we assume constant over time for all t:

$$\hat{\pi}(a_t) = \sum_{s_t \in \mathscr{S}} \pi(a_t|s_t) \cdot \hat{p}(s_t). \quad (19)$$

The prior $\hat{p}(s_t)$ is chosen as uniform distribution over the states for all t.

Stronger consistency assumptions, such as requiring the $\hat{p}(s_{t+1}), \hat{p}(s_{t+2}), \ldots$ to respect the transition probabilities $p(s_{t+1}|s_t,a_t)$ in the Bayesian Network (we call this *ergodic stationarity* in the special case of $\hat{p}(s_{t+1}), \hat{p}(s_{t+2}), \ldots$ being identical

[4] Note that, unless stated otherwise, we always imply that the distributions $\hat{p}(s_{t+1}), \hat{p}(s_{t+2}), \ldots$ as well as $\hat{\pi}(a_{t+1}), \hat{\pi}(a_{t+2}), \ldots$ can be different for different t

distributions) will be considered in the future, but are outside of the remit of the present paper.

With above comments, the information-to-go will be defined in the following as the Kullback-Leibler divergence of the of the future sequence of states and actions, starting from s_t, a_t, with respect to stationary prior state distributions over the state sequence $\hat{p}(s_{t+1}), \hat{p}(s_{t+2}), \ldots$ and policy-consistent (Eq. (19)) action distributions $\hat{\pi}(a_{t+1}), \hat{\pi}(a_{t+2}), \ldots$:

$$\Im^{\pi}(s_t, a_t) \stackrel{\Delta}{=} \mathbb{E}_{p(s_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}, \ldots | s_t, a_t)} \log \frac{p(s_{t+1}, a_{t+1}, s_{t+2}, a_{t+2}, \ldots | s_t, a_t)}{\hat{p}(s_{t+1})\hat{\pi}(a_{t+1})\hat{p}(s_{t+2})\hat{\pi}(a_{t+2})\ldots}. \quad (20)$$

The interpretation of this quantity is as follows: $\Im^{\pi}(s_t, a_t)$ measures the informational cost for the system to carry out the policy π, starting at time t into the indefinite future with respect to the prior. In general, this quantity will grow with the length of the future. It measures how much information is processed by the whole agent-environment system in pursuing the given policy π. This quantity can also be interpreted as the number of bits that all the states and actions share over the extent of the process as opposed to the prior. One motive for studying this quantity is that it provides important insights about how minimalistic agents can solve external tasks under limited informational resources.

The central result of the present paper is that the optimization of V^{π} under constrained information-to-go $\Im^{\pi}(s_t, a_t)$, although encompassing the whole future of the current agent, can be computed through a one-step-lookahead recursion relation; moreover, this recursion relation closely mirrors the Bellman recursion used in the value iteration algorithms of conventional Reinforcement Learning.

6.3 Recursion equation for the MDP Information-To-Go

We obtain a recursion relation for this function by separating the first expectation from the the rest. With Proposition 2 in the context of Eq. (12), it is easy to see that one has

$$\Im^{\pi}(s_t, a_t) = \mathbb{E}_{p(s_{t+1}, a_{t+1} | s_t, a_t)} \left[\log \frac{p(s_{t+1} | s_t, a_t)}{\hat{p}(s_{t+1})} + \log \frac{\pi(a_{t+1} | s_{t+1})}{\hat{\pi}(a_{t+1})} + \Im^{\pi}(s_{t+1}, a_{t+1}) \right], \quad (21)$$

with $p(s_{t+1}, a_{t+1} | s_t, a_t) = p(s_{t+1} | s_t, a_t)\pi(a_{t+1} | s_{t+1})$ (for more general Bayesian graphs, more general statements can be derived).

The terms associated with the first state-action-transition,

$$\Delta I_{s_t, a_t}^{s_{t+1}} = \log \frac{p(s_{t+1} | s_t, a_t)}{\hat{p}(s_{t+1})} + \log \frac{\pi(a_{t+1} | s_{t+1})}{\hat{\pi}(a_{t+1})} \quad (22)$$

can be interpreted as the *information gain* associated with this transition. In this recursion, the information gain takes on the role of the local reward, in complete analogy with the quantity $\mathbf{R}^{s'}_{s,a}$ from Reinforcement Learning.

Information gain-type quantities appear as natural Lyapunov functions for Master Equations and in the informational formulation of exploration and learning problems (Haken 1983; Vergassola et al. 2007). Note that the quantities in Eq. (22) can be both positive as well as negative (even if the prior is marginal, e.g. Lizier et al. 2007)[5]. Only by averaging the familiar nonnegativity property of informational quantities is obtained.

6.3.1 The Environmental Response Term

The information gain Eq. (22) consists of two terms which we discuss in turn. The first term quantifies the statistical surprise in the transition due to our action (relative to the prior). It can be seen as the *environmental response information* as it measures the response of the world to the agent's control action. It is also interpretable as the information gained if one can *observe* the next state (in a fully observed MDP), or as the information *processed* by the environment in this state transition. In Sec. 7.2 this term combined together with the MDP reward will give rise to the concept of the perfectly adapted environment which reflects the perception-action cycle version of the notion of optimally matched channels by Gastpar et al. (2003).

The environmental response information can be considered an information-theoretic generalization or a soft version of the control-theoretic concept of *controllability* (in this context, see also Ashby 1956; Touchette and Lloyd 2000, 2004; Klyubin et al. 2005b; Todorov 2009). As here we do not limit the agent's access to the world state and also do not model the sensoric cost, the information gain term does not contain an analogous information-theoretic term corresponding to observability, but this is only a restriction of the current scenario, not of the model in general.

Strictly spoken, when we talk above about controllability/observability, we refer only to the actual level of control (and observation) exerted, not about controllability (and observability) in the sense of the maximally achievable control/observation. For an information-theoretic treatment of *combined* controllability and observability in the latter (i.e. maximality) sense, see e.g. (Klyubin et al. 2008).

[5] The interpretation of a negative information gain is that under the presence/observation of a particular condition the subsequent distributions are blurred. One caricature example would be that, to solve a crime, one would have a probability distribution sharply concentrated on a particular crime suspect. If now additional evidence would exclude that suspect from consideration and reset the distribution to cover all suspects equally, this would be an example for negative information gain.

6.3.2 The Decision Complexity Term

We now turn briefly to the second term in Eq. (22); the second term reflects the decision complexity, i.e. the informational effort that the agent has to invest in the subsequent decision at time $t+1$. The average of this term according to Eq. (21) measures the information required for the selection of the agent's action at time $t+1$. Importantly, note that this value for the decision complexity at time $t+1$ as calculated from the recursive Eq. (21) and Eq. (22) is always conditional on the state s_t and the action a_t at the current time t.

These two components make clear that the information processing exhibited by the agent-environment system decomposes into two parts, one that captures the environmental information processing, and one that reflects the agent's decision. This decomposition is related to that known from compositional Markov chains (Wennekers and Ay 2005) and provides an elegant and transparent way of distinguishing which part of a system is responsible for which aspect of information processing.

7 Trading Information and Value

We can now calculate the minimal information-to-go (i.e., environmental information processing cost plus decision complexity) that is required to achieve a given level of value-to-go.

7.1 The "Free-Energy" functional

At this point, we remind the reader of Eq. (9) which is used in the Reinforcement Learning literature to characterize the value- or reward-to-go in terms of state-action pairs instead of states only:

$$Q^{\pi}(s_t, a_t) = \sum_{s_{t+1}} \mathbf{P}^{s_{t+1}}_{s_t,a_t} \cdot \left[\mathbf{R}^{s_{t+1}}_{s_t,a_t} + V^{\pi}(s_{t+1}) \right] . \tag{23}$$

As V^{π} quantifies the future expected cumulative reward when starting in state s_t and then following the policy π, the function Q^{π} separates out also the initial action a_t, in addition to the initial state s_t.

The constrained optimization problem of finding the minimal information-to-go at a given level of value-to-go can be turned into an unconstrained one using the Lagrange method; for this the quantity to minimize (the information-to-go) is complemented by the constraint (the value-to-go) multiplied by a Langrange multiplier β:

$$F^{\pi}(s_t, a_t, \beta) \stackrel{\Delta}{=} \Im^{\pi}(s_t, a_t) - \beta Q^{\pi}(s_t, a_t); . \tag{24}$$

This Lagrangian builds a link to the Free Energy formalism known from statistical physics: the Langrange multiplier β corresponds to the inverse temperature, the information-to-go $\Im^\pi$ corresponds to the physical entropy. The value-to-go (here expressed as Q^π) corresponds to the energy of a system, and F^π/β corresponds to the free energy from statistical physics. However, for simplicity we will apply the notion of *free energy* to F^π itself. The analogy with the free energy from statistical physics provides an additional justification for the minimization of the information-to-go under value-to-go constraints: the minimization of F^π identifies the least committed policy in the sense that the future is the least informative, i.e. the least constrained and thus the most robust.

This philosophy is closely related to the minimum information principle (Globerson et al. 2009): if one has an input-output relationship, one selects a model that processes the least information that is consistent with the observations. This corresponds again to the least committed solution that covers the observations (and is, in general, not identical to the maximum entropy solution with which it coincides only in certain cases, see Globerson et al. 2009).

For the later purposes, it is useful to expand the free energy as follows :

$$\begin{aligned}
&\Im^\pi(s_t,a_t) - \beta Q^\pi(s_t,a_t) = \\
&= \mathbb{E}_{p(s_{t+1}|s_t,a_t)\pi(a_{t+1}|s_{t+1})}\Bigg[\log\frac{p(s_{t+1}|s_t,a_t)}{\hat{p}(s_{t+1})} + \log\frac{\pi(a_{t+1}|s_{t+1})}{\hat{\pi}(a_{t+1})} \\
&\qquad\qquad - \beta\mathbf{R}_{s_t a_t}^{s_{t+1}} + \Im^\pi(s_{t+1},a_{t+1}) - \beta V^\pi(s_{t+1})\Bigg] \\
&\qquad = \mathbb{E}_{p(s_{t+1}|s_t,a_t)}\Bigg[\log\frac{p(s_{t+1}|s_t,a_t)}{\hat{p}(s_{t+1})} - \beta\mathbf{R}_{s_t a_t}^{s_{t+1}} \\
&\qquad\qquad + \mathbb{E}_{\pi(a_{t+1}|s_{t+1})}\Bigg[\log\frac{\pi(a_{t+1}|s_{t+1})}{\hat{\pi}(a_{t+1})} + \Im^\pi(s_{t+1},a_{t+1}) \\
&\qquad\qquad\qquad - \beta Q^\pi(s_{t+1},a_{t+1})\Bigg]\Bigg]
\end{aligned} \tag{25}$$

where the last equality follows from

$$V^\pi(s_{t+1}) = \mathbb{E}_{\pi(a_{t+1}|s_{t+1})}[Q^\pi(s_{t+1},a_{t+1})]\,.$$

This leads to the following recursive relation for the free energy:

$$\begin{aligned}
F^\pi(s_t,a_t,\beta) = \mathbb{E}_{p(s_{t+1}|s_t,a_t)}\Bigg[&\log\frac{p(s_{t+1}|s_t,a_t)}{\hat{p}(s_{t+1})} - \beta\mathbf{R}_{s_t a_t}^{s_{t+1}} \\
&+ \mathbb{E}_{\pi(a_{t+1}|s_{t+1})}\left[\log\frac{\pi(a_{t+1}|s_{t+1})}{\hat{\pi}(a_{t+1})} + F^\pi(s_{t+1},a_{t+1},\beta)\right]\Bigg]\,.
\end{aligned} \tag{26}$$

The task of finding the optimal policy, i.e. the one minimizing its information-to-go under a constraint on the attained value-to-go is solved by the unconstrained minimization of the corresponding Lagrangian, i.e. the free energy functional F^{π}:

$$\operatorname*{argmin}_{\pi} F^{\pi}(s_t, a_t, \beta) = \operatorname*{argmin}_{\pi} [\Im^{\pi}(s_t, a_t) - \beta Q^{\pi}(s_t, a_t)]$$

where the minimization ranges over all policies. A particular constraint on the value-to-go is imposed by selecting the respective "inverse temperature" Lagrange multiplier β.

Extending Eq. (26) by the Lagrange term for the normalization of π and taking the gradient with respect to π, and then setting the gradient of F^{π} to 0 (both for the entire term as well as inside the brackets) provides us with a Bellman-type recursion for the free energy functional as follows: an optimal policy π satisfies the recursive Eq. (26) as well as the relations

$$\pi(a|s) = \frac{\hat{\pi}(a)}{Z^{\pi}(s,\beta)} \exp(-F^{\pi}(s,a,\beta)) \tag{27}$$

$$Z^{\pi}(s,\beta) = \sum_{a} \hat{\pi}(a) \exp(-F^{\pi}(s,a,\beta)) \tag{28}$$

$$\hat{\pi}(a) = \sum_{s} \pi(a|s)\hat{p}(s)\,. \tag{29}$$

in a self-consistent fashion. In turn, iterating the system of self-consistent Equations (26) to (29) till convergence for every state will produce an optimal policy. This system of equations essentially unifies the Bellman Equation and the Blahut-Arimoto algorithm from rate-distortion theory.

Notice that as result of the algorithm, we obtain a non-trivial soft-max policy for every finite value of β. Furthermore, if the optimal policy is unique, the equations will recover it as a deterministic policy for the limit $\beta \to \infty$. The compound iterations converged to a unique policy for any finite value of β. While we believe that a convergence proof (possibly without uniqueness guarantees) could be developed along the lines of the usual convergence proofs for the Blahut-Arimoto algorithm, we defer this to a future paper.

It should be mentioned at this point that the reward structure determining the form of Q is an externally defined part of the system description. In the present paper, the reward can be of any type. However, the reward could be realized as a more specific quantity, e.g. an informational measure, such as, for example, a predictive information gain in which case the formalism would reduce to a particular form.

7.2 Perfectly adapted environments

An intriguing aspect of the free-energy formalism that we can consider the optimality not only of the agent's policy but also that of the environment. This is particularly relevant for the characterization of a "best match" between the organism's action space and the responses of the environment, which realizes a *perfectly adapted environment*. We already commented earlier on the close conceptual analogy between the concept of the perfectly adapted environment and what we suggest to be its information-theoretic counterpart: the environment being considered as the channel, and the agent's actions as the source in an optimally matched channel (Gastpar et al. 2003).

In the following paragraphs we characterize the optimal (i.e. perfectly adapted) environment in the language our formalism as the the MDP that minimizes the free energy. Define the notation

$$q_\beta(s'|s,a) = q_{s,a}^{s'} \stackrel{\Delta}{=} \frac{p(s')}{Z(\beta,s,a)} \exp\left(\beta \mathbf{R}_{s,a}^{s'}\right). \tag{30}$$

Then, with the free energy functional $F^\pi(s_t,a_t,\beta) = \Im^\pi(s_t,a_t) - \beta Q^\pi(s_t,a_t)$, the Bellman equation can be rewritten as:

$$F^\pi(s_t,a_t,\beta) = \mathbb{E}_{p(s_{t+1}|s_t,a_t)}\left[\log\frac{p(s_{t+1}|s_t,a_t)}{q_\beta(s_{t+1}|s_t,a_t)} - \log Z(\beta,s_t,a_t) + \mathbb{E}_{\pi(a_{t+1}|s_{t+1})}\left[\log\frac{\pi(a_{t+1}|s_{t+1})}{\hat{\pi}(a_{t+1})} + F^\pi(s_{t+1},a_{t+1},\beta)\right]\right].$$

Note that in this form the first term averages to the Kullback-Leibler divergence between the actual probability $p(s_{t+1}|s_t,a_t)$ and the "optimal distribution" $q_\beta(s_{t+1}|s_t,a_t)$ of the next state s_{t+1}, for fixed current state s_t and action a_t.

The first term in F is minimized[6] with respect to the environment transition probabilities precisely when the MDP is fully adapted to the reward, namely, when

$$p(s_{t+1}|s_t,a_t) = q_\beta(s_{t+1}|s_t,a_t). \tag{31}$$

In this case, the Kullback-Leibler divergence vanishes. Plugging in the optimal policy π (satisfying Equations (26) to (29)), and using the special relationship between the state transitions and the rewards (Eqs. (30) and (31)), the accumulated term reduces to the sum $-\log Z(\beta,s_t,a_t) - \log Z^\pi(s_{t+1},\beta)$, i.e. the local free energy purely of the current step which itself consists of the environmental and the agent component.

[6] Alternatively, one could minimize F^π by setting the gradient of F^π with respect to $p(s_{t+1}|s_t,a_t)$ to 0 similar to the derivation of Eqs. (27) to (29) under the assumption that π is already optimized. This implements the assumption that the adaptation of the environmental channel is "slow" corresponding to the adaptation of the agent policy.

In the particular case of perfectly adapted environments, all the future information is indeed valuable. In other words, minimizing the statistical surprise or maximizing the predictive information is equivalent to maximizing the reward. This can be interpreted as a generalization of the classical Kelly gambling scenario (Kelly 1956). Note that this is not the case for general reward structures $\mathbf{R}^{s'}_{s,a}$. In view of the hypothesized central role of information for biological systems, it will be of significant interest for future research to establish to which extent the environments of living organisms are indeed perfectly adapted.

8 Experiments and Discussion

8.1 Information-Value Trade-Off in a Maze

The recursion developed in Sec. 7.1 can be applied to various scenarios, of which we study one specific, but instructive case. We consider a simple maze (inset of Fig. 1) where an agent starts out at the bright spot in the lower left corner of the grid world and needs to reach the target in the right upper corner, marked by the red dot. The task is modeled through a usual Reinforcement Learning reward, by giving each step a "reward" (i.e. a penalty) of -1 until the target is reached. The target cell is an absorbing state, and once the agent reaches it, any subsequent step receives a 0 reward, realizing an episodic task in the non-episodic framework of the Bellman-recursion from Sec. 6.

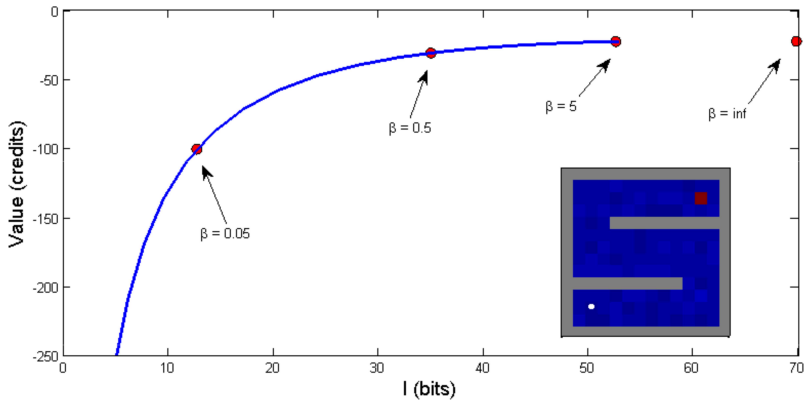


Fig. 1 Trade-off curve between value-to-go and information-to-go. This is in full analogy to the rate-distortion plots, if we consider (negative) distortion replaced by value-to-go and rate by information-to-go.

Figure 1 shows how as one permits increasing amounts of information-to-go, the future expected cumulated reward achieved also increases (it is the negative value of

the length of the route — i.e. as one is ready to invest more information bandwidth, one can shorten the route). Note that when β vanishes, this attempts to save on information-to-go while being indifferent to the achievement of a high value-to-go. As opposed to that, letting $\beta \to \infty$ aims for a policy that is indeed optimal in its value-to-go.

Now, in the case of $\beta \to \infty$, similar to (Polani et al. 2006), the informational Bellman recursion will find a policy which is optimal for the Reinforcement Learning task. However, unlike the conventional policy or value iteration algorithms, the algorithm will not be "satisfied" with a value-optimal solution, but select a policy among the optimal policies which at the same time minimizes the information-to-go.

8.2 Soft vs. Sharp Policies

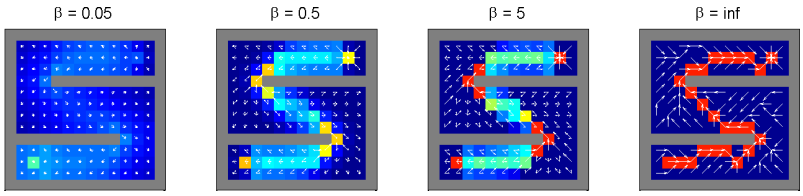


Fig. 2 Policies resulting from the trade-off between value-to-go and information-to-go

Figure 2 shows actual policies resulting for various values of β. For small β, the policy is almost a random walk. Such a walk will ultimately end up in the (absorbing) goal state, at little informational cost, but at quite negative reward values, since it takes a long time to find the goal.

As one increases β and thus increases the available information capacity, sharper, more refined and accurate policies emerge. Note that, in general, the policies we obtain by the informational trade-off algorithm from Secs. 6 and 7 will be soft policies for finite β and an agent following them will produce state trajectories which increasingly expand and branch out over time which will typically reduce the future expected cumulated reward. This may allow additional scenarios to those mentioned in Footnote 1 (Sec. 4.2) to exhibit converging rewards without having to use the usual discount factor, as long as β is finite. In these cases, if the cumulated rewards diverge for $\beta \to \infty$ (zero temperature, i.e. optimal policy), this would only constitute a "pathological" boundary case.

Under the PAC-Bayes perspective (McAllester 1999), our free energy is composed of the cumulated Kullback-Leibler distances between posterior and prior distribution of S_{t+1} and A_{t+1} (see Eqs. (21) and (22)). This gives rise to another interesting interpretation of the soft policies obtained by the above formalism: namely, the policies minimizing the respective Kullback-Leibler expressions in these equations provide a bound on the variation of the accumulated reward over different episodes

of the agent's run; in fact, those policies provide *stable* results in the sense that the bound on variations from run to run is the tightest (Rubin et al. 2010).

The soft policies we obtained above are similar to the "softened paths" suggested in (Saerens et al. 2009), derived from entropic quantities used as regularization term for the Reinforcement Learning task. In the present paper, however, as in (Polani et al. 2006), we use Shannon information not just as a regularization quantity, but with a specific interpretation as an information processing cost: the minimal informational cost of the decision process that the agent (or, in the present paper, the agent-environment) system has to undergo in order to achieve a particular reward.

In the light of this interpretation, the study of information has immediate repercussions for the biological picture of information processing, i.e. the "information metabolism" of organisms. If one adopts the view that organisms tend to implement an information parsimony principle (Laughlin 2001; Polani 2009), then this implies that biological systems will exhibit a tendency to achieve a given level of performance at the lowest informational cost possible (or perform as well as possible under a given informational bandwidth). In our formalism, this would correspond to operating close to the optimal reward/information (strictly spoken, decision complexity) trade-off curve, always assuming that a suitable reward function can be formulated (Taylor et al. 2007; Bialek et al. 2007).

In the present paper, we demonstrated how the trade-off curve between value-to-go and information-to-go can be computed for the agent-environment system over whole sequence histories using a Bellman-type recursive backup rule. In the future, we will apply these techniques introduced here to other variants of the problem. One is the calculation of the decision complexity (i.e. the relevant information) only, the minimal amount of information that needs to acquired and processed by the agent itself, but not by the environment, to achieve a certain reward. In (Polani et al. 2006), the relevant information was computed only for a single-step action sequence. With the Information-Bellman backup rule introduced here, we will be able in the future to generalize the relevant information calculation to multi-step action sequences. To quantify parsimonious information acquisition in the multi-step case, we will use Massey's concept of *directed information* (Massey 1990).

At this point, some comments are in place concerning the Markovicity of the world state in our models. The Markovicity condition seems, at first sight, a comparatively strong assumption which might seem limit the applicability of the formalism for modeling the subjective knowledge of an organism or agent. However, note that, while we compute various quantities from a "eagle's eye perspective" under knowledge of the full state, in the model the agent itself is not assumed to have full access to the state. Rather, the information bandwidth but not its precise form is constrained in the present paper. Finally, using the full formalism from Eq. (11), more complex structural constraints on the information acquisition can easily be incorporated in the form of sensors.

Let us here emphasize another final point: the presented perception-action cycle formalism implements an information-theoretic analogy for the classical treatment of optimal control problems. However, as we propose to consider information not merely as an auxiliary quantity, but in fact as a "first class" quantity in its own right,

the present treatment aims to go beyond just an equivalent restatement of stochastic optimal control: rather, to provide a conceptually enriched framework, in which the informational view gives rise to a refined set of notions, insights, tools, and, ultimately, research questions.

9 Conclusions

In the paper, we have treated the reward-driven decision process in the perception-action cycle of an agent in a consistently information-theoretic framework. This was motivated by increasing biological evidence for the importance of (Shannon) information as resource and by the universality that the language of information is able to provide.

We consider a particular incarnation of this problem, namely an agent situated in an MDP defining a concrete task; this task is encoded as a cumulated reward which the agent needs to maximize. The information-theoretic view transforms this problem into a trade-off between the reward achieved at a given informational cost. This extends classic rate-distortion theory into the context of a full-fledged perception-action cycle. At the same time, the methodology gives a precise quantitative meaning to J.M. Fuster's above quoted intuition about the perception-action cycle being the "circular flow of information between an organism and its environment".

The paper shows that not only it is possible and natural to reframe the treatment of perception-action cycles in this way, but that MDP formalisms such as the Bellman recursion can be readily extended to provide a unified Blahut-Arimoto/Value Iteration hybrid that computes the quantities of interest. In the current paper, we illustrated this idea in a simple setting. More comprehensive settings which are of significant interest for both biology as well as for artificial intelligence can be readily incorporated due to the flexibility of the formalism and will be treated in future work.

We hypothesize that the ability to trade off the value and the informational cost of whole behaviours lies at the core of any understanding of organismic behaviours. The hypothesis is that organisms attempt to realize valuable behaviours at the lowest possible informational cost, and that they will seek slightly suboptimal solutions if these solutions can be afforded at a significantly lower informational cost. Thus, the informational treatment of the perception-action cycle promises to open a quantitative and predictive path to understand the structure of behaviours and information processing in living organisms. At the same time it can provide a systematic handle on how to develop AI systems according to principles which are both biologically plausible and relevant.

Acknowledgement

The authors would like to thank Jonathan Rubin for carrying out the simulations and the preparation of the corresponding diagrams.

References

Ashby, W. R., (1956). *An Introduction to Cybernetics*. Chapman & Hall Ltd.

Ay, N., Bertschinger, N., Der, R., Güttler, F., and Olbrich, E., (2008). Predictive Information and Explorative Behavior of Autonomous Robots. *European Journal of Physics B*, 63:329–339.

Ay, N., and Polani, D., (2008). Information Flows in Causal Networks. *Advances in Complex Systems*, 11(1):17–41.

Ay, N., and Wennekers, T., (2003). Dynamical Properties of Strongly Interacting Markov Chains. *Neural Networks*, 16(10):1483–1497.

Berger, T., (2003). Living Information Theory — The 2002 Shannon Lecture. *IEEE Information Theory Society Newsletter*, 53(1):1,6–19.

Bialek, W., de Ruyter van Steveninck, R. R., and Tishby, N., (2007). Efficient representation as a design principle for neural coding and computation. arXiv.org:0712.4381 [q-bio.NC].

Bialek, W., Nemenman, I., and Tishby, N., (2001). Predictability, complexity and learning. *Neural Computation*, 13:2409–2463.

Brenner, N., Bialek, W., and de Ruyter van Steveninck, R., (2000). Adaptive rescaling optimizes information transmission. *Neuron*, 26:695–702.

Cover, T. M., and Thomas, J. A., (1991). *Elements of Information Theory*. New York: Wiley.

Csiszár, I., and Körner, J., (1986). *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Budapest: Academiai Kiado.

Der, R., Steinmetz, U., and Pasemann, F., (1999). Homeokinesis – A new principle to back up evolution with learning. In Mohammadian, M., editor, *Computational Intelligence for Modelling, Control, and Automation*, vol. 55 of *Concurrent Systems Engineering Series*, 43–47. IOS Press.

Ellison, C., Mahoney, J., and Crutchfield, J., (2009). Prediction, Retrodiction, and the Amount of Information Stored in the Present. *Journal of Statistical Physics*, 136(6):1005–1034.

Engel, Y., Mannor, S., and Meir, R., (2003). Bayes meets Bellman: The Gaussian Process Approach to Temporal Difference Learning. In *Proc. of ICML 20*, 154–161.

Friston, K., (2009). The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.*, 13(7):293–301.

Friston, K., Kilner, J., and Harrison, L., (2006). A free energy principle for the brain. *Journal of Physiology-Paris*, 100:70–87.

Fry, R. L., (2008). Computation by Neural and Cortical Systems. Presentation at the Workshop at CNS*2008, Portland, OR: Methods of Information Theory in Computational Neuroscience.

Fuster, J. M., (2001). The Prefrontal Cortex — An Update: Time Is of the Essence. *Neuron*, 30:319–333.

Fuster, J. M., (2006). The cognit: A network model of cortical representation. *International Journal of Psychophysiology*, 60(2):125–132.

Gastpar, M., Rimoldi, B., and Vetterli, M., (2003). To Code, or Not to Code: Lossy Source-Channel Communication Revisited. *IEEE Transactions on Information Theory*, 49(5):1147– 1158.

Globerson, A., Stark, E., Vaadia, E., and Tishby, N., (2009). The Minimum Information principle and its application to neural code analysis. *PNAS*, 106(9):3490–3495.

Haken, H., (1983). *Advanced synergetics*. Berlin: Springer-Verlag.

Howard, R. A., (1966). Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, SSC-2:22–26.

Jung, T., and Polani, D., (2007). Kernelizing LSPE(λ). In *Proc. 2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning, April 1-5, Hawaii*, 338–345.

Kappen, B., Gomez, V., and Opper, M., (2009). Optimal control as a graphical model inference problem. arXiv:0901.0633v2 [cs.AI].

Kelly, J. L., (1956). A New Interpretation of Information Rate. *Bell System Technical Journal*, 35:917–926.

Klyubin, A., Polani, D., and Nehaniv, C., (2007). Representations of Space and Time in the Maximization of Information Flow in the Perception-Action Loop. *Neural Computation*, 19(9):2387–2432.

Klyubin, A. S., Polani, D., and Nehaniv, C. L., (2004). Organization of the Information Flow in the Perception-Action Loop of Evolved Agents. In *Proceedings of 2004 NASA/DoD Conference on Evolvable Hardware*, 177–180. IEEE Computer Society.

Klyubin, A. S., Polani, D., and Nehaniv, C. L., (2005a). All Else Being Equal Be Empowered. In *Advances in Artificial Life, European Conference on Artificial Life (ECAL 2005)*, vol. 3630 of *LNAI*, 744–753. Springer.

Klyubin, A. S., Polani, D., and Nehaniv, C. L., (2005b). Empowerment: A Universal Agent-Centric Measure of Control. In *Proc. IEEE Congress on Evolutionary Computation, 2-5 September 2005, Edinburgh, Scotland (CEC 2005)*, 128–135. IEEE.

Klyubin, A. S., Polani, D., and Nehaniv, C. L., (2008). Keep Your Options Open: An Information-Based Driving Principle for Sensorimotor Systems. *PLoS ONE*, 3(12):e4018.
`http://dx.doi.org/10.1371/journal.pone.0004018`, Dec 2008

Laughlin, S. B., (2001). Energy as a constraint on the coding and processing of sensory information. *Current Opinion in Neurobiology*, 11:475–480.

Lizier, J., Prokopenko, M., and Zomaya, A., (2007). Detecting non-trivial computation in complex dynamics. In Almeida e Costa, F., Rocha, L. M., Costa, E.,

Harvey, I., and Coutinho, A., editors, *Advances in Artificial Life (Proc. ECAL 2007, Lisbon)*, vol. 4648 of *LNCS*, 895–904. Berlin: Springer.

Lungarella, M., and Sporns, O., (2005). Information Self-Structuring: Key Principle for Learning and Development. In *Proceedings of 4th IEEE International Conference on Development and Learning*, 25–30. IEEE.

Lungarella, M., and Sporns, O., (2006). Mapping Information Flow in Sensorimotor Networks. *PLoS Computational Biology*, 2(10).

Massey, J., (1990). Causality, feedback and directed information. In *Proc. Int. Symp. Inf. Theory Applic. (ISITA-90)*, 303–305.

McAllester, D. A., (1999). PAC-Bayesian model averaging. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory, Santa Cruz, CA*, 164–170. New York: ACM.

Pearl, J., (2000). *Causality: Models, Reasoning and Inference*. Cambridge, UK: Cambridge University Press.

Pfeifer, R., and Bongard, J., (2007). *How the Body Shapes the Way We think: A New View of Intelligence*. Bradford Books.

Polani, D., (2009). Information: Currency of Life?. *HFSP Journal*, 3(5):307–316. `http://link.aip.org/link/?HFS/3/307/1`, Nov 2009

Polani, D., Martinetz, T., and Kim, J., (2001). An Information-Theoretic Approach for the Quantification of Relevance. In Kelemen, J., and Sosik, P., editors, *Advances in Artificial Life (Proc. 6th European Conference on Artificial Life)*, vol. 2159 of *LNAI*, 704–713. Springer.

Polani, D., Nehaniv, C., Martinetz, T., and Kim, J. T., (2006). Relevant Information in Optimized Persistence vs. Progeny Strategies. In (Rocha et al. 2006), 337–343.

Prokopenko, M., Gerasimov, V., and Tanev, I., (2006). Evolving Spatiotemporal Coordination in a Modular Robotic System. In Nolfi, S., Baldassarre, G., Calabretta, R., Hallam, J. C. T., Marocco, D., Meyer, J.-A., Miglino, O., and Parisi, D., editors, *From Animals to Animats 9: 9th International Conference on the Simulation of Adaptive Behavior (SAB 2006), Rome, Italy*, vol. 4095 of *Lecture Notes in Computer Science*, 558–569. Berlin, Heidelberg: Springer.

Rocha, L. M., Bedau, M., Floreano, D., Goldstone, R., Vespignani, A., and Yaeger, L., editors, (2006). *Proc. Artificial Life X*.

Rubin, J., Shamir, O., and Tishby, N., (2010). A PAC-Bayesian Analysis of Reinforcement Learning. In preparation.

Saerens, M., Achbany, Y., Fuss, F., and Yen, L., (2009). Randomized Shortest-Path Problems: Two Related Models. *Neural Computation*, 21:2363–2404.

Seldin, Y., and Tishby, N., (2009). PAC-Bayesian Generalization Bound for Density Estimation with Application to Co-clustering. In *Proc. 12th International Conference on Artificial Intelligence and Statistics (AIStats 2009)*, vol. 5 of *JMLR Workshop and Conference Proceedings*.

Shalizi, C. R., and Crutchfield, J. P., (2002). Information Bottlenecks, Causal States, and Statistical Relevance Bases: How to Represent Relevant Information in Memoryless Transduction. *Advances in Complex Systems*, 5:1–5.

Shannon, C. E., (1949). The Mathematical Theory of Communication. In Shannon, C. E., and Weaver, W., editors, *The Mathematical Theory of Communication*. Urbana: The University of Illinois Press.

Slonim, N., Friedman, N., and Tishby, N., (2006). Multivariate Information Bottleneck. *Neural Computation*, 18(8):1739–1789.

Sporns, O., and Lungarella, M., (2006). Evolving coordinated behavior by maximizing information structure. In (Rocha et al. 2006), 323–329.

Still, S., (2009). Information-theoretic approach to interactive learning. *EPL (Europhysics Letters)*, 85(2):28005–28010.

Strens, M., (2000). A Bayesian Framework for Reinforcement Learning. In Langley, P., editor, *Proc. 17th Intl. Conf. on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*. Morgan Kaufmann.

Sutton, R. S., and Barto, A. G., (1998). *Reinforcement Learning*. Cambridge, Mass.: MIT Press.

Taylor, S. F., Tishby, N., and Bialek, W., (2007). Information and Fitness. arXiv.org:0712.4382 [q-bio.PE].

Tishby, N., Pereira, F. C., and Bialek, W., (1999). The Information Bottleneck Method. In *Proc. 37th Annual Allerton Conference on Communication, Control and Computing, Illinois*. Urbana-Champaign.

Todorov, E., (2009). Efficient computation of optimal actions. *PNAS*, 106(28):11478–11483.

Touchette, H., and Lloyd, S., (2000). Information-Theoretic Limits of Control. *Phys. Rev. Lett.*, 84:1156.

Touchette, H., and Lloyd, S., (2004). Information-theoretic approach to the study of control systems. *Physica A*, 331:140–172.

van Dijk, S. G., Polani, D., and Nehaniv, C. L., (2009). Hierarchical Behaviours: Getting the Most Bang for your Bit. In Kampis, G., and Szathmáry, E., editors, *Proc. European Conference on Artificial Life 2009, Budapest*. Springer.

Vergassola, M., Villermaux, E., and Shraiman, B. I., (2007). 'Infotaxis' as a strategy for searching without gradients. *Nature*, 445:406–409.

Wennekers, T., and Ay, N., (2005). Finite State Automata Resulting From Temporal Information Maximization. *Neural Computation*, 17(10):2258–2290.