A Unified Bellman Equation for Causal Information and Value in Markov Decision Processes

Stas Tiomkin [1] **Naftali Tishby** [1] [2]

Abstract

The interaction between an artificial agent and its environment is bi-directional. The agent extracts relevant information from the environment, and affects the environment by its actions in return to accumulate high expected reward. Standard reinforcement learning (RL) deals with the expected reward maximization. However, there are always information-theoretic limitations that restrict the expected reward, which are not properly considered by the standard RL.

In this work we consider RL objectives with information-theoretic limitations. For the first time we derive a Bellman-type recursive equation for the causal information between the environment and the agent, which is combined plausibly with the Bellman recursion for the value function. The unified equitation serves to explore the typical behavior of artificial agents in an infinite time horizon.

1. Introduction

The interaction between an organism and the environment consists of three major components - utilization of past experience, observation of the current environmental state, and generation of a behavioral policy. The latter component, *planning* is an essential feature of intelligent systems for survival in limited-resource environments. Organisms with long-term planning can accumulate more resources and avoid possible catastrophic states in the future, which makes than evolutionarily superior to 'short-term planners'. An intelligent agent combines past experience with the current observations to act upon the environment. This feedback interaction induces complex statistical correlations over different time scales between environment state trajectories and the agent's action trajectories. Infinite-time

interaction patterns, (infinite state and action trajectories) define the typical behavior of an organism in a given environment. This typical behavior is crucial for the design and analysis of intelligent systems. In this work we derive typical behavior within the formalism of a reinforcement learning (RL) model subject to information-theoretic constraints.

In the standard RL model, (Sutton & Barto, 1998), an artificial organism generates an optimal policy through an interaction with the environment. Typically, the optimality is taken with regard to a reward, such as energy, money, social network 'likes/dislikes', time etc. Intriguingly, the reward-associated value function, (an average accumulated reward), possesses the same property for different types of rewards - the optimal value function is a *Lyapunov function*, (Perkins & Barto, 2002), a *generalized energy* function.

A principled way to solve RL (as well as Optimal Control) problems is to find a corresponding generalized energy function. Specifically, the Bellman recursive equation, (Bertsekas, 1995), is the *Lyapunov function* for Markov decision processes. Strictly speaking, the standard RL framework is about the discovery and minimization (maximization) of an *energetic* quantity, whereas the Bellman recursion is an elegant tool for finding the optimal policy both for finite and infinite time horizons for planning.

However, this model is incomplete. Physics suggests that there is always an interplay between energetic and entropic quantities. Recent advances in the artificial intelligence show that the behavior of *realistic* artificial agents is indeed affected simultaneously by energetic and entropic quantities (Tishby & Polani, 2011; Rubin et al., 2012). The finite rate of information transfer is a fundamental constraint for any value accumulation mechanism.

Information-theoretic constraints are ubiquitous in the interaction between an agent and the environment. For example, a sensor has a limited bandwidth, a processor has a limited information processing rate, a memory has a limited information update rate, etc. All these are not pure theoretical considerations, but rather the practical requirements to build and analyze realistic intelligent agents. Moreover, often, an artificial agent needs to limit the available information to the information for a particular task, (Polani et al., 2006). For example, when looking at a picture, the visual stimuli strikes the retina at the rate of the speed of light,

---

[1]The Benin School of Computer Science and Engineering, The Hebrew University, Jerusalem, Israel. [2]The Edmond and Lilly Safra Center for Brain Sciences, The Hebrew University, Jerusalem, Israel. Correspondence to: Stas Tiomkin <stas.tiomkin@mail.huji.ac.il>, Naftali Tishby <tishby@cs.huji.ac.il>.

which is decreased dramatically to leave only the relevant information rate, which is essential for understanding the picture.

Different types of entropic constraints have been studied in the context of RL. The KL-control, (Todorov et al., 2006; Tishby & Polani, 2011; Kappen et al., 2012) introduces a D_{KL} cost to each state, which penalizes complex policies, where complexity is measured by the D_{KL} divergence between a prior policy, or uncontrolled dynamics, and an actual policy. Importantly, the KL-control does not directly address the question of the information transfer rate.

Some recent works do explicitly consider the effects of information constraints on the performance of a controller (Tatikonda & Mitter, 2004; Tanaka & Sandberg, 2015; Tanaka et al., 2015) for an infinite time horizon. However, these works are mostly limited to linear dynamics.

Obviously, feedback implies causality - the current action depends on previous and/or present information alone. Consequently, the information constraints should address the interaction causality.

In this work we consider directed information constraints over an infinite time horizon in the framework of Markov decision processes, where the directed information is between the environment state trajectories and the agent's action trajectories.

For the first time we derive a Bellman type equation for the causal information, and combine it with the Bellman recursion for value. The unified Information-Value Bellman equation enables us to derive the typical optimal behavior over an infinite time horizon in the MDP setting.

In addition, this work has practical implications for the design criteria of optimal intelligent agents. Specifically, an information processing rate of a 'brain' (processor) of an artificial agent should be higher than the minimal information rate required to solve a particular MDP problem.

The interaction between an agent and the environment is bi-directional - an agent extracts information and affects the environment by its actions in return. To provide a comprehensive analysis of this bi-directional interaction, we consider both information channels. These channels are dual, comprising the action-perception cycle of reinforcement learning under the information-theoretic constraints. We derive a Bellman type equation for the directed information from action trajectories to state trajectories.

The minimal directed information rate from the environment to the agent comprises a constraint for the expected reward maximization in standard goal-oriented RL test bench tasks. By contrast, the maximal directed information rate from the agent to the environment comprises a *proxy* for solving different AI test bench tasks without specifying the target, which can be considered as self-motivated behavior to be ready for any possible target. We elaborate on the difference between these two opposite information rates by their corresponding analytic expressions, and by numer-

ical simulations of different types of problems, which can be solved by them.

The paper is organized as follows. In Section 2 we provide background on causal conditioning, directed information, and Markov decision processes. In Section 3 we represent the Bellman-type recursive equation for directed information from the agent's actions to the environment. In Section 4 we show the unified recursion for the directed information and the common recursion for the value function. The optimization problem for finding the minimal directed information rate required to achieve the maximal expected reward in an infinite time horizon is stated in Section 4.1. The numerical simulation of the unified Bellman equation is provided in Section 5, where we consider the standard maze-escaping problem for a predefined target. In Section 6 we consider the dual problem of the maximal directed information rate from the agent to the environment, and derive the dual Bellman equation. In Section 7 we provide a numerical simulation for the dual Bellman equation, where we consider tasks without a predefined target. We compare our solution to the standard algorithms for finding the average all-pairs shortest path problem. Finally, in Section 8 we summarize the paper, and provide directions for the continuation of this work.

2. Background

In this section we overview the theoretical background for this work. Specifically, we briefly review the framework of reinforcement learning and Markov decision processes. Then, we review causal conditioning and directed information, and define the required quantities.

2.1. Interaction Model

We assume that the agent and environment interact at discrete time steps for $t \in (-\infty, \ldots, T)$. At time t, the environment appears at state $S_t \in \mathcal{S}$, the agent observes the state, and affects the environment by its action $A_t \in \mathcal{A}$ according to the policy $\pi(A_t \mid S_t)$. The environmental evolution is given by the transition probability distribution, $p(S_{t+1} \mid A_t, S_t)$. This interaction model is plausibly described by the probabilistic graph, shown at Figure 1, where the arrows denote directions of causal dependency.
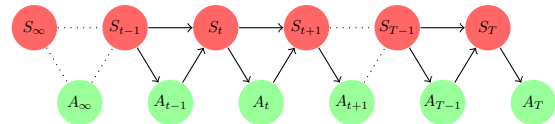

Figure 1. A probabilistic directed graph for a Markov Decision Process.

As a results of this feedback interaction, the environment state trajectories, $S_{-\infty}^T$ defined by (1), and the agent action trajectories, $A_{-\infty}^T$ defined by (2), become statistically correlated.

$$S_{-\infty}^T \doteq (\ldots, S_{t-1}, S_t, S_{t+1}, \ldots, S_T), \qquad (1)$$
$$A_{-\infty}^T \doteq (\ldots, A_{t-1}, A_t, A_{t+1}, \ldots, A_T).. \qquad (2)$$

The corresponding trajectories of finite length length T, and their time-delayed versions are given by:

$$S_{t+1}^T \doteq (S_{t+1}, S_{t+2}, \ldots, S_T),$$
$$A_{t+1}^T \doteq (A_{t+1}, A_{t+2}, \ldots, A_T),$$
$$DS_{t+1}^T \doteq (S_t, S_{t+1}, \ldots, S_{T-1}),$$
$$DA_{t+1}^T \doteq (A_t, A_{t+1}, \ldots, A_{T-1}),$$

where DS denotes the time-delay operator.

Remark 1. *These definitions imply that the original sequence, S_{t+1}^T, includes the time-delayed sequence, DS_{t+1}^T, except for its first component, S_t. The union between S_{t+1}^T and DS_{t+1}^T is a sequence S_t^T with length $t + 1$.*

2.1.1. BELLMAN RECURSION FOR VALUE

Extending the interaction model in Section 2.1 by introducing the reward function, $r(S_{t+1} \mid A_t, S_t)$, makes the agent to plan 'consciously' its actions by adapting the policy, $\pi(A_t \mid S_t)$, in order to achieve a higher reward in a given time-horizon. The optimal policy, $\pi(A_t \mid S_t)$, that guarantees the maximal expected reward, is found by the *policy iteration algorithm*, (Sutton & Barto, 1998), where a value of $\pi(A_t \mid S_t)$ is estimated by the value function. It can be computed recursively:

$$Q^\pi(S_t, A_t) = \left\langle r(S_{t+1}; A_t, S_t) \right\rangle_{p(S_{t+1}|S_t, A_t)} \qquad (3)$$
$$+ \gamma_t \left\langle Q(S_{t+1}, A_{t+1}) \right\rangle_{\pi(A_{t+1}|S_{t+1})p(S_{t+1}|S_t, A_t)},$$

where, γ_t is the discounting factor. The Bellman recursion is an fundamental concept in optimal planning. A comprehensive review of reinforcement learning can be found in (Bertsekas, 1995; Sutton & Barto, 1998; Puterman, 2014).

2.2. Causal Conditioning

A random process can be conditioned by another random process either causally or non-causally, respectively:

$$p(A_{t+1}^T \parallel S_{t+1}^T) \doteq \prod_{i=t+1}^T p(A_i \mid A_{t+1}^{i-1}, \mathbf{S_{t+1}^i}) \qquad (4)$$

$$p(A_{t+1}^T \mid S_{t+1}^T) \doteq \prod_{i=t+1}^T p(A_i \mid A_{t+1}^{i-1}, \mathbf{S_{t+1}^T}). \qquad (5)$$

The causal conditioning assumes *time-synchronization* between the sequences. A random process can be causally-conditioned by a random variable at a particular time as well. Compare the following causal

$$p(A_{t+1}^T \parallel S_{t+\tau}) = \prod_{i=t+1}^{t+\tau-1} p(A_i \mid A_{t+1}^{i-1}) \prod_{i=t+\tau}^T p(A_i \mid A_{t+1}^{i-1}, S_{t+\tau}),$$

vs. non-causal scalar conditioning:

$$p(A_{t+1}^T \mid S_{t+\tau}) = \prod_{i=t+1}^T p(A_i \mid A_{t+1}^{i-1}, S_{t+\tau}).$$

A random sequence can be conditioned simultaneously by another random sequence and by a scalar random variable. In particular, the following two expressions will be useful in the next section:

$$p(A_{t+1}^T \parallel S_t^T, A_t) = \prod_{i=t+1}^T p(A_i \mid S_t^i, A_{t+1}^{i-1}, A_t)$$

$$p(A_{t+1}^T \parallel S_t^{T-1}, A_t) = \prod_{i=t+1}^T p(A_i \mid S_t^{i-1}, A_{t+1}^{i-1}, A_t),$$

which, due to the graph in Figure 1, are reduced to:

$$p(A_{t+1}^T \parallel S_t^T, A_t) = \prod_{i=t+1}^T p(A_i \mid S_i) \qquad (6)$$

$$p(A_{t+1}^T \parallel S_t^{T-1}, A_t) = \prod_{i=t+1}^T p(A_i \mid S_{i-1}, A_{i-1}). \qquad (7)$$

2.2.1. CAUSAL ENTROPY

Following the causally conditioned distribution (4), the causally conditioned entropy is defined by:

$$H(A_{t+1}^T \parallel S_{t+1}^T) = \left\langle p(A_{t+1}^T \parallel S_{t+1}^T) \right\rangle_{p(A_{t+1}^T, S_{t+1}^T)}. \qquad (8)$$

The causally conditioned entropy satisfies the chain rule, (and other properties of entropy, (Kramer, 1998)):

$$H(A_{t+1}^T \parallel S_{t+1}^T) \sum_{i=t+1}^T H(A_i \mid S_{t+1}^i, A_{t+1}^{i-1}). \qquad (9)$$

The causally conditioned entropy enables us to define the directed (causal) information between two sequences of random variables.

2.2.2. DIRECTED INFORMATION

The Shannon information between two sequences of random variables, X_1^T, and, Y_1^T, is given by a decrease in the entropy of one variable by the conditional entropy of this variable (Cover & Thomas, 2012):

$$\mathcal{I}\Big[X; Y\Big] \doteq H(X) - H(X \mid Y). \qquad (10)$$

The directed information is defined similarly as a decrease in the entropy. But, in this case, the conditional entropy is the causally conditional entropy (Massey & Massey, 2005; Kramer, 1998):

$$\vec{\mathcal{I}}\Big[X \to Y\Big] \doteq H(X) - H(X \parallel Y). \qquad (11)$$

The essential difference between ordinary mutual information (10) and directed information (11) is the lack of sym-

metry:

$$\vec{\mathcal{I}}\Big[X \to Y\Big] \neq \vec{\mathcal{I}}\Big[Y \to X\Big]. \tag{12}$$

A comprehensive review of directed information properties can be found in (Kramer, 1998; Marko, 1973; Massey & Massey, 2005).

3. Information from Environment to Agent

We equip the agent with the ability to remember the previous state and the action. Memory is a desired feature of intelligent agents. Memoryless agents can act only reactively without learning from the experience. Specifically, we model the memory by *causally-conditioning* the directed information flow from the environments to the agent. Assuming that at time t the environment and the agent action were S_t and A_t, respectively, the causally-conditioning directed information is given by:

$$\vec{\mathcal{I}}\Big[S_{t+1}^T \to A_{t+1}^T \parallel DS_{t+1}^T, A_t\Big] =$$
$$= \Big\langle \ln\Big[\frac{p(A_{t+1}^T \parallel S_{t+1}^T, DS_{t+1}^T, A_t)}{p(A_{t+1}^T \parallel DS_{t+1}^T, A_t)}\Big]\Big\rangle_{p(S_{t+1}^T, A_{t+1}^T \mid S_t, A_t)},$$

which, following Remark 1, is equivalent to:

$$= \Big\langle \ln\Big[\frac{p(A_{t+1}^T \parallel S_t^T, A_t)}{p(A_{t+1}^T \parallel S_t^{T-1}, A_t)}\Big]\Big\rangle_{p(S_{t+1}^T, A_{t+1}^T \mid S_t, A_t)}, \tag{13}$$

and, according to Eq. (6) and Eq. (7) is decomposed to:

$$= \Big\langle \ln\Big[\frac{\prod_{i=t+1}^{T} p(A_i \mid S_i)}{\prod_{i=t+1}^{T} p(A_i \mid S_{i-1}, A_{i-1})}\Big]\Big\rangle_{p(S_{t+1}^T, A_{t+1}^T \mid S_t, A_t)}$$

$$= \sum_{i=t+1}^{T} \Big\langle \ln\Big[\frac{p(A_i \mid S_i)}{p(A_i \mid S_{i-1}, A_{i-1})}\Big]\Big\rangle_{p(S_{t+1}^T, A_{t+1}^T \mid S_t, A_t)}. \tag{14}$$

3.1. Bellman Equation for $\vec{\mathcal{I}}$[States $\to$ Actions]

The expression (14) enables us to write the directed information recursively:

$$\vec{\mathcal{I}}\Big[S_{t+1}^T \to A_{t+1}^T \parallel DS_{t+1}^T, A_t\Big] =$$
$$= \Big\langle \ln\Big[\frac{p(A_{t+1} \mid S_{t+1})}{p(A_{t+1} \mid S_t, A_t)}\Big]\Big\rangle_{p(S_{t+1}, A_{t+1} \mid S_t, A_t)}$$
$$+ \sum_{i=t+2}^{T} \Big\langle \ln\Big[\frac{p(A_i \mid S_i)}{p(A_i \mid S_{i-1}, A_{i-1})}\Big]\Big\rangle_{p(S_{t+1}^T, A_{t+1}^T \mid S_t, A_t)}$$
$$= \mathcal{I}\left[S_{t+1}; A_{t+1} \mid S_t, A_t\right] \tag{15}$$
$$+ \Big\langle \vec{\mathcal{I}}\Big[S_{t+2}^T \to A_{t+2}^T \parallel DS_{t+2}^T, A_{t+1}\Big]\Big\rangle_{p(S_{t+1}, A_{t+1} \mid S_t, A_t)}.$$

The directed information in (13) is a function of S_t and A_t, whereas the directed information in (15) within the expectation operator $< \cdot >_{p(S_{t+1}, A_{t+1} \mid S_t, A_t)}$ is a function of

S_{t+1} and A_{t+1}. Consequently, the recursive equation in (15) can be decomposed into the first step and all the future steps:

$$\vec{\mathcal{I}}_T(S_t, A_t) = \mathcal{I}\left[S_{t+1}; A_{t+1} \mid S_t, A_t\right] \tag{16}$$
$$+ \Big\langle \vec{\mathcal{I}}_T(S_{t+1}, A_{t+1})\Big\rangle_{\pi(A_{t+1} \mid S_{t+1})p(S_{t+1} \mid S_t, A_t)},$$

where the first term is the local step of the recursion and the second term is the future directed information. This expression comprises the Bellman type equation for the directed information between the environment and the agent. The time horizon of the future state and action trajectories is marked by a subscript in the directed information, $\vec{\mathcal{I}}_T(S_t, A_t)$.

Information processing is an integral part of the interaction between the agent and its environment. An optimal agent strives to accumulate as much value as possible under the constraints of limited bandwidth and information transfer rate. In the next section we show that the Bellman recursion for the directed information and the Bellman recursion for the value naturally combine into a unified Bellman recursion.

4. Unified Bellman Recursion

As mentioned above, the optimal control solution should adhere to the information processing constraints, which are always present in real-life intelligent systems. For example, the limited bandwidth and/or time-delays due to the distance between a controller and a plant are ubiquitous information-theoretic constraints, which may reduce the maximal achievable value.

Consequently, the designer of an intelligent agent should always address the question - *what is the minimal information rate required to achieve the maximal value?* This question is formulated by the unconstrained Lagrangian which comprises the unified Bellman recursion.

We found that the Bellman recursion for the directed information (16) and the Bellman recursion for the value (3) combine naturally to form the unified *Information-Value Bellman Recursion*:

$$\mathcal{G}_T^\pi(S_t, A_t, \beta) = \vec{\mathcal{I}}_T^\pi(S_t, A_t) - \beta Q^\pi(S_t, A_t) =$$

which is decomposed into the first step and all the future steps:

$$= \mathcal{I}_T^\pi\Big[S_{t+1}; A_{t+1} \mid S_t, A_t\Big] \tag{17}$$
$$- \beta \left\langle r(S_{t+1}; S_t, A_t)\right\rangle_{p(S_{t+1} \mid S_t, A_t)}$$
$$+ \Big\langle \vec{\mathcal{I}}_T^\pi(S_{t+1}, A_{t+1})$$
$$- \beta Q^\pi(S_{t+1}, A_{t+1})\Big\rangle_{\pi(A_{t+1} \mid S_{t+1})p(S_{t+1} \mid S_t, A_t)}.$$

Writing explicitly the mutual information and identifying the terms within the second expectation operator with the future averaged value of $\mathcal{G}_T^\pi(S_{t+1}, A_{t+1}, \beta)$, we get the unified Bellman recursion:

$$\mathcal{G}_T^\pi(S_t, A_t, \beta) = \left\langle \pi(A_{t+1} \mid S_{t+1}) \ln\left[\frac{\pi(A_{t+1} \mid S_{t+1})}{p(A_{t+1} \mid S_t, A_t)}\right] \right. \quad (18)$$
$$\left. - \beta r(S_{t+1}; S_t, A_t) \right\rangle_{p(S_{t+1}\mid S_t, A_t)}$$
$$+ \left\langle \mathcal{G}_T^\pi(S_{t+1}, A_{t+1}) \right\rangle_{\pi(A_{t+1}\mid S_{t+1})p(S_{t+1}\mid S_t, A_t)}.$$

This recursive equation can be computed for any T. The stationary policy, $\pi(A \mid S)$ is achieved for $T \to \infty$. In the next section we show how to find the optimal policy, $\pi(A \mid S)$, that maximizes the unified Bellman recursion in (18).

4.1. Optimization Problem

To find the optimal policy (that minimizes the directed information under a constraint on the attained value), one needs to solve the unconstrained minimization of the corresponding Lagrangian in (17).

$$\arg\min_\pi \mathcal{G}_T^\pi(S_t, A_t, \beta) = \arg\min_\pi \left[\vec{\mathcal{I}}_T^\pi(S_t, A_t) - \beta Q^\pi(S_t, A_t)\right]. \quad (19)$$

The minimization with regard to π is done by augmenting the equation in (19) by the Lagrange term for the normalization of $\pi(A \mid S)$, and computing the variational derivative, $\delta\mathcal{G}_T^\pi/\delta\pi$. The minimization with regard to $q(A_{t'} \mid S, A) \doteq p(A_{t'} \mid S, A)$ is done by setting the corresponding variational derivative $\delta\mathcal{G}_T^q/\delta q$ to zero. Straightforward computation of these variational derivatives gives the following set of self-consistent equations:

$$\pi_T^*(A \mid S) = \frac{\exp\left(-\mathcal{G}_T^{\pi^*,q^*}(S_t, A_t, \beta)\right)}{\sum_A \exp\left(-\mathcal{G}_T^{\pi^*,q^*}(S_t, A_t, \beta)\right)}, \quad (20)$$

$$p^*(A' \mid S, A) = \frac{1}{|\mathcal{A}|}, \quad (21)$$

4.2. Stationary Policy

The typical behavior of the agent in its environment is defined by the stationary distribution over the states and the actions. This stationary distribution is given by the policy with the infinite time horizon:

$$\pi_\infty^*(A \mid S) = \frac{\exp\left(-\mathcal{G}_\infty^{\pi^*,q^*}(S_t, A_t, \beta)\right)}{\sum_A \exp\left(-\mathcal{G}_T^{\pi^*,q^*}(S_t, A_t, \beta)\right)}. \quad (22)$$

The unified Bellman recursion enables us to compute the stationary policy by iterating the recursion in Eq. (18) in the same way as for the Bellman recursion for the value

function. The stationary policy for $T \to \infty$ was derived for the LQG setting (Tanaka et al., 2015); in this work we derive the optimal stationary policy for $T \to \infty$ in the MDP setting.

5. Numerical Simulations

To explore a long-term (typical) interaction between an agent and the environment we solve different benchmark RL problems, such as *maze escaping problems* (Sutton & Barto, 1998), using the information-value recursion in Eq. (18).

We emphasize that even though, in these experiments the target goal appears in a particular state (or at a number of states), there are different possible scenarios where a goal state can appear in any state. In these cases, the agent should be prepared to reach any goal. This omni-goal strategy will be discussed in Section 7.

5.1. Maze with predefined targets

To investigate the solutions in (18) we simulated different mazes where the agent is asked to find a target cell. The agent has full information about the world (it knows its exact location within the maze), which is proportional to the logarithm of the maze size. But it cannot utilize all this information to generate the optimal actions due to its slow 'brain'. For example, the agent's sensor retrieves 8 bits of information about its location in each interaction with the environment, whereas the processor can transform only 3 bits of information to the policy.

The dynamics is given by a transition probability $p(S_{t+1} \mid A_t, S_t)$. If a certain action at time t is impossible, such as stepping into a wall, the agent remains in its current position. The target state is an absorbing state - the agent stays in this state for any action $p(S_{target} \mid A_t, S_t) = \delta_{S_{target}, S_t}$. The agent receives a reward, 0, when it is in the target state. In all other states the reward is given by $r(S_{t+1} \mid A_t, S_t) = -1$. This is the general setting for maze escaping problems, (Sutton & Barto, 1998).

Figure 2 shows the value function, and the directed information for different values of parameter β. The optimal solutions to equation (18) are characterized by *phase-transitions* - the qualitative difference in the typical behavior as a function of the value of parameter β. The optimal path is indicated by the solid white line from the start state, marked by S, to the goal state, marked by G.

The low values of β correspond to the high stochasticity in the system. The more stochastic the system is, the harder it is for the agent to affect the environment. Specifically, the agent prefers not to take the risk of going through the narrow door for $\beta = 0.005$, as shown in Figure (2-b). However, it does go through the narrow door for $\beta = 0.005$, as shown in Figure (2-a).

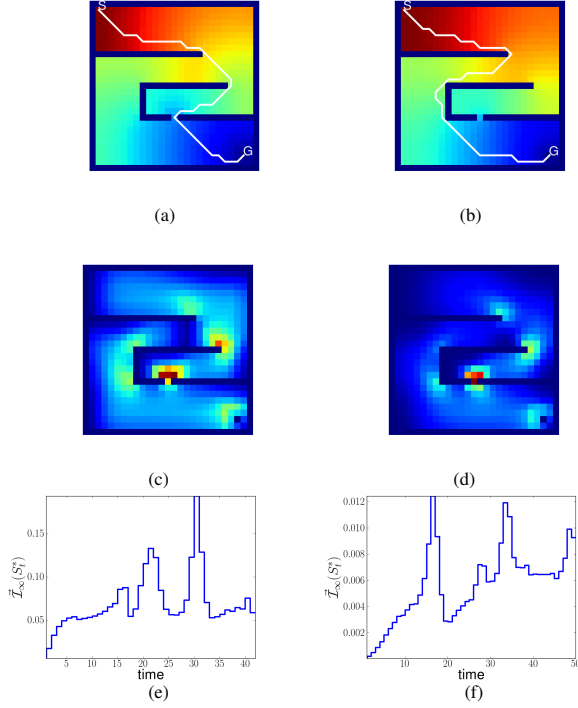The typical distribution of the directed information along the optimal path is shown in the bottom plots. The informa-

(a)   (b)

(c)   (d)

(e)   (f)

Figure 2. Numerical simulation of the maze problem with 625 cells for an infinite time-horizon, $T \to \infty$. The sub figures (a), (c), and (e) correspond to $\beta = 0.05$. They show the value landscape, the directed information landscape, and the directed information along the optimal path. The sub figures (b), (d), and (f) correspond to $\beta = 0.05$. A phase-transition occurs between $\beta = 0.05$ and $\beta = 0.005$, where a typical behavior *"to pass through the narrow door"*, changes to another qualitatively different behavior *"to take a longer path without entering the narrow door"*. The peaks at the sub figures (e) and (f) correspond to the states with high information consumption. For instance, as seen at (e), the 31st state along the optimal path (the entrance to the narrow door), corresponds to the maximal information rate that the agent should process to reach the goal state.

tion is higher when the agent needs to move carefully, e.g., near the corners and/or through the narrow door. The solution is derived for the infinite time horizon, which enables us to explore the typical stationary policy of the agent.

As is seen in the sub figure (d), the agent needs different amounts of information at different times over the optimal path. For example, it must be more careful (needs more information), at the narrow hole in order to exit the compartment in the middle of the maze. This is the maximum of the minimal directed information from the states to the actions, which should be taken into consideration when choosing a processor for the agent.

6. Information from Agent to Environment

In this section we address the question of the maximal directed information from the agent to the environment,

which is a dual question to the question of the maximal directed information from the environment to the agent. These two directed informations differ, because the directed information functional is asymmetric in general. We argue and demonstrate by numerical simulations that the maximal directed information between the agent and the environment is useful for acting optimally without an externally provided reward (unknown targets).

6.1. Background

The information capacity of the channel from the agent's actions to the environment states was shown to be useful in solving various AI benchmark problems without providing an external reward/utility function, (Salge et al., 2014; Klyubin et al., 2005b;a; Mohamed & Rezende, 2015). This capacity, known as *empowerment*, evaluates a maximal number of environment states that the agent can potentially achieve by its actions.

The higher the value of empowerment is, the more potential states are reachable by actions. As in the empowerment method, when a goal state is not predefined but may occur in any state as in Chess, Go and other interactive games, the optimal strategy for the agent is to keep as many *reachable states* as possible. As was shown in (Klyubin et al., 2005a) Figure 5 and Figure 6, there is a qualitative similarity between the landscape of the empowerment and the landscape of the average shortest distance from a given state to any other state.

In this paper we extend the empowerment method by considering in the feedback control policy by applying the directed information from the action trajectories to the environment trajectories. In the next section we describe the causal conditioning of a sequence of states of length T on a sequence of actions of length T. We derive the Bellman-type recursion for the directed information from the action to the states, which together with the Bellman recursion from the states to the action in Eq.(18) provide a complete picture of the bi-directional information exchange between the agent and the environment.

6.1.1. CAUSAL CONDITIONING

We assume that at time t the environmental state is S_t. Given S_t, the agent computes action A_t, and the environment moves to S_{t+1}, according to the given transition probability, $p(S_{t+1} \mid S_t, A_t)$. This interaction repeats $T-1$ cycles, driving the system to its final state, S_T, by a sequence of actions A_1^{T-1}.

The joint probability between A_t^{T-1} and S_{t+1}^T conditioned on S_t is given by $p(S_{t+1}^T, A_t^{T-1} \mid S_t)$. The causal dependency between the time series, A_t^{T-1}, and S_t^T differs from the non-causal one in its conditional probability (for details see e.g., (Kramer, 1998; Massey & Massey, 2005; Marko, 1973; Permuter et al., 2009)), as following:

$$p(S_{t+1}^T \mid A_t^{T-1}, S_t) = \prod_{i=t}^{T} p(S_i \mid S_t^{i-1}, A_t^{T-1}), \qquad (23)$$

$$p(S_{t+1}^T \| A_1^{T-1}, S_t) = \prod_{i=t}^{T} p(S_i \mid S_t^{i-1}, A_t^{i-1}). \qquad (24)$$

The join probability, $p(S_{i+1}^T, A_i^{T-1} \mid S_i)$, is given by:

$$p(S_{t+1}^T, A_t^{T-1} \mid S_t) =$$
$$= p(S_{t+1}^T \| A_t^{T-1}, S_t) p(A_t^{T-1} \| S_t^{T-1}), \qquad (25)$$

which is decomposed alternatively according to Bayes' rule as:

$$= p(A_t^{T-1} \mid S_{t+1}^T, S_t) p(S_{t+1}^T \mid S_t). \qquad (26)$$

6.1.2. DIRECTED INFORMATION

The two alternative decompositions of the joint distribution, $p(S_{t+1}^T, A_t^{T-1} \mid S_t)$, enable us to decompose the directed information as follows

$$\mathcal{I}\left[A_t^{T-1} \to S_{t+1}^n \mid s_t\right] = \qquad (27)$$
$$= \left\langle \ln \frac{p(S_{t+1}^T \| A_t^{T-1}, S_t)}{p(S_{i+1}^n \mid S_i)} \right\rangle_{p(S_{i+1}^n, A_i^{n-1} \mid s_i)}$$
$$= \left\langle \ln \frac{\prod_{j=i}^{n-1} p(A_j \mid S_j, S_{j+1})}{\prod_{j=i}^{n-1} \pi(A_j \mid S_j)} \right\rangle_{p(S_{i+1}^n, A_i^{n-1} \mid s_i)},$$

which follows from the alternative decompositions of the joint distribution in (25) and (26):

$$\frac{p(S_{t+1}^T \| A_t^{T-1}, S_t)}{p(S_{t+1}^T \mid S_t)} = \frac{p(A_t^{T-1} \mid S_{t+1}^T, S_t)}{p(A_t^{T-1} \| S_t^{T-1})}, \qquad (28)$$

and the graph property:

$$1 \le i < n : A_i \perp\!\!\!\perp \left(A_1^{i-1}, A_{i+1}^{n-1}, S_1^{i-1}, S_{i+1}^n\right) \mid S_i, S_{i+1}.$$

This decomposition leads to the Bellman equation for the directed information from the action trajectories to the state trajectories.

6.2. Bellman Equation for $\vec{\mathcal{I}}[\text{Actions} \to \text{States}]$

Separating the first term in (27) from the rest we get the following recursion:

$$\mathcal{I}\left[A_i^{n-1} \to S_{i+1}^n \mid s_i\right] = \mathcal{I}\left[A_i; s_i \mid S_{i+1}\right] \qquad (29)$$
$$+ \left\langle \mathcal{I}\left[A_{i+1}^{n-1} \to S_{i+2}^n \mid S_{i+1}\right]\right\rangle_{p(S_{i+1} \mid s_i)}.$$

where $\mathcal{I}\left[A_t; S_t \mid S_{t+1}\right]$ is the mutual information between the t-th state and action, conditioned on the $(t+1)$-th state. The conditional directed information in (29) is a function

of the state. To improve readability, the recursion in (29) may be represented concisely as follows, (where the right arrow in $\vec{\mathcal{I}}$ distinguishes the directed information from the ordinary/symmetric mutual information):

$$\vec{\mathcal{I}}(s_t) = \mathcal{I}\left[A_t; s_t \mid S_{t+1}\right] + \left\langle \vec{\mathcal{I}}(S_{t+1}) \right\rangle_{p(S_{t+1} \mid s_t)}, \qquad (30)$$

where the free distribution is the feedback policy distribution, $\pi(A \mid S)$. Assuming stationarity, the time index in (30) may be reduced to:

$$\vec{\mathcal{I}}(s) = \mathcal{I}\left[A; s \mid S'\right] + \left\langle \vec{\mathcal{I}}(S') \right\rangle_{p(S' \mid s)}. \qquad (31)$$

The maximum of $\vec{\mathcal{I}}(s)$ with regard to $\pi(A \mid S)$ gives the maximal number of states that the agent can achieve by its action. In the next section we provide the formal solution.

6.3. Optimal solution

To find the maximum of the directed information in (31), we apply the alternating maximization between the policy distribution, $\pi(A_i \mid S_i)$, and the *inverse-channel* distribution $p(A_i \mid S_i, S_{i+1})$ (which is conventionally denoted by $q(A_i \mid S_i, S_{i+1})$), as in the standard problem for finding the channel capacity, (Cover & Thomas, 2012; Blahut, 1972). The optimal formal solution is given by the following set of self-consistent equations:

$$q^*(A \mid S, S') = \frac{p(S' \mid S, A)\pi^*(A \mid S)}{\sum_A p(S' \mid S, A)\pi^*(A \mid S)}, \qquad (32)$$

$$\vec{\mathcal{I}}_{\pi^*}[S] = \mathcal{I}_{\pi^*}[A, S' \mid S] + \left\langle \vec{\mathcal{I}}_{\pi^*}[S']\right\rangle_{p(S' \mid A, S)\pi^*(A \mid S)}, \qquad (33)$$

$$\pi^*(A \mid S) = \qquad (34)$$
$$= \frac{\exp \sum_{S'} p(S' \mid S, A)\left(\ln q^*(A \mid S, S') + \vec{\mathcal{I}}^*[S']\right)}{\sum_A \exp \sum_{S'} p(S' \mid S, A)\left(\ln q^*(A \mid S, S') + \vec{\mathcal{I}}^*[S']\right)}.$$

This set of self-consistent equations is an extension of the common Blahut-Arimoto algorithm by the directed information of the future action and state trajectories, which appears in the exponent of $\pi^*(A \mid S)$ in Eq.34. A numerical solution is derived by iterating equations (32-34) until convergence. The typical convergence is shown by the numeric simulations, provided in the next section.

7. Numerical Simulations

To demonstrate the optimal solution given by equations (32-34) we consider the problem of the optimal location of a firehouse in a city. We model the city by a grid world, where a fire engine is an agent moving across the city according to a given transition probability, $p(S_{t+1} \mid S_t, A_t)$. The goal is to locate the firehouse at a position where the firefighters can reach each block in the city in the fastest way on average.
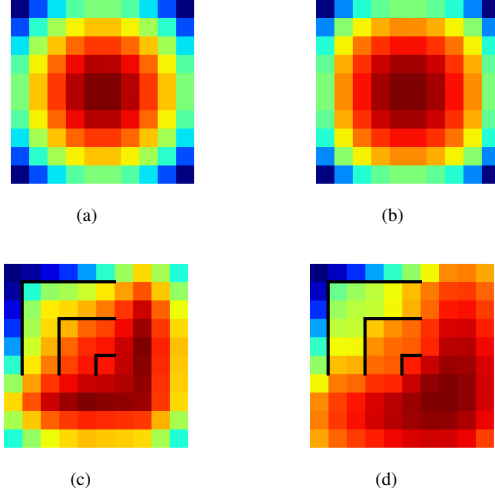
(a)   (b)

(c)   (d)

Figure 4. Convergence of the Bellman recursion for the directed information. x-axis: number of blocks, y-axis: number of iterations until convergence.

Figure 3. Sub figures (a) and (c) show the exact values of the minimal average path from a given block to all the other blocks in the city. Sub figures (b) and (d) show the value of the maximal directed information from the agent to the environment in the long-time horizon on every block in the city. Black lines represent the fences in the city. Red corresponds to high values.

This is an example where the agent should be prepared optimally to reach any possible target state on average, rather than to be able to get to a particular state (or a number of states).

The naive solution is to compute the average shortest distance between each block in the city and all the other blocks, and to locate the firehouse on the block with the minimal average shortest distance. The algorithms for solving the all-pairs shortest path problem have a time complexity of $O(V^3)$ (Cormen, 2009), where V is the number vertices in a graph.

The sub figures (3.a) and (3.c) show the exact values of the average shortest paths from a particular block to all the other blocks in two different cities, a city without any walls, and a city with perpendicular black walls, respectively. The sub figures (3.b) and (3.d) show the values of the directed information computed by iterating the self-consistent equation (32-34). Red blocks denote high values, and blue blocks denote low values.

The two approaches suggest qualitatively similar patterns on the location of the firehouse. The preferable locations are the brightest red blocks. In the case of the city without walls both solutions give the same block.

Figure (4) shows the convergence of the iteration of the self-consistent equations (32-34). The convergence is sublinear in the number of blocks in the city, where each iteration consumes $O\big(|\mathcal{S}| \times |\mathcal{A}|\big)$ multiplications of real numbers, which is overall faster in order of magnitude than the Floyd/Dijkstra-type algorithms. Convergence is said to oc-
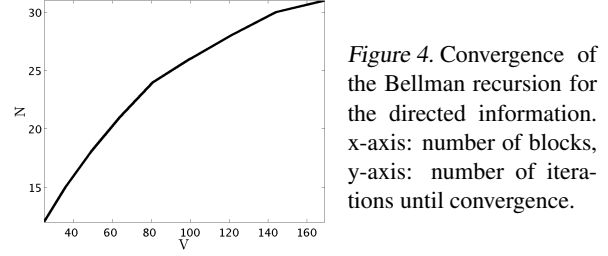
cur when the relative change in the directed information between two consequent iterations is smaller than tolerance, $\tau = 10^{-6}$, $abs(\vec{\mathcal{I}}_{k+1} - \vec{\mathcal{I}}_k)/\vec{\mathcal{I}}_k \leq \tau$, where $\mathcal{I}_{k+1}$ is given by Eq.31 in the k-th iteration.

8. Conclusion

The feedback interaction between agent and the environment includes two asymmetric information channels. Directed information from the environment constraints the maximal expected reward obtained by the agent in the standard RL setting. Directed information is a functional of a joint probability distribution of agent's action trajectories and environmental state trajectories in a given time horizon. As a result, the value of the directed information depends on a particular time horizon. However, it is worth exploring the typical behavior of an agent in an infinite time horizon, when all transients events have settled down or become statistically insignificant. In this work we derive the typical behavior of an agent that strives to increase its expected reward under the limitations of directed information. This is shown to be possible through the unified Information-Value Bellman-type recursion, which we represent in this work.

The second channel transfers directed information from the agent to the environment. It has been shown that the capacity of this channel is useful for solving various AI test bench problems, when the agent needs to be ready to solve a problem optimally on average for any possible target. We derived a Bellman recursion for this channel as well, which enables us to explore the typical behavior of the agent in an infinite time horizon.

A direct continuation of this work would be to explore the entire action-perception cycle of the interaction between the intelligent agent and the environment when both information channels operate simultaneously.

References

Bertsekas, Dimitri P., et al. *Dynamic programming and optimal control*. Athena Scientific Belmont, MA, 1995.

Blahut, Richard. Computation of channel capacity and rate-distortion functions. *IEEE transactions on Information Theory*, 18(4):460–473, 1972.

Cormen, Thomas H. *Introduction to algorithms*. MIT press, 2009.

Cover, Thomas M and Thomas, Joy A. *Elements of information theory*. John Wiley & Sons, 2012.

Kappen, Hilbert J, Gómez, Vicenç, and Opper, Manfred. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.

Klyubin, Alexander S, Polani, Daniel, and Nehaniv, Chrystopher L. All else being equal be empowered. In *European Conference on Artificial Life*, pp. 744–753. Springer, 2005a.

Klyubin, Alexander S, Polani, Daniel, and Nehaniv, Chrystopher L. Empowerment: A universal agent-centric measure of control. In *Evolutionary Computation, 2005. The 2005 IEEE Congress on*, volume 1, pp. 128–135. IEEE, 2005b.

Kramer, Gerhard. *Directed information for channels with feedback*. PhD thesis, University of Manitoba, Canada, 1998.

Marko, Hans. The bidirectional communication theory–a generalization of information theory. *IEEE Transactions on communications*, 21(12):1345–1351, 1973.

Massey, James L and Massey, Peter C. Conservation of mutual and directed information. In *Information Theory, 2005. ISIT 2005. Proceedings. International Symposium on*, pp. 157–158. IEEE, 2005.

Mohamed, Shakir and Rezende, Danilo Jimenez. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in neural information processing systems*, pp. 2125–2133, 2015.

Perkins, Theodore J and Barto, Andrew G. Lyapunov design for safe reinforcement learning. *Journal of Machine Learning Research*, 3(Dec):803–832, 2002.

Permuter, Haim Henry, Weissman, Tsachy, and Goldsmith, Andrea J. Finite state channels with time-invariant deterministic feedback. *IEEE Transactions on Information Theory*, 55(2):644–662, 2009.

Polani, Daniel, Nehaniv, C, Martinetz, Thomas, and Kim, Jan T. Relevant information in optimized persistence vs. progeny strategies. In *Artificial Life X: Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems*, pp. 337–343. Citeseer, 2006.

Puterman, Martin L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Rubin, Jonathan, Shamir, Ohad, and Tishby, Naftali. Trading value and information in mdps. In *Decision Making with Imperfect Decision Makers*, pp. 57–74. Springer, 2012.

Salge, Christoph, Glackin, Cornelius, and Polani, Daniel. Empowerment–an introduction. In *Guided Self-Organization: Inception*, pp. 67–114. Springer, 2014.

Sutton, Richard S and Barto, Andrew G. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

Tanaka, Takashi and Sandberg, Henrik. Sdp-based joint sensor and controller design for information-regularized optimal lqg control. In *Decision and Control (CDC), 2015 IEEE 54th Annual Conference on*, pp. 4486–4491. IEEE, 2015.

Tanaka, Takashi, Esfahani, Peyman Mohajerin, and Mitter, Sanjoy K. Lqg control with minimal information: Three-stage separation principle and sdp-based solution synthesis. *arXiv preprint arXiv:1510.04214*, 2015.

Tatikonda, Sekhar and Mitter, Sanjoy. Control under communication constraints. *IEEE Transactions on Automatic Control*, 49(7):1056–1068, 2004.

Tishby, Naftali and Polani, Daniel. Information theory of decisions and actions. In *Perception-action cycle*, pp. 601–636. Springer, 2011.

Todorov, Emanuel et al. Linearly-solvable markov decision problems. In *NIPS*, pp. 1369–1376, 2006.