# Trading value and information in MDPs

Jonathan Rubin, Ohad Shamir and Naftali Tishby

**Abstract** Interactions between an organism and its environment are commonly treated in the framework of Markov Decision Processes (MDP). While standard MDP is aimed at maximizing expected future rewards (*value*), the circular flow of information between the agent and its environment is generally ignored. In particular, the information gained from the environment by means of perception and the information involved in the process of action selection are not treated in the standard MDP setting. In this paper, we focus on the *control information* and show how it can be combined with the reward measure in a unified way. Both of these measures satisfy the familiar Bellman recursive equations, and their linear combination (the *free-energy*) provides an interesting new optimization criterion. The tradeoff between value and information, explored using our INFO-RL algorithm, provides a principled justification for stochastic (soft) policies. We use computational learning theory to show that these optimal policies are also robust to uncertainties in the reward values.

## 1 Introduction

Modeling an agent's interaction with the environment is commonly treated in the framework of Markov Decision Processes: given a statistical model of the environment which includes transition and rewarding rules, the agent is expected to find an optimal policy which will maximize its future accumulated rewards [8].

Jonathan Rubin
Hebrew University Jerusalem, e-mail: rubinj@cs.huji.ac.il

Ohad Shamir
Microsoft Research Cambridge e-mail: ohadsh@microsoft.com

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

While this framework is rather general, the explicit flow of information between the agent and its environment is ignored. This circular flow of information (also referred in the literature as the *perception-action cycle* [4]) includes two terms: the information gained from the environment in response to the agent's actions and the *control information* associated with the decisions the agent make. The first term corresponds to the flow of information from the environment to the agent (sensory perception) and the second term corresponds to the flow of information from the agent back to the environment (by means of action selection).

In this work we focus mainly on the *control information* term (presented in Section 2). We show how this information measure can be treated side-by-side with traditional *value* measures used in control theory. This treatment gives rise to a tradeoff between *value* and *information*, which differs from standard MDP as the information term is an explicit function of the unknown policy itself [10]. Here we develop this framework further. In Section 3 we show that this new optimization problem can be solved by dynamic programming with *global* convergence to a unique optimum, using our INFO-RL algorithm. In a special setting of deterministic states-transition model we show that the problem reduces to a simple linear form. We illustrate our approach on a simple grid-world navigation task in Section 4.

Moreover, trading *value* and *information* is not restricted to problems where information explicitly carries a price tag. In Section 5 we consider a setting, in which the exact parameters of the MDP are not fully known. Utilizing recent theorems from computational learning theory [6] we show how the control information actually serves as the proper regularization term leading to a more robust policy.

Our approach is related to other algorithms that combine information theoretic functions with optimal control [3, 11, 5], but its setting and scope are different. In our case the information theoretic components, quantifying the information flow between the agent and its environment, serves as an integral part of the reward that drives the action selection. Treating information quantities thus allows us to explore the tradeoff between value and information in an explicit principled way.

## 2 Preliminaries

This section presents our notation and introduces the *control information* term and its motivation.

A word about notations, we use $\mathbb{E}_{x|y}[\cdot]$ to denote expectations with respect to the distribution of $x$ given $y$.

### *2.1 Markov Decision Processes*

A finite Markov decision process (MDP) is defined by a tuple $\langle \mathcal{S}, \mathcal{A}, R, P \rangle$ where: $\mathcal{S} = \{1, ..., n\}$ is a finite set of $n$ states; $\mathcal{A}$ is a finite set of actions; $R$ is a scalar reward

function, such that $R(s,a)$ represents the immediate reward obtained in state $s$ after taking action $a$; and $P$ is a Markovian transition model, where $P_{s,a}(s')$ represents the probability of transition to state $s'$ when taking action $a$ at state $s$. The agent chooses action $a$ in state $s$ according to a stationary probability $\pi_s(a)$, which is known as the agent's *policy*.

In this work, we focus on a setting where the aim is to reach some *terminal state* with maximum expected rewards. This is known in the literature as the 'stochastic shortest path' or 'first exit' problem. To this aim, we introduce a new terminal state $s_{goal}$ which is an absorbing state, $P_{s_{goal},a}\left(s_{goal}\right)=1$ for all $a\in\mathcal{A}$. We denote the set of MDP states including this terminal state by $\mathcal{S}^{+}=\left\{1,2,...,n,s_{goal}\right\}$. We assume here that all 'rewards' are negative (i.e., 'costs'): $R\left(s_{goal},a\right)<0$. for all $a\in\mathcal{A},\ s\in\mathcal{S}$ and that the absorbing state is 'cost free', $R\left(s_{goal},a\right)=0$.

We define a *proper policy* as a policy with the following property: there is a positive number $m<\infty$, such that for any initial state, the probability of reaching the terminal state after at most $m$ steps is some $\varepsilon>0$. In particular, this guarantees that we reach the terminal state $s_{goal}$ with probability one after finitely many steps, regardless of the initial state.

The *value function* of a policy $\pi$ is defined as the expected accumulated rewards for executing $\pi$ starting from state $s_0$,

$$V_{\pi}(s_0)=\lim_{T\to\infty}\mathbb{E}\left[\sum_{t=0}^{T-1}R(s_t,a_t)\right] \tag{1}$$

where the expectation is taken with respect to the probability of all future trajectories, starting in $s_0\in\mathcal{S}$ and executing the stationary policy $\pi$ thereafter,

$$\Pr(a_0,s_1,a_1,s_2,...|s_0)=\prod_{t\geq 0}\pi_{s_t}(a_t)P_{s_t,a_t}(s_{t+1}) \tag{2}$$

Note that $V_{\pi}\left(s_{goal}\right)=0$ under any policy $\pi$, as the terminal state is 'cost free'.

The *optimal value function*, $V^{*}(s)=\max_{\pi}V_{\pi}(s)$, is defined as the maximal achievable value (for each state $s$) by any stationary policy. This optimal value function is the unique fixed-point solution of Bellman's optimality criterion [1],

$$\begin{aligned}V^{*}(s)&=\max_{\pi_s(\cdot)}\sum_{a}\pi_s(a)\sum_{s'}P_{s,a}\left(s'\right)\left[R(s,a)+V^{*}\left(s'\right)\right]\\&=\max_{a}\left[R(s,a)+\sum_{s'\in\mathcal{S}}P_{s,a}\left(s'\right)V^{*}\left(s'\right)\right]\end{aligned}$$

In this case, a deterministic optimal policy $\pi^{*}$ can be obtained by acting greedily with respect to $V^{*}$: at each state $s$ the selected action maximizes the optimal *states-actions value* function $Q^{*}(s,a)$,

$$Q^*(s,a) = R(s,a) + \sum_{s' \in \mathcal{S}} P_{s,a}(s') V^*(s')$$

$$\pi_s^*(a) = \begin{cases} 1 & a = \arg\max_{a'} Q^*(s,a') \\ 0 & otherwise \end{cases}$$

## 2.2 Control information

We consider scenarios where the controller and the actuator are separated by some communication channel. This could be transmitting radio signals to a distant robot or sending control commands from the brain to the muscles through the nervous system. Sending information through a channel doesn't come free, and an optimal policy should take these communication costs into account.

We use information theory to quantify the expected (information) cost for executing policy $\pi$ in state $s \in \mathcal{S}$ as,

$$\Delta I(s) = \sum_a \pi_s(a) \log \frac{\pi_s(a)}{\rho_s(a)}$$

and define $\Delta I = 0$ at the terminal state.

This measure is the relative entropy at state $s$ between the controller's policy $\pi_s(a)$ and some *default policy* $\rho_s(a)$. This default policy could represent a naive policy used by the actuator in the absence of information from the controller. Without loss of generality we set $\rho_s(a)$ to be uniformly distributed over the available actions at state $s$. This measure, $\Delta I(s)$, corresponds to the minimal number of bits required to describe the outcome of the random variable $a \sim \pi_s(\cdot)$. It also corresponds to the *minimal capacity* of a communication channel (between the controller and the actuator) capable of transmitting this control without an error [2]. Thus, it serves here as a measure for the *cost of control*. For example, when only two actions are available, a deterministic control (such as 'turn left here') 'costs' $\Delta I = 1$ bit, while executing a 'random walk' control is essentially free, $\Delta I = 0$. It follows that sending deterministic control is more expensive then sending vague (stochastic) control through the communication channel. In cases where different actions result in little change in the expected value – stochastic control might suffice.

In analogy with the value function $V_\pi(s_0)$, we define the total *control information* involved in executing policy $\pi$ starting from $s_0$,

$$I_\pi(s_0) = \lim_{T\to\infty} \mathbb{E}\Big[\Delta I(s_t)\Big] \tag{3}$$

$$= \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \log \frac{\pi_{s_t}(a_t)}{\rho_{s_t}(a_t)}\right] \tag{4}$$

$$= \lim_{T\to\infty} \mathbb{E}\left[\log \frac{\Pr(a_0,a_1,...,a_{T-1}|s_0;\pi)}{\Pr(a_0,a_1,...,a_{T-1}|s_0;\rho)}\right] \tag{5}$$

with the expectation taken with respect to all future trajectories as in Eq. (2), and $I_\pi = 0$ at the terminal state.

Deterministic policies, like those resulting from maximizing the *value function* $V_\pi$ alone in the standard MDP framework (acting greedily with respect to the value function), are usualy expensive in terms of the *control information* $I_\pi$. The tradeoff between these two quantities is the subject of our next section.

## 3 Trading value and information

We define optimal policies as policies that achieve maximal *value* given a constraint on the *control information*. In this way, optimal policies reflect a balance between maximizing expected rewards (value) and minimizing the information cost involved in control. To this aim we define a *free-energy* function and show how it can be used to derive optimal policies and explore the tradeoff between value and control information.

### *3.1 Free-energy formulation*

Borrowing terminology from statistical mechanics, we define a *free-energy* function $F_\pi(s_0;\beta)$ that combines both the value term $V_\pi(s_0)$ and our information term $I_\pi(s_0)$,

$$\begin{aligned} F_\pi(s_0;\beta) &= I_\pi(s_0) - \beta V_\pi(s_0) \\ &= \lim_{T\to\infty} \mathbb{E}\sum_{t=0}^{T-1}\left[\log \frac{\pi_{s_t}(a_t)}{\rho_{s_t}(a_t)} - \beta R(s_t,a_t)\right] \end{aligned}$$

with the expectation taken with respect to all future trajectories as in Eq. (2), and $F_\pi = 0$ at the terminal state. The parameter $\beta > 0$ controls the tradeoff between information and value.

In practice, by minimizing the free-energy with respect to the policy $\pi$ for a given $\beta > 0$, we solve the following constrained optimization problem. Out of all policies with control information below some threshold, find the policy achieving maximal value,

$$\max_\pi V_\pi(s_0) \quad s.t. \;\; I_\pi(s_0) \leq threshold$$

This formulation is similar to the one used in *rate-distortion theory* (RDT) in information theory (c.f., Chapter 13 of [2]) with the expected value replacing the expected block distortion.

In analogy with the value function, the free-energy can also be shown to satisfy Bellman's optimality equation as suggested in the following.

**Theorem 1.** *The optimal free-energy vector $F^*(s;\beta)$ satisfies Bellman's equation, $F^* = \mathcal{B}F^*$, where the mapping $\mathcal{B}:\mathbb{R}^n \mapsto \mathbb{R}^n$ is defined as follows,*

$$[\mathcal{B}F](s) = \min_{\pi_s(\cdot)} \sum_{a\in\mathcal{A}} \pi_s(a) \left[ \log \frac{\pi_s(a)}{\rho_s(a)} - \beta R(s,a) + \sum_{s'\in\mathcal{S}} P_{s,a}(s') F(s';\beta) \right] \tag{6}$$

*Furthermore, $F^*$ is the unique solution of this self-consistent equation.*

The proof of this theorem is given in the appendix. Following the theorem, we use standard dynamic programming to solve the modified Bellman's equation. In practice, we start from $F_0$ (a zeros vector) and iteratively apply the mapping $\mathcal{B}$ until convergence to the unique fixed point $F^*(s;\beta)$,

$$F_{k+1}(s;\beta) = [\mathcal{B}F_k](s) \ , \ k = 0,1,... \tag{7}$$

**Lemma.** *Applying the mapping $\mathcal{B}$ on a vector $F \in R^n$ is equivalent to,*

$$[\mathcal{B}F](s) = -\log Z(s;\beta) \tag{8}$$

*where $Z(s;\beta)$ is the partition function,*

$$Z(s;\beta) = \sum_a \rho_s(a) \exp\Big[\beta R(s,a) - \mathbb{E}_{s'|s,a} F(s';\beta)\Big]$$

*Proof.* The minimization in the mapping $\mathcal{B}$ is over the set of *normalized* conditional distributions. For this purpose, we introduce the following Lagrangian,

$$\mathcal{L}[\pi_s(\cdot)] = \sum_a \pi_s(a) \left[ \log \frac{\pi_s(a)}{\rho_s(a)} - \beta R(s,a) + \sum_{s'} P_{s,a}(s') F(s';\beta) \right] + \lambda_s \sum_a \pi_s(a)$$

taking the derivative of $\mathcal{L}$ with respect to $\pi_s(a)$ for a given $a$ and $s$ we obtain,

$$\frac{\delta\mathcal{L}}{\delta\pi_s(a)} = \log \frac{\pi_s(a)}{\rho_s(a)} - \beta R(s,a) + \mathbb{E}_{s'|s,a} F(s';\beta) + \lambda_s + 1$$

and setting the derivative to zero we have,

$$\begin{aligned} \pi_s(a) &= \frac{\rho_s(a)}{Z(s;\beta)} \exp\left[\beta R(s,a) - \mathbb{E}_{s'|s,a} F(s';\beta)\right] \\ Z(s;\beta) &= \sum_a \rho_s(a) \exp\left[\beta R(s,a) - \mathbb{E}_{s'|s,a} F(s';\beta)\right] \end{aligned}$$

**Algorithm 1** INFO-RL $(\mathcal{S}, \mathcal{A}, P, R, \rho, \beta)$

**initialize** $F(s) \leftarrow 0,\ \forall s \in \mathcal{S}$
**repeat**
  **for** $s = 1$ to $n$ **do**
    $Z(s;\beta) \leftarrow \sum_a \rho_s(a)\, e^{\beta R(s,a) - \mathbb{E}_{s'|s,a} F(s';\beta)}$
    $F(s;\beta) \leftarrow -\log Z(s;\beta)$
  **end for**
**until** $F$ has converged $(F^* \leftarrow F)$
**for each** $a \in \mathcal{A}, s \in \mathcal{S}$
  $\pi^*(a|s) \leftarrow \frac{\rho_s(a)}{Z(s;\beta)} \exp\left[\beta R(s,a) - \mathbb{E}_{s'|s,a} F^*(s';\beta)\right]$
**return** $\pi^*$

where $Z(s;\beta)$ is a partition function. Substituting the solution back in the Lagrangian establishes the Lemma.

Finally, we introduce our INFO-RL algorithm. For a given MDP model and a tradeoff parameter $\beta > 0$, it calculates $F^*$ by iterations of the Bellman equation, and returns the optimal policy $\pi^*$.

Due to the explicit non-linear dependence of the *free-energy* function on $\pi$, the solutions of this optimization problem are *stochastic*. This result of stochastic solutions is similar in nature to the results obtained in other information minimization problems like RDT [2] and the *information-bottleneck method* [9].

## *3.2 The value-information curve*

The tradeoff between value and information can be explored by solving the optimization problem for different values of $\beta > 0$. The solutions form a concave curve in the value-information plane (Fig. 1, left panel). This result is similar to the convexity of the rate-distortion function in RDT [2].

The tradeoff curve is the set of all solutions to the constraint optimization problem. It separates the plane into two regions: above the curve is the non-achievable region, where there is no corresponding policy to satisfy the constraints; below the curve are all sub-optimal solutions that achieve less value with the same level of control information. The rightmost point along the curve ($\beta \to \infty$) represents the maximal *value* any policy can achieve and the minimal level of *control information* required to achieve that value.

Generally, the tradeoff between information and value is far from being linear, allowing agents to find policies that compromise very little expected value while being much cheaper in terms of the *control information*.

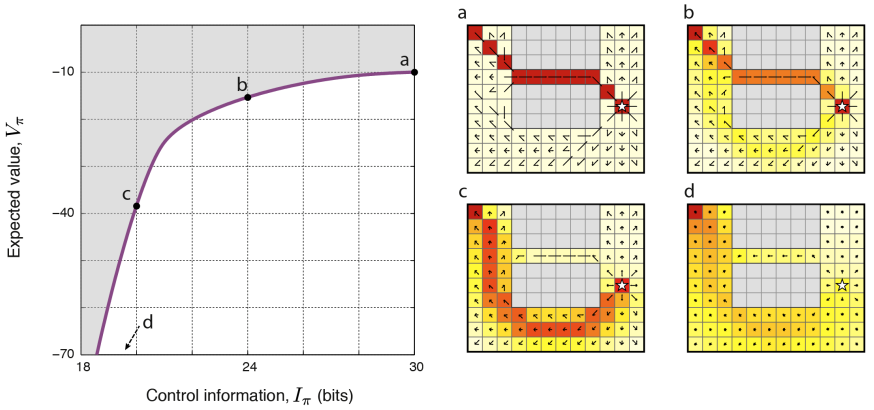


**Fig. 1** Trading value and information in a $10 \times 10$ grid world example. An agent is required to reach the goal (marked by a white star) starting from the top-left corner of the grid, without bumping into the walls. **Left:** The tradeoff curve calculated for this problem (using the INFO-RL algorithm with different tradeoff values). Each point along the curve represents a solution (an optimal policy) to the constrained optimization problem, achieving the maximal expected value $V_\pi$ for the specified level of control information $I_\pi$. **Right:** Four explicit solutions (optimal policies) are shown for control information levels of $I_\pi = 30$, 24, 20 and 2 bits (**a-d**). These policies are depicted by black arrows, with each arrow representing the probability of taking action $a \in \{\uparrow, \nearrow, \rightarrow, ...\}$ at a location $s$ along the grid. Colors represent the probability density of the agent's position along the grid as induced by its policy.

### *3.3 Deterministic environments*

Consider the special case where the states-transition function of the MDP is deterministic. In this particular case, the optimization problem takes a simple *linear* form as shown below. Let $x_{s,a}$ denote the state to which the agent transfers after taking action $a$ in state $s$,

$$P_{s,a}\left(s'\right) = \begin{cases} 1 & s' = x_{s,a} \\ 0 & otherwise \end{cases}$$

**Proposition.** *The update step in the* INFO-RL *algorithm takes the following linear form,*

$$Z_{k+1}\left(s;\beta\right) = \sum_{s' \in \mathcal{S}} \Phi_{s,s'}\left(\beta\right) Z_k\left(s';\beta\right)$$

*where the $n \times n$ matrix $\Phi$ is defined as,*

$$\Phi_{s,s'}\left(\beta\right) = \begin{cases} \sum_a \rho_s\left(a\right) e^{\beta R(s,a)}, & s' = x_{s,a} \\ 0 & otherwise \end{cases}$$

*Proof.* Since $P_{s,a}\left(s'\right) \in \{0,1\}$, we have that $\mathbb{E}_{s'|s,a} F\left(s';\beta\right) = F\left(x_{s,a};\beta\right)$. Substituting $\Phi$ in the update rule of $Z$ establishes the proposition,

$$Z_{k+1}(s;\beta) = \sum_a \rho_s(a)\, e^{\beta R(s,a)}\; e^{-F_{k+1}(x_{s,a};\beta)}$$
$$= \sum_{s'} \Phi_{s,s'}(\beta)\, Z_k(x_{s,a};\beta)$$

The problem of solving a nonlinear set of equations, $F^* = \mathcal{B}F^*$, thus reduces to solving a set of *linear* equations $Z = \Phi Z$, in resemblance to the Z-LEARNING algorithm given in [11]. Furthermore, in many problems the states of the MDP are far from being fully connected resulting in a sparse matrix $\Phi$.

## 4 Grid world example

In order to demonstrate our formalism we proceed with an illustrative example. We explore trading expected *value* and *control information* in a simple grid world problem.

In this setting, states represent the agent's location on a grid with a single state denoting the goal. In each step the agent chooses from eight actions corresponding to eight possible directions $\mathcal{A} = \{\uparrow, \nearrow, \rightarrow, ...\}$. The states-transition function is deterministic (as in Section 3.3): the agent is transferred to an adjacent grid cell according to its action, unless it attempts to move into a wall. The agent receives a negative reward of $R = -1$ on each step in order to favor short paths and a punishment of $R = -100$ for an attempt to move into a wall.

Each deterministic control (such as "go west") adds $\log \frac{\pi_s(a)}{\rho_s(a)} = \log \frac{1}{1/8} = 3$ bits to the total *control information*. As shown in Figure 1 (right panel a), a deterministic policy can lead the agent to the goal in ten steps. This, however, requires $10 \times 3 = 30$ bits of *control information* (see point **a** on the tradeoff curve). What happens when we restrict the control information to lower values?

We calculated the full value-information tradeoff by applying our INFO-RL algorithm for various values of the tradeoff parameter $\beta$. The resulting tradeoff curve is shown in Figure 1 (left). Each point along the curve represents an optimal solution $\pi^*$ of the constraint optimization problem for a specific value of $\beta$. It represents the maximal expected *value* that can be achieved for each and every level of *control complexity*. Four such solutions are presented explicitly. Panel **a** shows the case of $\beta \to \infty$ corresponding to the maximum possible expected value ($V_{max} = -10$). Panels **b**-**d** correspond to solutions obtained for decreasing values of $\beta$. As $\beta \to 0$ the policy becomes uniform over actions (i.e., random walk policy). Executing this random walk policy will eventually reach the goal, but with an expected value of $V < -15,000$ (not shown).

The path through the narrow corridor is indeed the shortest. However, it requires costly deterministic instructions in order not to bump into the walls. Constraining the control complexity to lower values (by decreasing $\beta$) favors the longer and 'safer' path.

## 5 Robustness

In settings with only partial knowledge of the world, the *control information* emerges as a natural regularization term to improve the robustness of the policy to sample fluctuations.

We explore a scenario in which the rewards are drawn from a state-dependent distribution $r \sim P_s(r)$ which is *unknown* to the agent. After a learning phase, in which the agent collects a sample of $m$ realizations of each $r(s)$, it constructs an empirical (and unbiased) estimate $\hat{R}(s) = \frac{1}{m}\sum_i r_i(s)$, of the expected reward function $R(s) = \mathbb{E}_{r|s}[r]$. Thus, the agent doesn't have access to the underlying distribution of the rewards, but only to a *noisy* estimate of it based on its experience. A policy chosen to maximize the value alone may suffer from over-fitting with respect to the *noisy* model, leading to inferior performance in the real world.

We use the *probably approximately correct* (PAC)-Bayesian approach to quantify the ability to learn a good policy from a finite sample. PAC-Bayesian bounds are a generalization of the Occams razor bound for algorithms which output a distribution over classifiers rather than just a single classifier, and are thus suitable for our analysis. We begin by recalling the PAC-Bayesian bound [6].

**Theorem 2.** *Let $x_1,..,x_m \in \mathcal{X}$ be a set of i.i.d samples from a distribution $D$ over $\mathcal{X}$. Also, let $Q$ be a distribution on a set $\mathcal{H}$ and let $l : \mathcal{H} \times \mathcal{X} \mapsto [0,1]$ be a bounded loss function. Under these conditions, it holds with a probability of at least $1-\delta$ over the choice of the sample $x_1,..,x_m$ that for any distribution $P$ over $\mathcal{H}$,*

$$\tilde{\mathrm{D}}_{\mathrm{KL}}\left[l(P,x_1,...,x_m)\,\|\,l(P,D)\right] \;\leq\; \frac{\mathrm{D}_{\mathrm{KL}}[P\|Q] + \log(2m/\delta)}{m-1}$$

*where $l(P,x_1,...,x_m) = \mathbb{E}_{h\sim P}\left[\frac{1}{m}\sum_{i=1}^{m} l(h,x_i)\right]$ is considered an empirical loss and $l(P,D) = \mathbb{E}_{h\sim P, x\sim D}[l(h,x)]$ is considered a generalization loss.*

We use the notation $\tilde{\mathrm{D}}_{\mathrm{KL}}[a\|b]$ for scalars $a,b \in [0,1]$ to denote the Kullback-Leibler (KL) divergence between two Bernoulli distributions with parameters $a$ and $b$.

To utilize the PAC-Bayesian approach in our framework, we make the following assumptions. Let $\mathcal{H} = \{\mathcal{A}\times\mathcal{S}\}^{\infty}$ be the class of possible trajectories. The agent's policy $\pi_s(a)$ and the states-transition probabilities $P_{s,a}(s')$ induce some distribution $P$ over $\mathcal{H}$, for a given initial state $s_0$. Similarly, the default policy $\rho_s(a)$ and $P_{s,a}(s')$ induce another distribution over $\mathcal{H}$, denoted by $Q$. Finally, we note that the KL-divergence between these two distributions is, by construction, our control information term (Eq. 3):

$$\begin{aligned}\mathrm{D}_{\mathrm{KL}}\left[P||Q\right] &= \lim_{T\to\infty}\mathbb{E}\left[\log\frac{\Pr\left(a_0,s_{1,},..,s_T|s_0;\pi\right)}{\Pr\left(a_0,s_1,..,s_T|s_0;\rho\right)}\right] \\ &= \lim_{T\to\infty}\mathbb{E}\left[\sum_{t=0}^{T-1}\log\frac{\pi_{s_t}\left(a_t\right)}{\rho_{s_t}\left(a_t\right)}\right] = I_\pi\left(s\right)\end{aligned}$$

where the expectation is taken with respect to $\Pr\left(a_0,s_{1,},..,s_T|s_0;\pi\right)$.

**Theorem 3.** *Suppose an agent has an a-priori stochastic policy $\rho_s\left(a\right)$. If the agent collects an empirical sample of rewards as described above (with m samples per reward), it holds with a probability of at least $1-\delta$ that for any new proper policy $\pi_s\left(a\right)$ and initial state $s_0$,*

$$\tilde{\mathrm{D}}_{\mathrm{KL}}\left[\hat{V}_\pi\left(s_0\right)\|V_\pi\left(s_0\right)\right] \;\leq\; \frac{I_\pi\left(s_0\right)+\log\left(2m/\delta\right)}{m-1}$$

*where $I_\pi$ is defined as in Eq. (3) and*

$$\begin{aligned}V_\pi\left(s_0\right) &= \lim_{T\to\infty}\mathbb{E}\left[\sum_{t=1}^{T}R\left(s_t\right)\right] \\ \hat{V}_\pi\left(s_0\right) &= \lim_{T\to\infty}\mathbb{E}\left[\sum_{t=1}^{T}\hat{R}\left(s_t\right)\right]\end{aligned}$$

*with the expectations taken with respect to* $\Pr\left(a_0,s_{1,},..,s_T|s_0;\pi\right)$.

The theorem tells us the following. Without any regularization, a policy that maximizes the rewards alone (based on the empirical data) might be very costly in terms of its control information $I_\pi$ (see Section 2.2). As a result, the bound in the theorem will be loose and the true expected reward $V_\pi$ might be much smaller than the empirical reward $\hat{V}_\pi$ which was maximized (i.e., low generalization). On the other hand, if the chosen policy is such that $P$ and $Q$ are very similar then the bound in the theorem will be tight (low *control information*). Nevertheless, both $\hat{V}_\pi$ and $V_\pi$ might still be similarly low. Thus, the theorem implies that in order to find a policy with maximal expected reward $V_\pi$, one should explore the tradeoff between maximizing the reward based on the empirical data (i.e., make $\hat{V}_\pi$ as large as possible), and minimizing the control information $I_\pi$ (i.e., reducing the divergence between $P$ and $Q$). This can be done in practice using the INFO-RL algorithm as demonstrated in the following..

To illustrate this idea, we used a simple $20\times 20$ grid world that the agent needs to cross with a 'mine field' of size $12\times 12$ situated in the middle. Stepping on a mine is punished by $r=-20$ with probability of 50%. In this setting, the preferred solution should be to bypass the 'mine field'. We sample one realization from $r\left(s\right)$ for $s\in\mathcal{S}$ and use it to construct an unbiased estimate of the reward function $\hat{R}\left(s\right)$. Based on this estimate, an optimal policy is calculated using the INFO-RL algorithm for different values of $\beta$. With $\beta\to\infty$, where the focus is on maximizing the value alone, the resulting policy passes *through* the 'mine field' (in between sampled mines). This is

far from being a good solution in the real world, as the entire field is dangerous. As we set $\beta$ to lower values, the information term regularizes the solutions, resulting in better policies (see Fig. 2).

## 6 Discussion

Information processing and control are known to be related, at least since Shannon's 1959 Rate-Distortion theory paper [7]. In this work we establish a direct link between information and control by combining an information theoretic measures within the MDP formulation. We explore the tradeoff between value and information explicitly by solving an optimization problem for optimal planning under information constraints. The suggested INFO-RL algorithm for solving the optimization problem is shown to converge to the *global* fixed-point solution. Furthermore, in the case of deterministic state-transitions, the problem is shown to take a very simple linear form.

Demonstrating the algorithm in a simple grid-world problem we show how *stochastic policies* can dramatically reduce the *control information* while main-

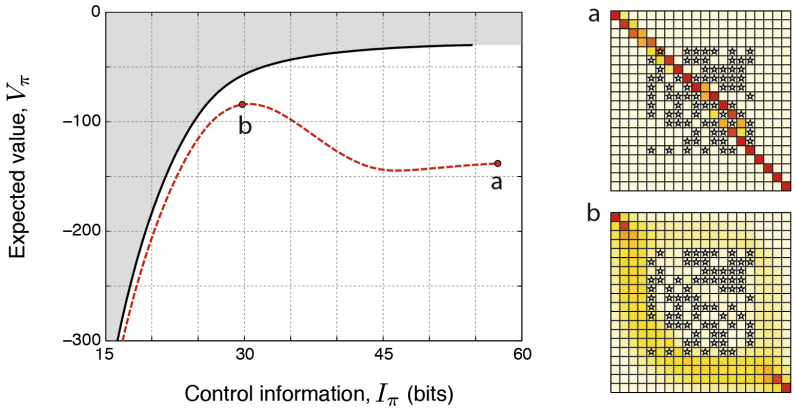


**Fig. 2** Robustness of the INFO-RL policies under a partial knowledge of the MDP parameters. An agent is required to cross a $20 \times 20$ grid world with a dangerous $12 \times 12$ 'mine field' in the center. The locations of mines are unknown to the agent, which can only utilize an empirical estimate of the reward function to plan ahead (see Section 5 for details). **Left:** The tradeoff curve calculated for this problem based on the underlying reward function (unavailable to the agent) is shown as a solid line. The curve is calculated based on the (noisy) empirical estimate of the reward function, shown as a dashed line. It shows the over-training at high control information levels with respect to the expected value under the 'full' model of the MDP. **Right:** Two optimal policies (with respect to the noisy estimate) are shown, with the probability density of the agent's position along the grid in grayscale. The sample of the 'mines' used to build the empirical estimate of the reward is indicated by stars. The solution in **b** is clearly better than the one in **a** in terms of the expected value (i.e., more robust to the noisy estimate of the underlying reward function).

taining close to maximal values. Stochastic policies can also be addressed by using the *softmax* action selection [8]. In contrast with our optimization principle, the *softmax* policy is constructed in two *independent* steps: first a maximal value solution $Q^*(s,a)$ is calculated; only then the *softmax* policy is calculated through $\pi_s(a) \propto e^{Q^*(s,a)/\beta}$, which results in a 'softer' version of the deterministic maximal value solution.

Finally, we use the PAC-Bayesian generalization theorem to show that the solutions are robust to sample fluctuations of the rewards, by providing a better generalization to the training episode sample.

This work is focused on the control term of the circular flow of information between the agent and its environment. Treatment of the complementary term, the information gained from the environment, is subject to ongoing work and will be published separately.

# Appendix

## *Proof of Theorem 1*

Our proof follows [1]. We begin with some preliminary results. We introduce the mapping $T_\pi : \mathbb{R}^n \mapsto \mathbb{R}^n$,

$$[T_\pi F](s) = \sum_{a \in \mathcal{A}} \pi_s(a) \left[ \log \frac{\pi_s(a)}{\rho_s(a)} - \beta R(s,a) + \sum_{s' \in \mathcal{S}} P_{s,a}(s') F(s') \right] \tag{9}$$

and we define the matrix $P_\pi$ (indexed by $s, s' \in \mathcal{S}$) and the vector $g_\pi$ (indexed by $s \in \mathcal{S}$),

$$\begin{aligned} P_\pi(s,s') &= \sum_a P_{s,a}(s') \pi_s(a) \\ g_\pi(s) &= \sum_a \pi_s(a) \left[ \log \frac{\pi_s(a)}{\rho_s(a)} - \beta R(s,a) \right] \end{aligned}$$

to rewrite Eq. (9) in compact vector notation,

$$T_\pi F = g_\pi + P_\pi F \tag{10}$$

and the *free-energy* of a policy $\pi$ as,

$$F_\pi = \lim_{k \to \infty} T_\pi^k F_0 = \lim_{N \to \infty} \sum_{k=0}^{N} P_\pi^k g_\pi \tag{11}$$

where $F_0$ denotes a zeros vector, and high superscripts $k$ indicate raising to the power of $k$.

From the definition of *proper policy* we have that for any *proper policy* $\pi$ and any vector $F$,

$$\lim_{k\to\infty} P_\pi^k F = 0 \tag{12}$$

**Assumption.** *There exists at least one proper policy. Furthermore, for every improper policy $\pi$, the corresponding vector $F_\pi(s)$ is $\infty$ for at least one state $s$.*

In the case that the policy is improper, there is at least one initial state from which the trajectory will never reach the terminal state, and thus we assume that the infinite sum diverges.

**Proposition 4.** *For a proper policy $\pi$, the associated free-energy vector $F_\pi$ satisfies,*

$$\lim_{k\to\infty}\left[T_\pi^k F\right](s) = F_\pi(s),\ s = 1,...,n \tag{13}$$

*for every vector $F$. Furthermore, $F_\pi = T_\pi F_\pi$, and $F_\pi$ is the unique solution of this equation.*

*Proof.* By an induction argument, we have for all $F$,

$$T_\pi^k F = P_\pi^k F + \sum_{t=0}^{k-1} P_\pi^t g_\pi,\ k \geq 1 \tag{14}$$

and using Eq. (12) and Eq. (11) we get,

$$F_\pi = \lim_{k\to\infty} T_\pi^k F = 0 + \lim_{k\to\infty} \sum_{t=0}^{k-1} P_\pi^t g_\pi \tag{15}$$

Also, we have by definition,

$$T_\pi^{k+1} F = g_\pi + P_\pi T_\pi^k F \tag{16}$$

and by taking the limit as $k \to \infty$, we obtain,

$$F_\pi = g_\pi + P_\pi F_\pi = T_\pi F_\pi \tag{17}$$

Finally, to show uniqueness, note that if $F = T_\pi F$, then we have $F = T_\pi^k F$ for all $k$, and so,

$$F = \lim_{k\to\infty} T_\pi^k F = F_\pi \tag{18}$$

**Proposition 5.** *A stationary policy $\pi$ satisfying,*

$$F(s) \geq (T_\pi F)(s),\ s = 1,...,n$$

*for some vector $F$, is proper.*

*Proof.* By Eq. (14) and the proposition's hypothesis, we have that,

$$F \geq T_\pi F \geq T_\pi^k F = P_\pi^k F + \sum_{t=0}^{k-1} P_\pi^t g_\pi \tag{19}$$

If $\pi$ was not *proper*, then by the assumption, some components of the sum in the right-hand side of the above relation will diverge to $\infty$ as $k \to \infty$, which is a contradiction.

Recall the mapping $\mathcal{B} : \mathbb{R}^n \mapsto \mathbb{R}^n$ on $F$ (see Theorem 1 and Eq. 9),

$$[\mathcal{B}F](s) = \min_{\pi_s(\cdot)} [T_\pi F](s)$$

The following proposition establishes the uniqueness of the solution.

**Proposition 6.** *The equation, $F = \mathcal{B}F$, has at most one fixed point solution.*

*Proof.* If $F$ and $F'$ are two fixed points, then we select $\pi$ and $\pi'$ such that,

$$\begin{aligned} \pi_s(\cdot) &= \arg\min_{\mu_s(\cdot)} \left[T_\mu F\right](s) \\ \pi'_s(\cdot) &= \arg\min_{\mu_s(\cdot)} \left[T_\mu F'\right](s) \end{aligned} \tag{20}$$

By this construction, $T_\pi F = F$ and $T_{\pi'} F' = F'$. By Proposition 5 we have that both $\pi$ and $\pi'$ are proper, and by Proposition 4 we have that $F = F_\pi$ and $F' = F_{\pi'}$. Also, we have,

$$F = \mathcal{B}F = \mathcal{B}^k F \leq T_{\pi'}^k F \ , \ k \geq 1$$

taking $k \to \infty$ and using Prop, 4, we obtain,

$$F \leq \lim_{k\to\infty} T_{\pi'}^k F = F_{\pi'} = F' \tag{21}$$

Similarly, $F' \leq F$, showing that $F = F'$ and that $F = \mathcal{B}F$ has at most one fixed point.

Lastly, we show that the optimal *free-energy* vector $F^* = \min_\pi F_\pi$ satisfies Bellman's equation, assuming that a proper optimal policy $\pi^*$ indeed exists.

**Proposition 7.** *The optimal free-energy vector $F^*$ satisfies $F^* = \mathcal{B}F^*$.*

*Proof.* Let $\pi^*$ denote the optimal proper policy,

$$\pi^* = \arg\min_\pi F_\pi$$

consequently, for any policy $\pi$ we have, $F^* \leq F_\pi$. Applying the mapping $\mathcal{B}$ on $F^*$ we have,

$$\mathcal{B}F^* = \min_\pi T_\pi F^* \leq T_{\pi^*} F^* = F^* \tag{22}$$

Next, we select a policy $\mu$ such that $T_\mu F^* = \mathcal{B}F^*$, and using Eq. (22) we have that $F^* \geq \mathcal{B}F^* = T_\mu F^*$. Thus for any $k \geq 1$ we have $F^* \geq T_\mu^k F^*$, and taking $k \to \infty$ we have by Proposition 4,

$$F^* \geq \lim_{k\to\infty} T_\mu^k F^* = F_\mu$$

and because $F^* \leq F_\pi$ for any policy $\pi$, we have that $F^* = F_\mu$. Lastly, by the construction of $\mu$ we have,

$$F^* = F_\mu = T_\mu F_\mu = T_\mu F^* = \mathcal{B}F^*$$

Finally, Proposition 6 and 7 establishes Theorem 1.

## References


1. Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 1995.
2. Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.
3. Karl Friston. The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.*, 13(7):293–301, June 2009.
4. Joaqn M. Fuster. The prefrontal cortex — an update: Time is of the essence. *Neuron*, 30:319–333, May 2001.
5. B. Kappen, V. Gomez, and M. Opper. Optimal control as a graphical model inference problem. *ArXiv e-prints*, January 2009.
6. David Mcallester. Simplified pac-bayesian margin bounds. In *Proc. of the 16th Annual Conference on Learning Theory*, 2003.
7. C.E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE NATO Conv. Rec.*, 4:142– 163, March 1959.
8. R. S. Sutton and A. G. Barto. *Reinforcement Learning*. MIT Press, Cambridge, Mass., 1998.
9. N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. 37th Annual Allerton Conference on Communication, Control and Computing*, 1999.
10. N. Tishby and D. Polani. Information theory of decisions and actions. In Vassilis, Hussain, and Taylor, editors, *Perception-Reason-Action*, Cognitive Neuroscience. Springer, 2010.
11. Emanuel Todorov. Efficient computation of optimal actions. *PNAS*, 106(28):11478–11483, 2009.