
A Unified Bellman Optimality Principle Combining Reward Maximization and Empowerment

Felix Leibfried, Sergio Pascual-Díaz, Jordi Grau-Moya
PROWLER.io
Cambridge, UK
{felix,sergio.diaz,jordi}@proowler.io

Abstract

Empowerment is an information-theoretic method that can be used to intrinsically motivate learning agents. It attempts to maximize an agent’s control over the environment by encouraging visiting states with a large number of reachable next states. Empowered learning has been shown to lead to complex behaviors, without requiring an explicit reward signal. In this paper, we investigate the use of empowerment in the presence of an extrinsic reward signal. We hypothesize that empowerment can guide reinforcement learning (RL) agents to find good early behavioral solutions by encouraging highly empowered states. We propose a unified Bellman optimality principle for empowered reward maximization. Our empowered reward maximization approach generalizes both Bellman’s optimality principle as well as recent information-theoretical extensions to it. We prove uniqueness of the empowered values and show convergence to the optimal solution. We then apply this idea to develop off-policy actor-critic RL algorithms which we validate in high-dimensional continuous robotics domains (MuJoCo). Our methods demonstrate improved initial and competitive final performance compared to model-free state-of-the-art techniques.

1 Introduction

In reinforcement learning [62] (RL), agents identify policies to collect as much reward as possible in a given environment. Recently, leveraging parametric function approximators has led to tremendous success in applying RL to high-dimensional domains such as Atari games [40] or robotics [56]. In such domains, inspired by the policy gradient theorem [63, 13], actor-critic approaches [36, 41] attain state-of-the-art results by learning both a parametric policy and a value function.

Empowerment is an information-theoretic framework where agents maximize the mutual information between an action sequence and the state that is obtained after executing this action sequence from some given initial state [26, 27, 53]. It turns out that the mutual information is highest for such initial states where the number of reachable next states is largest. Policies that aim for high empowerment can lead to complex behavior, e.g. balancing a pole in the absence of any explicit reward signal [23].

Despite progress on learning empowerment values with function approximators [42, 12, 49], there has been little attempt in the combination with reward maximization, let alone in utilizing empowerment for RL in the high-dimensional domains it has become applicable just recently. We therefore propose a unified principle for reward maximization and empowerment, and demonstrate that empowered signals can boost RL in large-scale domains such as robotics. In short, our contributions are:

- a generalized Bellman optimality principle for joint reward maximization and empowerment,
- a proof for unique values and convergence to the optimal solution for our novel principle,
- empowered actor-critic methods boosting RL in MuJoCo compared to model-free baselines.

2 Background

2.1 Reinforcement Learning

In the discrete RL setting, an agent, being in state $s \in \mathcal{S}$, executes an action $\mathbf{a} \in \mathcal{A}$ according to a behavioral policy $\pi_{\text{behave}}(\mathbf{a}|s)$ that is a conditional probability distribution $\pi_{\text{behave}} : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$. The environment, in response, transitions to a successor state $s' \in \mathcal{S}$ according to a (probabilistic) state-transition function $\mathcal{P}(s'|s, \mathbf{a})$, where $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$. Furthermore, the environment generates a reward signal $r = \mathcal{R}(s, \mathbf{a})$ according to a reward function $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. The agent's aim is to maximize its expected future cumulative reward with respect to the behavioral policy $\max_{\pi_{\text{behave}}} \mathbb{E}_{\pi_{\text{behave}}, \mathcal{P}} [\sum_{t=0}^{\infty} \gamma^t r_t]$, with t being a time index and $\gamma \in (0, 1)$ a discount factor. Optimal expected future cumulative reward values for a given state s obey then the following recursion:

$$V^*(s) = \max_{\mathbf{a}} (\mathcal{R}(s, \mathbf{a}) + \gamma \mathbb{E}_{\mathcal{P}(s'|s, \mathbf{a})} [V^*(s')]) =: \max_{\mathbf{a}} Q^*(s, \mathbf{a}), \quad (1)$$

referred to as Bellman's optimality principle [4], where V^* and Q^* are the optimal value functions.

2.2 Empowerment

Empowerment is an information-theoretic method where an agent executes a sequence of k actions $\vec{\mathbf{a}} \in \mathcal{A}^k$ when in state $s \in \mathcal{S}$ according to a policy $\pi_{\text{empower}}(\vec{\mathbf{a}}|s)$ which is a conditional probability distribution $\pi_{\text{empower}} : \mathcal{S} \times \mathcal{A}^k \rightarrow [0, 1]$. This is slightly more general than in the RL setting where only a single action is taken upon observing a certain state. The agent's aim is to identify an optimal policy π_{empower} that maximizes the mutual information $I[\vec{\mathbf{A}}, S'|s]$ between the action sequence $\vec{\mathbf{a}}$ and the state s' to which the environment transitions after executing $\vec{\mathbf{a}}$ in s , formulated as:

$$E^*(s) = \max_{\pi_{\text{empower}}} I[\vec{\mathbf{A}}, S'|s] = \max_{\pi_{\text{empower}}} \mathbb{E}_{\pi_{\text{empower}}(\vec{\mathbf{a}}|s) \mathcal{P}^{(k)}(s'|s, \vec{\mathbf{a}})} \left[\log \frac{p(\vec{\mathbf{a}}|s', s)}{\pi_{\text{empower}}(\vec{\mathbf{a}}|s)} \right]. \quad (2)$$

Here, $E^*(s)$ refers to the optimal empowerment value and $\mathcal{P}^{(k)}(s'|s, \vec{\mathbf{a}})$ to the probability of transitioning to s' after executing the sequence $\vec{\mathbf{a}}$ in state s , where $\mathcal{P}^{(k)} : \mathcal{S} \times \mathcal{A}^k \times \mathcal{S} \rightarrow [0, 1]$.

Importantly, $p(\vec{\mathbf{a}}|s', s) = \frac{\mathcal{P}^{(k)}(s'|s, \vec{\mathbf{a}}) \pi_{\text{empower}}(\vec{\mathbf{a}}|s)}{\sum_{\vec{\mathbf{a}}} \mathcal{P}^{(k)}(s'|s, \vec{\mathbf{a}}) \pi_{\text{empower}}(\vec{\mathbf{a}}|s)}$ is the inverse dynamics model of π_{empower} . The implicit dependency of p on the optimization argument π_{empower} renders the problem non-trivial.

From an information-theoretic perspective, optimizing for empowerment is equivalent to maximizing the capacity [58] of an information channel $\mathcal{P}^{(k)}(s'|s, \vec{\mathbf{a}})$ with input $\vec{\mathbf{a}}$ and output s' w.r.t. the input distribution $\pi_{\text{empower}}(\vec{\mathbf{a}}|s)$, as outlined in the following [11, 10]. Define the functional $I_f(\pi_{\text{empower}}, \mathcal{P}^{(k)}, q) := \mathbb{E}_{\pi_{\text{empower}}(\vec{\mathbf{a}}|s) \mathcal{P}^{(k)}(s'|s, \vec{\mathbf{a}})} \left[\log \frac{q(\vec{\mathbf{a}}|s', s)}{\pi_{\text{empower}}(\vec{\mathbf{a}}|s)} \right]$, where q is a conditional probability $q : \mathcal{S} \times \mathcal{S} \times \mathcal{A}^k \rightarrow [0, 1]$. Then the mutual information is recovered as a special case of I_f with $I[\vec{\mathbf{A}}, S'|s] = \max_q I_f(\pi_{\text{empower}}, \mathcal{P}^{(k)}, q)$ for a given π_{empower} . The maximum argument

$$q^*(\vec{\mathbf{a}}|s', s) = \frac{\mathcal{P}^{(k)}(s'|s, \vec{\mathbf{a}}) \pi_{\text{empower}}(\vec{\mathbf{a}}|s)}{\sum_{\vec{\mathbf{a}}} \mathcal{P}^{(k)}(s'|s, \vec{\mathbf{a}}) \pi_{\text{empower}}(\vec{\mathbf{a}}|s)} \quad (3)$$

is the true Bayesian posterior $p(\vec{\mathbf{a}}|s', s)$ —see [10] Lemma 10.8.1 for details. Similarly, maximizing $I_f(\pi_{\text{empower}}, \mathcal{P}^{(k)}, q)$ with respect to π_{empower} for a given q leads to:

$$\pi_{\text{empower}}^*(\vec{\mathbf{a}}|s) = \frac{\exp(\mathbb{E}_{\mathcal{P}^{(k)}(s'|s, \vec{\mathbf{a}})} [\log q(\vec{\mathbf{a}}|s', s)])}{\sum_{\vec{\mathbf{a}}} \exp(\mathbb{E}_{\mathcal{P}^{(k)}(s'|s, \vec{\mathbf{a}})} [\log q(\vec{\mathbf{a}}|s', s)])}. \quad (4)$$

As explained e.g. in [10] page 335 similar to [46]. The above yields the subsequent proposition.

Proposition 1 *Maximum Channel Capacity. Iterating through Equations (3) and (4) by computing q given π_{empower} and vice versa in an alternating fashion converges to an optimal pair $(q^*, \pi_{\text{empower}}^*)$ that maximizes the mutual information $\max_{\pi_{\text{empower}}} I[\vec{\mathbf{A}}, S'|s] = I_f(\pi_{\text{empower}}^*, \mathcal{P}^{(k)}, q^*)$. The convergence rate is $\mathcal{O}(1/N)$, where N is the number of iterations, for any initial $\pi_{\text{empower}}^{\text{ini}}$ with support in $\mathcal{A}^k \forall s$ —see [10] Chapter 10.8 and [11, 16]. This is known as Blahut-Arimoto algorithm [2, 7].*

Remark. Empowerment is similar to curiosity concepts of predictive information that focus on the mutual information between the current and the subsequent state [6, 48, 69, 61, 43, 54].

3 Motivation: Combining Reward Maximization with Empowerment

The Blahut-Arimoto algorithm presented in the previous section solves empowerment for low-dimensional discrete settings but does not readily scale to high-dimensional or continuous state-action spaces. While there has been progress on learning empowerment values with parametric function approximators [42], how to combine it with reward maximization or RL remains open. In principle, there are two possibilities for utilizing empowerment. The first is to directly use the policy π_{empower}^* obtained in the course of learning empowerment values $E^*(\mathbf{s})$. The second is to train a behavioral policy to take an action in each state such that the expected empowerment value of the next state is highest (requiring E^* -values as a prerequisite). Note that the two possibilities are conceptually different. The latter seeks states with a large number of reachable next states [23]. The first, on the other hand, aims for high mutual information between actions and the subsequent state, which is not necessarily the same as seeking highly empowered states [42].

We hypothesize empowered signals to be beneficial for RL, especially in high-dimensional environments and at the beginning of the training process when the initial policy is poor. In this work, we therefore combine reward maximization with empowerment inspired by the two behavioral possibilities outlined in the previous paragraph. Hence, we focus on the cumulative RL setting rather than the non-cumulative setting that is typical for empowerment. We furthermore use one-step empowerment as a reference, i.e. $k = 1$, because cumulative one-step empowerment learning leads to high values in such states where the number of possibly reachable next states is high, and preserves hence the original empowerment intuition *without* requiring a multi-step policy—see Section 4.3. The first idea is to train a policy that trades off reward maximization and *learning* cumulative empowerment:

$$\max_{\pi_{\text{behave}}} \mathbb{E}_{\pi_{\text{behave}}, \mathcal{P}} \left[\sum_{t=0}^{\infty} \gamma^t \left(\alpha \mathcal{R}(\mathbf{s}_t, \mathbf{a}_t) + \beta \log \frac{p(\mathbf{a}_t | \mathbf{s}_{t+1}, \mathbf{s}_t)}{\pi_{\text{behave}}(\mathbf{a}_t | \mathbf{s}_t)} \right) \right], \quad (5)$$

where $\alpha \geq 0$ and $\beta \geq 0$ are scaling factors, and p indicates the inverse dynamics model of π_{behave} in line with Equation (3). Note that p depends on the optimization argument π_{behave} , similar to ordinary empowerment, leading to a non-trivial Markov decision problem (MDP).

The second idea is to learn cumulative empowerment values *a priori* by solving Equation (5) with $\alpha = 0$ and $\beta = 1$. The outcome of this is a policy π_{empower}^* (and its inverse dynamics model p) that can be used to construct an intrinsic reward signal which is then added to the external reward:

$$\max_{\pi_{\text{behave}}} \mathbb{E}_{\pi_{\text{behave}}, \mathcal{P}} \left[\sum_{t=0}^{\infty} \gamma^t \left(\alpha \mathcal{R}(\mathbf{s}_t, \mathbf{a}_t) + \beta \mathbb{E}_{\pi_{\text{empower}}^*(\mathbf{a} | \mathbf{s}_t), \mathcal{P}(\mathbf{s}' | \mathbf{s}_t, \mathbf{a})} \left[\log \frac{p(\mathbf{a} | \mathbf{s}', \mathbf{s}_t)}{\pi_{\text{empower}}^*(\mathbf{a} | \mathbf{s}_t)} \right] \right) \right]. \quad (6)$$

Importantly, Equation (6) poses an ordinary MDP since the reward signal is merely extended by another stationary state-dependent signal.

Both proposed ideas require to solve the novel MDP as specified in Equation (5). In Section 4, we therefore prove the existence of unique values and convergence of the corresponding value iteration scheme (including a grid world example). We also show how our formulation generalizes existing formulations from the literature. In Section 5, we carry our ideas over to high-dimensional continuous state-action spaces by devising off-policy actor-critic-style algorithms inspired by the proposed MDP formulation. We evaluate our novel actor-critic-style algorithms in MuJoCo demonstrating better initial and competitive final performance compared to model-free state-of-the-art baselines.

4 Joint Reward Maximization and Empowerment Learning in MDPs

We state our main theoretical result *in advance*, proven in the remainder of this section (an intuition follows): the solution to the MDP from Equation (5) implies unique optimal values V^* obeying the Bellman recursion

$$\begin{aligned} V^*(\mathbf{s}) &= \max_{\pi_{\text{behave}}} \mathbb{E}_{\pi_{\text{behave}}, \mathcal{P}} \left[\sum_{t=0}^{\infty} \gamma^t \left(\alpha \mathcal{R}(\mathbf{s}_t, \mathbf{a}_t) + \beta \log \frac{p(\mathbf{a}_t | \mathbf{s}_{t+1}, \mathbf{s}_t)}{\pi_{\text{behave}}(\mathbf{a}_t | \mathbf{s}_t)} \right) \middle| \mathbf{s}_0 = \mathbf{s} \right] \\ &= \max_{\pi_{\text{behave}}, q} \mathbb{E}_{\pi_{\text{behave}}(\mathbf{a} | \mathbf{s})} \left[\alpha \mathcal{R}(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\mathcal{P}(\mathbf{s}' | \mathbf{s}, \mathbf{a})} \left[\beta \log \frac{q(\mathbf{a} | \mathbf{s}', \mathbf{s})}{\pi_{\text{behave}}(\mathbf{a} | \mathbf{s})} + \gamma V^*(\mathbf{s}') \right] \right] \\ &= \beta \log \sum_{\mathbf{a}} \exp \left(\frac{\alpha}{\beta} \mathcal{R}(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\mathcal{P}(\mathbf{s}' | \mathbf{s}, \mathbf{a})} \left[\log q^*(\mathbf{a} | \mathbf{s}', \mathbf{s}) + \frac{\gamma}{\beta} V^*(\mathbf{s}') \right] \right), \end{aligned} \quad (7)$$

where

$$q^*(\mathbf{a}|\mathbf{s}', \mathbf{s}) = \frac{\mathcal{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a})\pi_{\text{behave}}^*(\mathbf{a}|\mathbf{s})}{\sum_{\mathbf{a}} \mathcal{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a})\pi_{\text{behave}}^*(\mathbf{a}|\mathbf{s})} = p(\mathbf{a}|\mathbf{s}', \mathbf{s}) \quad (8)$$

is the inverse dynamics model of the optimal behavioral policy π_{behave}^* that assumes the form:

$$\pi_{\text{behave}}^*(\mathbf{a}|\mathbf{s}) = \frac{\exp\left(\frac{\alpha}{\beta}\mathcal{R}(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\mathcal{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a})}\left[\log q^*(\mathbf{a}|\mathbf{s}', \mathbf{s}) + \frac{\gamma}{\beta}V^*(\mathbf{s}')\right]\right)}{\sum_{\mathbf{a}} \exp\left(\frac{\alpha}{\beta}\mathcal{R}(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\mathcal{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a})}\left[\log q^*(\mathbf{a}|\mathbf{s}', \mathbf{s}) + \frac{\gamma}{\beta}V^*(\mathbf{s}')\right]\right)}, \quad (9)$$

where the denominator is just $\exp((1/\beta)V^*(\mathbf{s}))$. While the remainder of this section explains how Equations (7) to (9) are derived in detail, it can be insightful to understand at a high level what makes our formulation non-trivial. The difficulty is that the inverse dynamics model $p = q^*$ depends on the optimal policy $\pi_{\text{behavioral}}^*$ and vice versa leading to a non-standard optimal value identification problem. Proving the existence of V^* -values and how to compute them poses therefore our main theoretical contribution, and implies the existence of at least one $(q^*, \pi_{\text{behave}}^*)$ -pair that satisfies the recursive relationship of Equations (8) and (9). This proof is given in Section 4.1 and leads naturally to a value iteration scheme to compute optimal values in practice. The convergence of this scheme is proven in Section 4.2 and we also demonstrate value learning in a grid world example—see Section 4.3. In Section 4.4, we elucidate how our formulation generalizes and relates to existing MDP formulations.

4.1 Existence of Unique Optimal Values

Following the second line from Equation (7), let's define the Bellman operator $B_* : \mathbb{R}^{|\mathcal{S}|} \rightarrow \mathbb{R}^{|\mathcal{S}|}$ as

$$B_*V(\mathbf{s}) := \max_{\pi_{\text{behave}}, q} \mathbb{E}_{\pi_{\text{behave}}(\mathbf{a}|\mathbf{s})} \left[\alpha\mathcal{R}(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\mathcal{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a})} \left[\beta \log \frac{q(\mathbf{a}|\mathbf{s}', \mathbf{s})}{\pi_{\text{behave}}(\mathbf{a}|\mathbf{s})} + \gamma V(\mathbf{s}') \right] \right]. \quad (10)$$

Theorem 1 *Existence of Unique Optimal Values.* Assuming a bounded reward function \mathcal{R} , the optimal value vector V^* as given in Equation (7) exists and is a unique fixed point $V^* = B_*V^*$ of the Bellman operator B_* from Equation (10).

Proof. The proof of Theorem 1 comprises three steps. First, we prove for a given (q, π_{behave}) -pair the existence of unique values $V^{(q, \pi_{\text{behave}})}$ which obey the following recursion

$$V^{(q, \pi_{\text{behave}})}(\mathbf{s}) = \mathbb{E}_{\pi_{\text{behave}}(\mathbf{a}|\mathbf{s})} \left[\alpha\mathcal{R}(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\mathcal{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a})} \left[\beta \log \frac{q(\mathbf{a}|\mathbf{s}', \mathbf{s})}{\pi_{\text{behave}}(\mathbf{a}|\mathbf{s})} + \gamma V^{(q, \pi_{\text{behave}})}(\mathbf{s}') \right] \right]. \quad (11)$$

This result is obtained through Proposition 2 following [5, 51, 18] where we show that the value vector $V^{(q, \pi_{\text{behave}})}$ is a unique fixed point of the operator $B_{q, \pi_{\text{behave}}} : \mathbb{R}^{|\mathcal{S}|} \rightarrow \mathbb{R}^{|\mathcal{S}|}$ given by

$$B_{q, \pi_{\text{behave}}}V(\mathbf{s}) := \mathbb{E}_{\pi_{\text{behave}}(\mathbf{a}|\mathbf{s})} \left[\alpha\mathcal{R}(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\mathcal{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a})} \left[\beta \log \frac{q(\mathbf{a}|\mathbf{s}', \mathbf{s})}{\pi_{\text{behave}}(\mathbf{a}|\mathbf{s})} + \gamma V(\mathbf{s}') \right] \right]. \quad (12)$$

Second, we prove in Proposition 3 that solving the right hand side of Equation (10) for the pair (q, π_{behave}) can be achieved with a Blahut-Arimoto-style algorithm in line with [16]. Third, we complete the proof in Proposition 4 based on Proposition 2 and 3 by showing that $V^* = \max_{\pi_{\text{behave}}, q} V^{(q, \pi_{\text{behave}})}$, where the vector-valued max-operator is well-defined because both π_{behave} and q are conditioned on \mathbf{s} . The proof completion follows again [5, 51, 18]. \square

Proposition 2 *Existence of Unique Values for a Given (q, π_{behave}) -Pair.* Assuming a bounded reward function \mathcal{R} , the value vector $V^{(q, \pi_{\text{behave}})}$ as given in Equation (11) exists and is a unique fixed point $V^{(q, \pi_{\text{behave}})} = B_{q, \pi_{\text{behave}}}V^{(q, \pi_{\text{behave}})}$ of the Bellman operator $B_{q, \pi_{\text{behave}}}$ from Equation (12).

As opposed to the Bellman operator B_* , the operator $B_{q, \pi_{\text{behave}}}$ does *not* include a max-operation that incurs a non-trivial recursive relationship between optimal arguments. The proof for existence of unique values follows hence standard methodology [5, 51, 18] and is given in Appendix A.1.

Proposition 3 *Blahut-Arimoto for One Value Iteration Step.* Assuming that \mathcal{R} is bounded, the maximization problem $\max_{\pi_{\text{behave}}, q}$ from Equation (10) in the Bellman operator B_* can be solved for (q, π_{behave}) by iterating through the following two equations in an alternating fashion:

$$q^{(m)}(\mathbf{a}|\mathbf{s}', \mathbf{s}) = \frac{\mathcal{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a})\pi_{\text{behave}}^{(m)}(\mathbf{a}|\mathbf{s})}{\sum_{\mathbf{a}} \mathcal{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a})\pi_{\text{behave}}^{(m)}(\mathbf{a}|\mathbf{s})}, \quad (13)$$

$$\pi_{\text{behave}}^{(m+1)}(\mathbf{a}|\mathbf{s}) = \frac{\exp\left(\frac{\alpha}{\beta}\mathcal{R}(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\mathcal{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a})}\left[\log q^{(m)}(\mathbf{a}|\mathbf{s}', \mathbf{s}) + \frac{\gamma}{\beta}V(\mathbf{s}')\right]\right)}{\sum_{\mathbf{a}} \exp\left(\frac{\alpha}{\beta}\mathcal{R}(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\mathcal{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a})}\left[\log q^{(m)}(\mathbf{a}|\mathbf{s}', \mathbf{s}) + \frac{\gamma}{\beta}V(\mathbf{s}')\right]\right)}, \quad (14)$$

where m is the iteration index. The convergence rate is $\mathcal{O}(1/M)$ for arbitrary initial $\pi_{\text{behave}}^{(0)}$ with support in $\mathcal{A} \forall \mathbf{s}$. M is the total number of iterations. The complexity for a single \mathbf{s} is $\mathcal{O}(M|\mathcal{S}||\mathcal{A}|)$.

Proof Outline. The problem in Proposition 3 is mathematically similar to the maximum channel capacity problem [58] from Proposition 1 and proving convergence follows similar steps that we outline here—details can be found in Appendix A.2. First, we prove that optimizing the right-hand side of Equation (10) w.r.t. q for a given π_{behave} results in Equation (13) according to [10] Lemma 10.8.1. Second, we prove that optimizing w.r.t. π_{behave} for a given q results in Equation (14) following standard techniques from variational calculus and Lagrange multipliers. Third, we prove convergence to a global maximum when iterating alternately through Equations (13) and (14) following [16].

Proposition 4 *Completing the Proof of Theorem 1. The optimal value vector is given by $V^* = \max_{\pi_{\text{behave}}, q} V^{(q, \pi_{\text{behave}})}$ and is a unique fixed point $V^* = B_* V^*$ of the Bellman operator B_* .*

Completing the proof of Theorem 1 requires two ingredients: the existence of unique $V^{(q, \pi_{\text{behave}})}$ -values for any (q, π_{behave}) -pair as proven in Proposition 2, and the fact that the optimal Bellman operator can be expressed as $B_* = \max_{\pi_{\text{behave}}, q} B_{q, \pi_{\text{behave}}}$ where $\max_{\pi_{\text{behave}}, q}$ is the max-operator from Proposition 3. The proof follows then standard methodology [5, 51, 18], see Appendix A.3.

4.2 Value Iteration and Convergence to Optimal Values

In the previous section, we have proven the existence of unique optimal values V^* that are a fixed point of the Bellman operator B_* . This section devises a value iteration scheme based on the operator B_* and proves its convergence. We commence by a corollary to express B_* more concisely.

Corollary 1 *Optimal Bellman Operator. The operator B_* from Equation (10) can be written as*

$$B_* V(\mathbf{s}) = \beta \log \sum_{\mathbf{a}} \exp\left(\frac{\alpha}{\beta}\mathcal{R}(\mathbf{s}, \mathbf{a}) + \mathbb{E}_{\mathcal{P}(\mathbf{s}'|\mathbf{s}, \mathbf{a})}\left[\log q^{\text{converged}}(\mathbf{a}|\mathbf{s}', \mathbf{s}) + \frac{\gamma}{\beta}V(\mathbf{s}')\right]\right), \quad (15)$$

where $q^{\text{converged}}(\mathbf{a}|\mathbf{s}', \mathbf{s})$ is the result of the converged Blahut-Arimoto scheme from Proposition 3.

This result is obtained by plugging the converged solution $\pi_{\text{behave}}^{\text{converged}}$ from Equation (14) into Equation (10) and leads naturally to a two-level value iteration algorithm that proceeds as follows: the outer loop updates the values V by applying Equation (15) repeatedly; the inner loop applies the Blahut-Arimoto algorithm from Proposition 3 to identify $q^{\text{converged}}$ required for the outer value update.

Theorem 2 *Convergence to Optimal Values. Assuming bounded \mathcal{R} and let $\epsilon \in \mathbb{R}$ be a positive number such that $\epsilon < \frac{\eta}{1-\gamma}$ where $\eta = \alpha \max_{\mathbf{s}, \mathbf{a}} |\mathcal{R}(\mathbf{s}, \mathbf{a})| + \beta \log |\mathcal{A}|$. If the value iteration scheme with initial values of $V(\mathbf{s}) = 0 \forall \mathbf{s}$ is run for $i \geq \left\lceil \log_{\gamma} \frac{\epsilon(1-\gamma)}{\eta} \right\rceil$ iterations, then $\|V^* - B_*^{(i)} V\|_{\infty} \leq \epsilon$, where the notation $B_*^{(i)} V$ means to apply B_* to V i -times consecutively.*

Proof. Via a sequence of inequalities, one can show that the following holds true: $\|V^* - B_*^{(i)} V\|_{\infty} \leq \gamma \|V^* - B_*^{(i-1)} V\|_{\infty} \leq \gamma^i \|V^* - V\|_{\infty} \leq \gamma^i \frac{1}{1-\gamma} \eta$ —see Appendix A.4 for a more detailed derivation. This implies that if $\epsilon \geq \gamma^i \frac{1}{1-\gamma} \eta$ then $i \geq \left\lceil \log_{\gamma} \frac{\epsilon(1-\gamma)}{\eta} \right\rceil$ presupposing $\epsilon < \frac{\eta}{1-\gamma}$. \square

Conclusion. Together, Theorems 1 and 2 prove that our proposed value iteration scheme converges to optimal values V^* in combination with a corresponding optimal pair $(q^*, \pi_{\text{behave}}^*)$ as described at the beginning of this section in the third line of Equation (7) and in Equations (8) and (9) respectively. The overall complexity is $\mathcal{O}(iM|\mathcal{S}|^2|\mathcal{A}|)$ where i and M refer to outer and inner iterations.

Remark. Our value iteration is required for both objectives from Section 3 to combine reward maximization with empowerment. Equation (5) motivated our scheme in the first place, whereas Equation (6) requires cumulative empowerment values without reward maximization ($\alpha = 0, \beta = 1$).

4.3 Practical Verification in a Grid World Example

In order to practically verify our value iteration scheme from the previous section, we conduct experiments on a grid world example. The outcome is shown in Figure 1 demonstrating how different configurations for α and β , that steer cumulative reward maximization versus empowerment learning, affect optimal values V^* . Importantly, the experiments show that our proposal to learn cumulative one-step empowerment values recovers the original intuition of empowerment in the sense that high values are assigned to states where many other states can be reached and low values to states where the number of reachable next states is low, *but without* the necessity to maintain a multi-step policy.

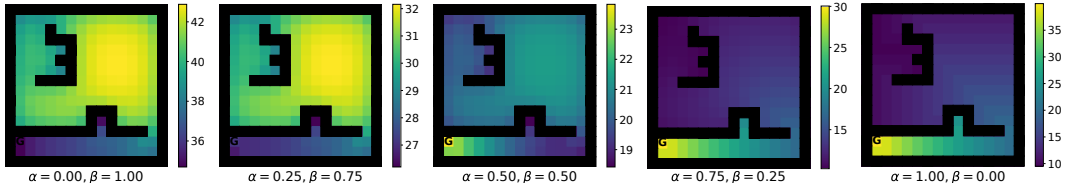


Figure 1: Value Iteration for a Grid World Example. The agent aims to arrive at the goal 'G' in the lower left—detailed information regarding the setup can be found in Appendix C.1. The plots show optimal values for different α and β : α increases from left to right while β decreases. The leftmost values show raw cumulative empowerment learning ($\alpha = 0.0$, $\beta = 1.0$). High values are assigned to states where many other states can be reached, i.e. the upper right; and low values to states where the number of reachable next states is low, i.e. close to corners and dead ends. The rightmost values recover ordinary cumulative reward maximization ($\alpha = 1.0$, $\beta = 0.0$) assigning high values to states close to the goal and low values to states far away from the goal.

4.4 Generalization of and Relation to Existing MDP formulations

Our Bellman operator B_* from Equation (10) relates to prior work as follows (see also Appendix A.5).

- Ordinary value iteration [52] is recovered as a special case for $\alpha = 1$ and $\beta = 0$.
- Cumulative one-step empowerment is recovered as a special case for $\alpha = 0$ and $\beta = 1$, with *non-cumulative* one-step empowerment [29] as a further special case of the latter ($\gamma \rightarrow 0$).
- When setting $q(\mathbf{a}|s', s) = q(\mathbf{a}|s)$, using a distribution that is *not* conditioned on s' and *omitting* maximizing w.r.t. q , one recovers as a special case the soft Bellman operator presented e.g. in [51]. Note that this soft Bellman operator also occurred in numerous other work on MDP formulations and RL [3, 14, 45, 55, 33].
- As a special case of the previous, when $q(\mathbf{a}|s', s) = \mathcal{U}(\mathcal{A})$ is the uniform distribution in action space, one recovers cumulative entropy regularization [70, 44, 34] that inspired algorithms such as soft Q-learning [20] and soft actor-critic [21, 22].
- When dropping the conditioning on s' and s by setting $q(\mathbf{a}|s', s) = q(\mathbf{a})$ but *without omitting* maximization w.r.t. q , one recovers a formulation similar to [65] based on mutual-information regularization [59, 60, 17, 31] that spurred RL algorithms such as [30, 19, 32].
- When replacing $q(\mathbf{a}|s', s)$ with $q(\mathbf{a}|s_t, \mathbf{a}_t)$, where s_t and \mathbf{a}_t refers to the state-action pair of the previous time step, one recovers a formulation similar to [64] based on the information-theoretic principle of directed information [38, 28, 39].

5 Scaling to High-Dimensional Environments

In the previous section, we presented a novel Bellman operator in combination with a value iteration scheme to combine reward maximization and empowerment. In this section, by leveraging parametric function approximators, we validate our ideas in high-dimensional state-action spaces and when there is no prior knowledge of the state-transition function. In Section 5.1, we devise novel actor-critic algorithms for RL based on our MDP formulation since they are naturally capable of handling both continuous state and action spaces. In Section 5.2, we practically confirm that empowerment can boost RL in the high-dimensional robotics simulator domain of MuJoCo using deep neural networks.

5.1 Empowered Off-Policy Actor-Critic Methods with Parametric Function Approximators

Contemporary off-policy actor-critic approaches for RL [36, 1, 15] follow the policy gradient theorem [63, 13] and learn two parametric function approximators: one for the behavioral policy $\pi_\phi(\mathbf{a}|\mathbf{s})$ with parameters ϕ , and one for the state-action value function $Q_\theta(\mathbf{s}, \mathbf{a})$ of the parametric policy π_ϕ with parameters θ . The policy learning objective usually assumes the form: $\max_\phi \mathbb{E}_{\mathbf{s} \sim \mathcal{D}} [\mathbb{E}_{\pi_\phi(\mathbf{a}|\mathbf{s})} [Q_\theta(\mathbf{s}, \mathbf{a})]]$, where \mathcal{D} refers to a replay buffer [37] that stores collected state transitions from the environment. Following [21], Q-values are learned most efficiently by introducing another function approximator V_ψ for state values of π_ϕ with parameters ψ using the objective:

$$\min_\theta \mathbb{E}_{\mathbf{s}, \mathbf{a}, r, \mathbf{s}' \sim \mathcal{D}} \left[(Q_\theta(\mathbf{s}, \mathbf{a}) - (\alpha r + \gamma V_\psi(\mathbf{s}')))^2 \right], \quad (16)$$

where $(\mathbf{s}, \mathbf{a}, r, \mathbf{s}')$ refers to an environment interaction sampled from the replay buffer (r stands for the observed reward signal). We multiply r by the scaling factor α from our formulation because Equation (16) can be directly used for the parametric methods we propose. Learning policy parameters ϕ and value parameters ψ requires however novel objectives with two additional approximators: one for the inverse dynamics model $p_\chi(\mathbf{a}|\mathbf{s}', \mathbf{s})$ of π_ϕ , and one for the transition function $\mathcal{P}_\xi(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ (with parameters χ and ξ respectively). While the necessity for p_χ is clear, e.g. from inspecting Equation (5), the necessity for \mathcal{P}_ξ will fall into place shortly as we move forward.

In order to preserve a clear view, let's define the quantity $f(\mathbf{s}, \mathbf{a}) := \mathbb{E}_{\mathcal{P}_\xi(\mathbf{s}'|\mathbf{s}, \mathbf{a})} [\log p_\chi(\mathbf{a}|\mathbf{s}', \mathbf{s})] - \log \pi_\phi(\mathbf{a}|\mathbf{s})$, which is short-hand notation for the empowerment-induced addition to the reward signal—compare to Equation (5). We then commence with the objective for value function learning:

$$\min_\psi \mathbb{E}_{\mathbf{s} \sim \mathcal{D}} \left[(V_\psi(\mathbf{s}) - \mathbb{E}_{\pi_\phi(\mathbf{a}|\mathbf{s})} [Q_\theta(\mathbf{s}, \mathbf{a}) + \beta f(\mathbf{s}, \mathbf{a})])^2 \right], \quad (17)$$

which is similar to the standard value objective but with the added term $\beta f(\mathbf{s}, \mathbf{a})$ as a result of joint cumulative empowerment learning. At this point, the necessity for a transition model \mathcal{P}_ξ becomes apparent. In the above equation, new actions \mathbf{a} need to be sampled from the policy π_ϕ for a given \mathbf{s} . However, the inverse dynamics model (inside f) depends on the subsequent state \mathbf{s}' as well, requiring therefore a prediction for the next state. Note also that $(\mathbf{s}, \mathbf{a}, r, \mathbf{s}')$ -tuples from the replay buffer as in Equation (16) can't be used here, because the expectation over \mathbf{a} is w.r.t. to the current policy whereas tuples from the replay buffer come from a mixture of policies at an earlier stage of training.

Extending the ordinary actor-critic policy objective with the empowerment-induced term f yields:

$$\max_\phi \mathbb{E}_{\mathbf{s} \sim \mathcal{D}} [\mathbb{E}_{\pi_\phi(\mathbf{a}|\mathbf{s})} [Q_\theta(\mathbf{s}, \mathbf{a}) + \beta f(\mathbf{s}, \mathbf{a})]]. \quad (18)$$

The remaining parameters to be optimized are χ and ξ from the inverse dynamics model p_χ and the transition model \mathcal{P}_ξ . Both problems are supervised learning problems that can be addressed by log-likelihood maximization using samples from the replay buffer, leading to $\max_\chi \mathbb{E}_{\mathbf{s} \sim \mathcal{D}} [\mathbb{E}_{\pi_\phi(\mathbf{a}|\mathbf{s}) \mathcal{P}_\xi(\mathbf{s}'|\mathbf{s}, \mathbf{a})} [\log p_\chi(\mathbf{a}|\mathbf{s}', \mathbf{s})]]$ and $\max_\xi \mathbb{E}_{\mathbf{s}, \mathbf{a}, \mathbf{s}' \sim \mathcal{D}} [\log \mathcal{P}_\xi(\mathbf{s}'|\mathbf{s}, \mathbf{a})]$.

Coming back to our motivation from Section 3, we propose two novel empowerment-inspired actor-critic approaches based on the optimization objectives specified in this section. The first combines cumulative reward maximization and empowerment learning following Equation (5) which we refer to as empowered actor-critic. The second learns cumulative empowerment values to construct intrinsic rewards following Equation (6) which we refer to as actor-critic with intrinsic empowerment.

Empowered Actor-Critic (EAC). In line with standard off-policy actor-critic methods [36, 15, 21], EAC interacts with the environment iteratively storing transition tuples $(\mathbf{s}, \mathbf{a}, r, \mathbf{s}')$ in a replay buffer. After each interaction, a training batch $\{(\mathbf{s}, \mathbf{a}, r, \mathbf{s}')^{(b)}\}_{b=1}^B \sim \mathcal{D}$ of size B is sampled from the buffer to perform a *single* gradient update on the objectives from Equation (16) to (18) as well as the log likelihood objectives for the inverse dynamics and transition model—see Appendix B for pseudocode.

Actor-Critic with Intrinsic Empowerment (ACIE). By setting $\alpha = 0$ and $\beta = 1$, EAC can train an agent merely focusing on cumulative empowerment learning. Since EAC is off-policy, it can learn with samples obtained from executing *any* policy in the real environment, e.g. the actor of *any other* reward-maximizing actor-critic algorithm. We can then extend external rewards r_t at time t of this actor-critic algorithm with intrinsic rewards $\mathbb{E}_{\pi_\phi(\mathbf{a}|\mathbf{s}_t) \mathcal{P}_\xi(\mathbf{s}'|\mathbf{s}_t, \mathbf{a})} \left[\log \frac{p_\chi(\mathbf{a}|\mathbf{s}', \mathbf{s}_t)}{\pi_\phi(\mathbf{a}|\mathbf{s}_t)} \right]$ according to Equation (6), where (ϕ, ξ, χ) are the result of *concurrent* raw empowerment learning with EAC. This idea is similar to the preliminary work of [29] using non-cumulative empowerment as intrinsic motivation for deep value-based RL with discrete actions in the Atari game Montezuma's Revenge.

5.2 Experiments with Deep Function Approximators in MuJoCo

We validate EAC and ACIE in the robotics simulator MuJoCo [66, 8] with deep neural nets under the same setup for each experiment following [67, 25, 50, 24, 56, 36, 68, 57, 1, 9, 15, 21]—see Appendix C.2 for details. While EAC is a standalone algorithm, ACIE can be combined with any RL algorithm (we use the model-free state of the art SAC [21]). We compare against DDPG [36] and PPO [57] from RLlib [35] as well as SAC on the MuJoCo v2-environments (ten seeds per run [47]).

The results in Figure 2 confirm that both EAC and ACIE can attain better initial performance compared to model-free baselines. While this holds true for both approaches on the pendulum benchmarks (balancing and swing up), our empowered methods can also boost RL in demanding environments like Hopper, Ant and Humanoid (the latter two being amongst the most difficult MuJoCo tasks). EAC significantly improves initial learning in Ant, whereas ACIE boosts SAC in Hopper and Humanoid. While EAC outperforms PPO and DDPG in almost all tasks, it is not consistently better than SAC. Similarly, the added intrinsic reward from ACIE to SAC does not always help. *This is not unexpected as it cannot be in general ruled out that reward functions assign high (low) rewards to lowly (highly) empowered states, in which case the two learning signals may become partially conflicting.*

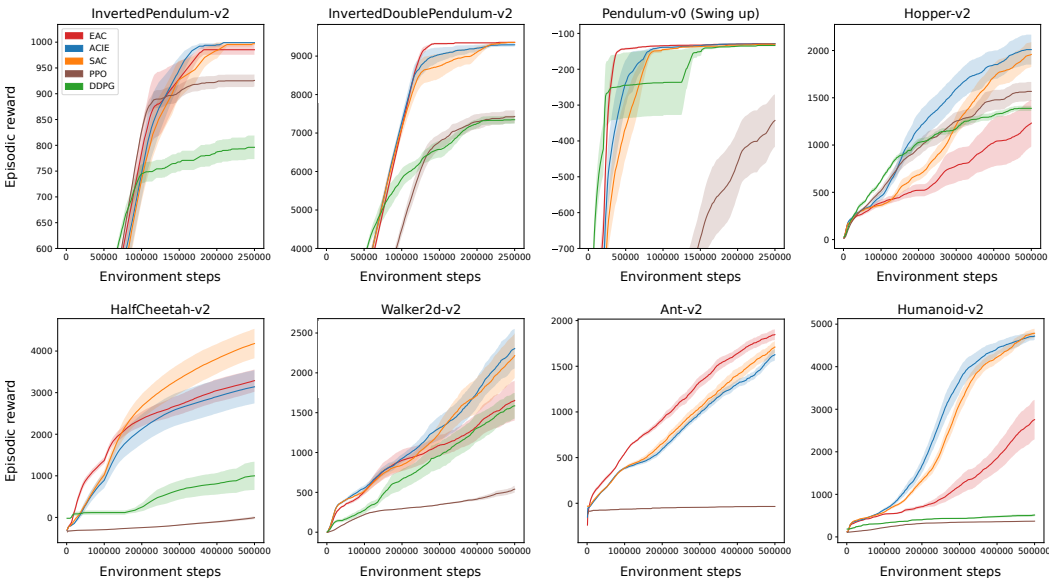


Figure 2: MuJoCo Experiments. The plots show maximum episodic rewards (averaged over the last 100 episodes) achieved so far [9] versus steps—*non-maximum* episodic reward plots can be found in Figure 3. EAC and ACIE are compared to DDPG, PPO and SAC (DDPG did not work in Ant, see [21] and Appendix C.2 for an explanation). Shaded areas refer to the standard error. Both EAC and ACIE improve initial learning over baselines in the three pendulum tasks (upper row). In demanding problems like Hopper, Ant and Humanoid, our methods can boost RL. In terms of final performance, EAC is competitive with the baselines: it consistently outperforms DDPG and PPO on all tasks except Hopper, but is not always better than SAC. Similarly, the ACIE-signal does not always help SAC. This is not unexpected as extrinsic and empowered rewards may partially conflict.

For the sake of completeness, we report Figure 3 which is similar to Figure 2 but shows episodic rewards and *not* maximum episodic rewards obtained so far [9]. Also, limits of y-axes are preserved for the pendulum tasks. Note that our SAC baseline is comparable with the SAC from [22] on Hopper-v2, Walker2d-v2, Ant-v2 and Humanoid-v2 after $5 \cdot 10^5$ steps (the SAC from [21] uses the earlier v1-versions of Mujoco and is hence not an optimal reference). However, there is a discrepancy on HalfCheetah-v2. This was earlier noted by others who tried to reproduce SAC results in HalfCheetah-v2 but failed to obtain episodic rewards as high as in [21, 22], leading to a GitHub issue <https://github.com/rail-berkeley/softlearning/issues/75>. The final conclusion of this issue was that differences in performance are caused by different seed settings and are therefore of statistical nature (comparing all algorithms under the same seed settings is hence valid).

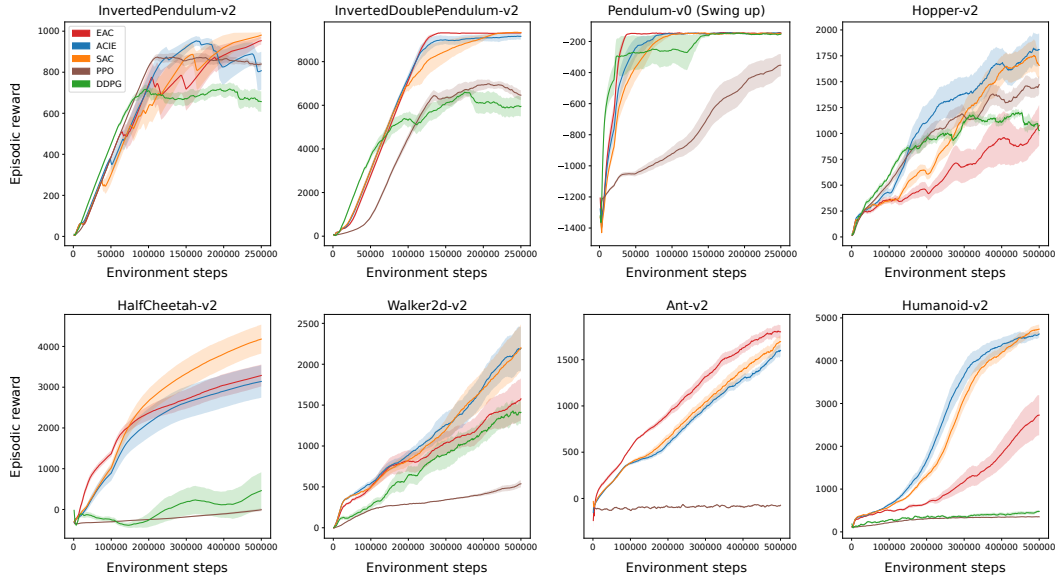


Figure 3: Raw Results of MuJoCo Experiments. The plots are similar to the plots from Figure 2, but report episodic rewards (averaged over the last 100 episodes) versus steps—*not* maximum episodic rewards seen so far as in [9]. For the pendulum tasks, the limits of the y-axes are preserved.

6 Conclusion

This paper provides a theoretical contribution via a unified formulation for reward maximization and empowerment that generalizes Bellman’s optimality principle and recent information-theoretic extensions to it. We proved the existence of and convergence to unique optimal values, and practically validated our ideas by devising novel parametric actor-critic algorithms inspired by our formulation. These were evaluated on the high-dimensional MuJoCo benchmark demonstrating that empowerment can boost RL in challenging robotics tasks (e.g. Ant and Humanoid).

The most promising line of future research is to investigate scheduling schemes that dynamically trade off rewards vs. empowerment with the prospect of obtaining better asymptotic performance. Empowerment could also be particularly useful in a multi-task setting where task transfer could benefit from initially empowered agents.

Acknowledgments

We thank Haitham Bou-Ammar for pointing us in the direction of empowerment.

References

- [1] A. Abdolmaleki, J. T. Springenberg, Y. Tassa, R. Munos, N. Heess, and M. Riedmiller. Maximum a posteriori policy optimisation. In *Proceedings of the International Conference on Learning Representations*, 2018.
- [2] S. Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Transactions on Information Theory*, 18(1):14–20, 1972.
- [3] M. G. Azar, V. Gomez, and H. J. Kappen. Dynamic policy programming with function approximation. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2011.
- [4] R. E. Bellman. *Dynamic Programming*. Princeton University Press, 1957.
- [5] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Springer, 1996.
- [6] W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity, and learning. *Neural Computation*, 13(11):2409–2463, 2001.

- [7] R. Blahut. Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory*, 18(4):460–473, 1972.
- [8] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. OpenAI gym. *arXiv*, 2016.
- [9] K. Chua, R. Calandra, R. McAllister, and S. Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, 2018.
- [10] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley & Sons, 2006.
- [11] I. Csiszar and G. Tusnady. Information geometry and alternating minimization procedures. *Statistics and Decisions*, Suppl.1:205–237, 1984.
- [12] I. M. de Abril and R. Kanai. A unified strategy for implementing curiosity and empowerment driven reinforcement learning. *arXiv*, 2018.
- [13] T. Degris, M. White, and R. S. Sutton. Off-policy actor-critic. In *Proceedings of the International Conference on Machine Learning*, 2012.
- [14] R. Fox, A. Pakman, and N. Tishby. Taming the noise in reinforcement learning via soft updates. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2016.
- [15] S. Fujimoto, H. van Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the International Conference on Machine Learning*, 2018.
- [16] R. G. Gallager. The Arimoto-Blahut algorithm for finding channel capacity. Technical report, Massachusetts Institute of Technology, USA, 1994.
- [17] T. Genewein, F. Leibfried, J. Grau-Moya, and D. A. Braun. Bounded rationality, abstraction, and hierarchical decision-making: An information-theoretic optimality principle. *Frontiers in Robotics and AI*, 2(27), 2015.
- [18] J. Grau-Moya, F. Leibfried, T. Genewein, and D. A. Braun. Planning with information-processing constraints and model uncertainty in Markov decision processes. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2016.
- [19] J. Grau-Moya, F. Leibfried, and P. Vrancx. Soft Q-learning with mutual-information regularization. In *Proceedings of the International Conference on Learning Representations*, 2019.
- [20] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine. Reinforcement learning with deep energy-based policies. *Proceedings of the International Conference on Machine Learning*, 2017.
- [21] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the International Conference on Machine Learning*, 2018.
- [22] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine. Soft actor-critic algorithms and applications. *arXiv*, 2019.
- [23] T. Jung, D. Polani, and P. Stone. Empowerment for continuous agent-environment systems. *Adaptive Behavior*, 19(1):16–39, 2011.
- [24] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.
- [25] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations*, 2014.
- [26] A. S. Klyubin, D. Polani, and C. L. Nehaniv. Empowerment: A universal agent-centric measure of control. In *IEEE Congress on Evolutionary Computation*, 2005.
- [27] A. S. Klyubin, D. Polani, and C. L. Nehaniv. Keep your options open: An information-based driving principle for sensorimotor systems. *PLoS ONE*, 3(12):p.e4018, 2008.
- [28] G. Kramer. *Directed information for channels with feedback*. PhD thesis, University of Manitoba, Canada, 1998.
- [29] N. M. Kumar. Empowerment-driven exploration using mutual information estimation. In *NIPS Workshop*, 2018.

- [30] F. Leibfried and D. A. Braun. A reward-maximizing spiking neuron as a bounded rational decision maker. *Neural Computation*, 27(8):1686–1720, 2015.
- [31] F. Leibfried and D. A. Braun. Bounded rational decision-making in feedforward neural networks. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2016.
- [32] F. Leibfried and J. Grau-Moya. Mutual-information regularization in Markov decision processes and actor-critic learning. In *Proceedings of the Conference on Robot Learning*, 2019.
- [33] F. Leibfried, J. Grau-Moya, and H. Bou-Ammar. An information-theoretic optimality principle for deep reinforcement learning. In *NIPS Workshop*, 2018.
- [34] S. Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv*, 2018.
- [35] E. Liang, R. Liaw, P. Moritz, R. Nishihara, R. Fox, K. Goldberg, J. E. Gonzalez, M. I. Jordan, and I. Stoica. RLlib: Abstractions for distributed reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2018.
- [36] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2016.
- [37] L.-J. Lin. *Reinforcement learning for robots using neural networks*. PhD thesis, Carnegie Mellon University, USA, 1993.
- [38] H. Marko. The bidirectional communication theory—a generalization of information theory. *IEEE Transactions on Communications*, 21(12):1345–1351, 1973.
- [39] J. L. Massey and P. C. Massey. Conversion of mutual and directed information. In *Proceedings of the International Symposium on Information Theory*, 2005.
- [40] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
- [41] V. Mnih, A. Puigdomenech Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2016.
- [42] S. Mohamed and D. J. Rezende. Variational information maximisation for intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems*, 2015.
- [43] G. Montufar, K. Ghazi-Zahedi, and N. Ay. Information theoretically aided reinforcement learning for embodied agents. *arXiv*, 2016.
- [44] O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans. Bridging the gap between value and policy based reinforcement learning. *Advances in Neural Information Processing Systems*, 2017.
- [45] G. Neu, V. Gomez, and A. Jonsson. A unified view of entropy-regularized Markov decision processes. *arXiv*, 2017.
- [46] P. A. Ortega and D. A. Braun. Thermodynamics as a theory of decision-making with information-processing costs. *Proceedings of the Royal Society A*, 469(2153), 2013.
- [47] J. Pineau. *Reproducible, reusable, and robust reinforcement learning*. NIPS Invited Talk, 2018.
- [48] M. Prokopenko, V. Gerasimov, and I. Tanev. Evolving spatiotemporal coordination in a modular robotic system. In *Proceedings of the International Conference on the Simulation of Adaptive Behavior*, 2006.
- [49] A. H. Qureshi, B. Boots, and M. C. Yip. Adversarial imitation via variational inverse reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2019.
- [50] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the International Conference on Machine Learning*, 2014.
- [51] J. Rubin, O. Shamir, and N. Tishby. Trading value and information in MDPs. In *Decision Making with Imperfect Decision Makers*, chapter 3. Springer, 2012.

- [52] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education Limited, 2016.
- [53] C. Salge, C. Glackin, and D. Polani. Empowerment—an introduction. In *Guided Self-Organization: Inception*, chapter 4. Springer, 2014.
- [54] J. Schossau, C. Adami, and A. Hintze. Information-theoretic neuro-correlates boost evolution of cognitive systems. *Entropy*, 18(1):6, 2016.
- [55] J. Schulman, P. Abbeel, and X. Chen. Equivalence between policy gradients and soft Q-learning. *arXiv*, 2017.
- [56] J. Schulman, S. Levine, P. Moritz, M. Jordan, and P. Abbeel. Trust region policy optimization. In *Proceedings of the International Conference on Machine Learning*, 2015.
- [57] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. In *arXiv*, 2017.
- [58] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.
- [59] C. E. Shannon. Coding theorems for a discrete source with a fidelity criterion. *Institute of Radio Engineers, International Convention Record*, 7:142–163, 1959.
- [60] C. A. Sims. Implications of rational inattention. *Journal of Monetary Economics*, 50(3):665–690, 2003.
- [61] S. Still and D. Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.
- [62] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.
- [63] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, 2000.
- [64] S. Tiomkin and N. Tishby. A unified Bellman equation for causal information and value in Markov decision processes. In *arXiv*, 2018.
- [65] N. Tishby and D. Polani. Information theory of decisions and actions. In *Perception-Action Cycle*, chapter 19. Springer, 2011.
- [66] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.
- [67] H. van Hasselt. Double Q-learning. In *Advances in Neural Information Processing Systems*, 2010.
- [68] H. van Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double Q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.
- [69] K. Zahedi, N. Ay, and R. Der. Higher coordination with less control—A result of information maximization in the sensorimotor loop. *Adaptive Behavior*, 18(3-4):338–355, 2010.
- [70] B. D. Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. PhD thesis, Carnegie Mellon University, USA, 2010.