Intrinsic Motivation in Dynamical Control Systems

Stas Tiomkin[1,*] Ilya Nemenman[2] Daniel Polani[3] and Naftali Tishby[4,†]
[1]Computer Engineering Department, Charles W. Davidson College of Engineering, San Jose State University, San Jose, California 95192, USA
[2]Department of Biology, Department of Physics, and Initiative in Theory and Modeling of Living Systems, Emory University, Atlanta, Georgia 30322, USA
[3]Adaptive Systems Research Group, University of Hertfordshire, Hatfield AL10 9AB, United Kingdom
[4]The Rachel and Selim Benin School of Computer Science and Engineering and Edmond and Lilly Safra Center for Brain Sciences (ELSC), Hebrew University of Jerusalem, Jerusalem 96906, Israel

(Received 18 January 2024; accepted 11 July 2024; published 29 August 2024)

Biological systems often choose actions without an explicit reward signal, a phenomenon known as intrinsic motivation. The computational principles underlying this behavior remain poorly understood. In this study, we investigate an information-theoretic approach to intrinsic motivation, based on maximizing an agent's empowerment (the mutual information between its past actions and future states). We show that this approach generalizes previous attempts to formalize intrinsic motivation, and we provide a computationally efficient algorithm for computing the necessary quantities. We test our approach on several benchmark control problems, and we explain its success in guiding intrinsically motivated behaviors by relating our information-theoretic control function to fundamental properties of the dynamical system representing the combined agent-environment system. This opens the door for designing practical artificial, intrinsically motivated controllers and for linking animal behaviors to their dynamical properties.

DOI: 10.1103/PRXLife.2.033009

I. INTRODUCTION

A. Motivation

Living organisms are able to generate behaviors that solve novel challenges without prior experience. Can this ability be explained by a single, generic mechanism? One proposal is that novel, useful behaviors can be generated through *intrinsic motivation* [1], which is defined informally as a set of computational algorithms that are derived directly from the intrinsic properties of the organism-environment dynamics and not specifically learned.

Increasingly, there is a move away from reinforcement learning and its extrinsically specified reward structure [2,3] in the theory and practice of artificial agents, robots, and machine learning more generally [4–20]. A specific class of such intrinsic motivation algorithms for artificial systems is known as *empowerment maximization*. It proposes that agents should maximize the mutual information [21] between their potential actions and a subsequent future state of the world [22]. This corresponds to maximizing the diversity of future world states achievable as a result of the chosen actions, potentiating a broader set of behavior options in the future. This measure replaces the traditional value function of reinforcement learning. However, importantly, it does not assume the definition of a problem-specific cost function, which would, for instance, encode additional domain knowledge. Instead, empowerment derives from the intrinsic properties of the system dynamics itself.

Intrinsically motivated synthetic agents develop behaviors that are atypical for inanimate engineered systems and often resemble those of simple living systems. Interestingly, potentiating future actions is also a key part of the success of modern reward-based training algorithms [8,23,24]. As an example relevant to the current work, consider a pendulum at the up-vertical orientation. Here, the agent can kick it easily in both directions—the information between the actions and the future states is now high, as the entropy of futures is high, while the variety of end states still can be controlled by the actions. Compare this to the pendulum pointing down. Here the futures are basically fluctuating around the equilibrium point. As any actions have to work uphill, the variety of end states is lower than for an up-pendulum.

Despite the successes of empowerment maximization, it remains unclear how well it can be used as a general intrinsic motivation principle. There are many different versions of intrinsic motivation related to empowerment, and their relation to each other is unknown [20,23,25]. Additionally, most work on empowerment maximization has relied on simulational case studies and ad hoc approximations, and analytical results are scarce. To gain insight, it is important to link empowerment to other, better-understood characterizations of the systems in question. Finally, calculating the mutual information between two interlinked processes in the general case

*Contact author: stas.tiomkin@sjsu.edu
†Prof. Naftali Tishby passed away when this work was in development. This project began under his leadership when Stas Tiomkin was a Ph.D. student in his group.

is a challenging task [26,27], which has so far limited the use of empowerment maximization to simple cases.

In this work, we unify different versions of intrinsic motivation related to the empowerment maximization paradigm. Here our main contribution is in showing analytically that empowerment-like quantities are linked to the sensitivity of the agent-environment dynamics, which is measured by the generalization of Lyapunov exponents that we introduce. This connects empowerment maximization to well-understood properties of dynamical systems. Since highly sensitive regions of the dynamics potentiate many diverse future behaviors, the connection to dynamical systems also explains why empowerment-based intrinsic motivations succeed in generating behaviors that resemble those of living systems.

The analytical results allow us to develop a practical computational algorithm for calculating empowerment for complex scenarios in continuous space and the continuous time limit, which is the second major contribution of the paper. We apply the algorithm to standard benchmarks used in intrinsic motivation research [14,16,28]. Specifically, a controller based on the efficient calculation of empowerment manages to balance an inverted pendula without extrinsic rewards, and without fine-tuning the control strategy to the dynamical equations describing the system. This opens the door for designing complex robotic intrinsically motivated agents with systematically computed—rather than heuristically estimated—empowerment.

B. Overview of the method

Consider the mutual information between a control process of a given duration ("time horizon") and a subsequent resulting process dynamics. If we start in a given system state x_0 and maximize the mutual information over all possible control processes, we obtain the empowerment in state x_0 for the given time horizon. While the computation of this quantity is, in principle, a numerically solvable problem in systems with discrete time and state spaces, its computational complexity scales exponentially with the time horizon, and transferring this to continuous time and space poses significant additional challenges.

In this paper, we propose an efficient method for the computation of empowerment in continuous space and the continuous time limit, under some assumptions. For this, we discretize time and consider the analyzed system in the linear regime approximation for small Gaussian control and perturbation signals around an unperturbed trajectory. This approximation allows us to formulate the mutual information maximization as calculating the capacity of a linear Gaussian channel, with the channel properties computed from the linearization of the dynamics around the zero-control trajectory. This capacity can be computed efficiently, and our numerical experiments show that its values converge benignly as the discretization time step interval approaches zero, thereby obtaining a numerical value for empowerment in the continuum.

To match the discretized representation with the formulation of the original continuous system, we introduce a parallel notation for the continuous versus the discretized version of the system in the next section.

II. RESULTS

A. Preliminaries

1. Notation

We consider an agent that takes on states $x(t) \in \mathcal{X} := \mathbb{R}^{d_x}$, evolving in time under the dynamics f with (small) stochastic perturbations $\eta(t) \in \mathbb{R}^{d_x}$. Via its (small) actions, $a(t) \in \mathcal{A} := \mathbb{R}^{d_a}$ filtered through the control gain g, the agent can affect the dynamics of the system:

$$dx(t) = f(x(t))dt + g(x(t))da(t) + d\eta(t)\,. \tag{1}$$

Here $d\eta$ denotes the system noise, modeled as a Wiener process. The agent's actions $a(t)$ are modeled by a stochastic control process with variance σ_t^2 controlled by the agent and with a mean of zero. This models the potential effect of actions centered around the null action.

To compute various quantities of interest, we will consider a discretized version of this system, for which we adopt a modified notation. To distinguish it from the continuous version, we replace the continuous time in parentheses by an integer index, $x_k := x(t + k \cdot \Delta t)$. Here Δt denotes the physical time step, and we adopted the convention that $x_0 = x(t)$, so that the index corresponding to the current physical time, t, is chosen as 0. We will consider trajectories of a fixed duration, and the agent will apply actions over a part of that trajectory. Note that we switch between referring to action as da or a depending on whether we consider the continuous or the discrete case, and we hope this does not lead to confusion. We denote by T_e the time index of the very last state of the trajectory, which we also refer to as the *time horizon*. We further use T_a to denote the (discretized) duration of the action sequence. Then state, control, and perturbation trajectories at finite equidistant times, $\{t + k \cdot \Delta t\}_{k=0}^{T}$, are denoted by $x_0^{T_e} \equiv \{x_k\}_{k=0}^{T_e}$, $a_0^{T_a} \equiv \{a_k\}_{k=0}^{T_a}$, and $\eta_0^{T_e} \equiv \{\eta_k\}_{k=0}^{T_e}$, respectively. For consistency with the control theory literature, we write a trajectory in the reverse order, e.g., $x_0^{T_e} = (x_{T_e}, \ldots, x_0)$. When we wish to emphasize the continuous nature of the underlying process, we will write $t_e \equiv t + T_e \cdot \Delta t$ and $t_a \equiv t + T_a \cdot \Delta t$ for explicitly continuous times.

2. Reinforcement learning vs intrinsic motivation

To elicit a desired behavior in an agent, one typically uses reinforcement learning (RL). RL is task-specific, and an agent needs an *extrinsic* feedback about its performance from a *reward* function to learn the behavior. The precise construction of this reward function is critical to achieve a desired performance in a short training time [2]. Some of the complications include a significant degree of arbitrariness when choosing among reward functions with equivalent performance [29] and the difficulty of translating an often vague desired behavior into a concrete reward function. Furthermore, complex behaviors consist of combinations of shorter sequences. Designing a reward function capable of partitioning the solution into such parts and hence learning it in a realistic time is hard [30]. In contrast to this, in living systems, acquisition of skills often starts with task-unspecific learning. This endows organisms with *potentiating* skills, which are not rewarding on their own. This is then followed by task-oriented specialization, which combines task-unspecific behaviors into complex and

explicitly rewarding tasks [1,31]. While specific tasks are often refined with the help of an extrinsic reinforcement, the potentiating tasks usually are intrinsically motivated [9,32].

3. Empowerment

The type of intrinsic motivation we focus on is *empowerment*. Empowerment is based on information-theoretic quantities [4,20,23,32–39]. It defines a pseudoutility function on the state space, based on the system dynamics only, without resorting to a reward. Formally, we express the dynamics of the system by the conditional probability distribution $p(x_{T_e} \mid a_0^{T_e-1}, x_0)$ of the resulting state when one starts in a state x_0 and subsequently carries out an action sequence $a_0^{T_e-1}$ (recall that the notation is defined in the Notation section above). Then the empowerment $\mathcal{C}(x_0)$ is a function of the starting state, x_0. It is given by the maximally achievable mutual information (the channel capacity [21]) between the control action sequence of length T_e and the final state when starting in the state x_0:

$$\mathcal{C}(x_0) := \max_{p(a_0^{T_e-1}|x_0)} I\big(X_{T_e}; A_0^{T_e-1} \big| x_0\big). \tag{2}$$

Here $p(\cdot)$ denotes a probability density or a probability distribution function, and I is the mutual information [21]

$$I\big(X_{T_e}; A_0^{T_e-1} \big| x_0\big) = H\big(X_{T_e} \big| x_0\big) - H\big(X_{T_e} \big| A_0^{T_e-1}, x_0\big). \tag{3}$$

H is the entropy, and conditioning an entropy on a random variable means the entropy of the conditional distribution, averaged over the conditioning variable.

In effect, empowerment measured information that the action sequence has about the end state. High empowerment requires high entropy of the end state, but also small entropy of the states conditional on the action sequence producing the final state. In other words, it is not enough to have diverse end states, but these must have been induced by the actions. Variability only counts in the empowerment if it can be specifically caused by the agent.

In contrast, if only the end state entropy were important, an agent would be induced to seek out, say, staying in front of a white-noise TV screen. However, unless the "pixels" of the screen are controllable by the agent, this white noise would not contribute to the empowerment.

The empowerment $\mathcal{C}(x_0)$ depends on both the state, x_0, and the time horizon, T_e. However, for notational convenience, we omit all parameters from the notation except for the dependency on x_0.

Locally maximizing empowerment (e.g., by following its gradient over x_0) guides an agent to perform actions atypical within the natural dynamics of the system. Indeed, since empowerment measures the diversity of achievable future states, maximizing it increases this diversity ("empowers" the agent—hence the name). Thus it is expected to be particularly useful for learning potentiating tasks [9]. Crucially, empowerment quantifies the relation between the final state and the *intentional* control, rather than the diversity of states due to the stochasticity of the system. In particular, it is not just the entropy of a passive diffusion process in the state variables, but of the subprocess that the agent can actively generate. Furthermore, it quantifies diversity due to *potential* future action sequences, which are not then necessarily carried out.

Empowerment is typically used in conjunction with a sensor through which the agent observes the states resulting from an action sequence. In the continuum, this can be modeled via observation noise applied to the outcome states.

Empowerment is typically used in the form of the *empowerment maximization principle* [17], where $\mathcal{C}(x_0)$ is treated as a pseudoutility function. At each time step, the agent chooses an action to greedily optimize its expected empowerment at the next time step, climbing up in its empowerment landscape, to eventually achieve a local maximum of $\mathcal{C}$,

$$a^*(x(t)) = \operatorname*{argmax}_{a\in\mathcal{A}} \mathbb{E}_\eta[\mathcal{C}(x(t) + f(x(t))\Delta t' + g(x(t))a\Delta t' + \eta_{\Delta t'})]. \tag{4}$$

Here $\mathcal{A}$ is the set of permitted actions, $\Delta t'$ is a small time step used to simulate the actual behavior of the system, $a \in \mathcal{A}$ is the candidate action kept fixed for the duration of $\Delta t'$, and $\eta_{\Delta t'}$ is the Wiener process integrated over the time interval $\Delta t'$. An empowerment-maximizing agent generates its behavior by repeating this action selection procedure for each decision step it takes.

The time step $\Delta t'$ for the empowerment-greedy action is selected as a small fixed value. Note that it is selected independently from the time step Δt that is used to discretize (1) for the purpose of computing empowerment. Empowerment in continuous scenarios will be computed by letting $\Delta t \to 0$.

Crucially, no general analytical solutions or efficient algorithms for numerical estimation of empowerment for arbitrary dynamical systems are known, limiting adoption of the empowerment maximization principle. Our goal is to provide a method to calculate it under specific approximations.

B. Empowerment in dynamical systems

1. The linear-response approximation

To relate empowerment to traditional quantities used to describe dynamical systems, we assume that the control signal a and process noise η in (1) are small. This is true in some of the most interesting cases, where the challenge is to solve a problem with only *weak controls* that cannot easily "force" a solution. Under this assumption, (1) is approximated by a linear time-variant dynamics around the trajectories of the autonomous dynamics (i.e., for $a = 0$ and $\eta = 0$). To proceed, we now introduce the following notation. We define $\bar{x}_s$ as the sth step of the trajectory in the discretized approximation of the dynamics (1), with $\bar{f}(\bar{x}) := \bar{x} + f(\bar{x})\Delta t$, $\bar{g}(\bar{x}) := g(\bar{x})\Delta t$, and $\bar{h}(\bar{x}) := \Delta t$ [40]:

$$\bar{x}_s = \bar{f}(\bar{x}_{s-1}) + \bar{g}(\bar{x}_{s-1})a_{s-1} + \bar{h}(\bar{x}_{s-1})\eta_{s-1}, \tag{5}$$

where $\bar{x}_0 = x_0 \equiv x(t)$. For example, $\bar{x}_3 = \bar{f}(\bar{f}(\bar{f}(\bar{x}_0) + \bar{g}(\bar{x}_0)a_0 + \bar{h}(\bar{x}_0)\eta_0) + \bar{g}(\bar{x}_1)a_1 + \bar{h}(\bar{x}_1)\eta_1) + \bar{g}(\bar{x}_2)a_2 + \bar{h}(\bar{x}_2)\eta_2$. We denote this recursive mapping from $\bar{x}_0$ to $\bar{x}_s$ by F, $\bar{x}_s = F(\bar{x}_0; a_0^{s-1}, \eta_0^{s-1})$. Then the sensitivity of the state at the time step s to the action at the time step r can be calculated via the iterated differentiation chain rule applied to the state derivative of the dynamics F at $a = 0$ and $\eta = 0$:

$$\frac{\partial \bar{x}_s}{\partial a_r} = \prod_{\tau=r+2}^{s} \nabla_{\bar{x}} \bar{f}(\bar{x}_{\tau-1})\, \bar{g}(\bar{x}_r), \tag{6}$$

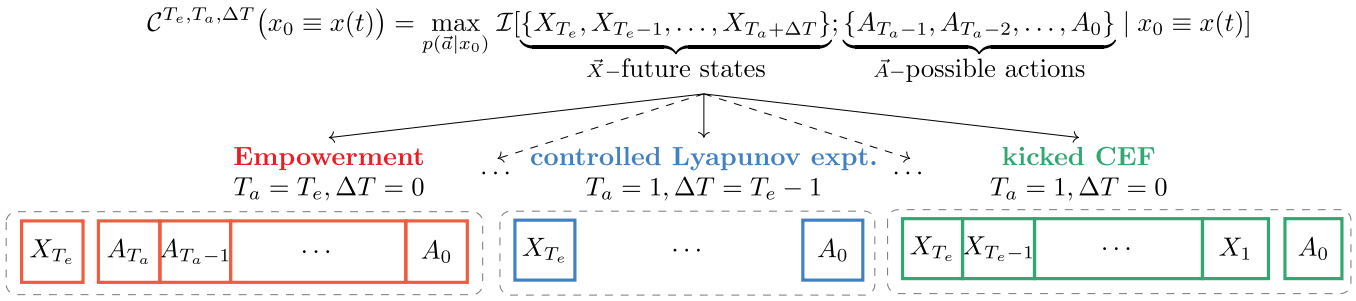

FIG. 1. Unified view on information theoretical intrinsic motivation, for discretized process sequences X_T (states) and A_T (actions). Starting at time x_0 [i.e., $x(t)$], potential actions A_T are applied for T_a times. Following that, after waiting for ΔT time steps, the future system trajectory is considered until T_e. A controlled Lyapunov exponent is a Lyapunov exponent, but only in directions controlled by the agent; cf. (11). "Kicked CEF" refers to a variant of Causal Entropic Forcing [20], with the addition that an action kicks the system at the beginning of a trajectory. The dashed-line blocks illustrate the correspondence in time between state and action sequences. For more details, see the section on Generalized Empowerment.

where $\nabla_{\bar{x}} \bar{f}(\bar{x}_\tau)$ is the $d_x \times d_x$ Jacobian matrix of $\bar{f}$. Specifically, the (i, j)th entry of $\nabla_{\bar{x}} \bar{f}(\bar{x}_\tau)$ is $\frac{\partial \bar{f}_i(\bar{x}_\tau)}{\partial \bar{x}_{\tau,j}}$, where indices i, j stand for components of the vectors x and f. For $s = r + 1$, the expression in (6) evaluates to $\frac{\partial \bar{x}_{r+1}}{\partial a_r} = \bar{g}(x_r)$. Analogously, the sensitivity of $\bar{x}_s$ to the perturbation, η_r, is given by $\frac{\partial \bar{x}_s}{\partial \eta_r} = \prod_{\tau=r+2}^{s} \nabla_{\bar{x}} \bar{f}(\bar{x}_{\tau-1})\, \bar{h}(\bar{x}_r)$.

Now we finally define the linear response of the sequence of the system's states $x_{s_1}^{s_2}$ to a sequence of small actions $\delta a_{r_1}^{r_2}$ by the agent

$$\mathcal{F}_{r_1,r_2}^{s_1,s_2}(x_0) = \begin{bmatrix} \frac{\partial \bar{x}_{s_2}}{\partial a_{r_2}} & \frac{\partial \bar{x}_{s_2}}{\partial a_{r_2-1}} & \cdots & \frac{\partial \bar{x}_{s_2}}{\partial a_{r_1}} \\ \frac{\partial \bar{x}_{s_2-1}}{\partial a_{r_2}} & \frac{\partial \bar{x}_{s_2-1}}{\partial a_{r_2-1}} & \cdots & \frac{\partial \bar{x}_{s_2-1}}{\partial a_{r_1}} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial \bar{x}_{s_1}}{\partial a_{r_2}} & \frac{\partial \bar{x}_{s_1}}{\partial a_{r_2-1}} & \cdots & \frac{\partial \bar{x}_{s_1}}{\partial a_{r_1}} \end{bmatrix}_{d_x \cdot s \times d_a \cdot r}, \tag{7}$$

where $s = s_2 - s_1 + 1$, $r = r_2 - r_1 + 1$, $s + \Delta T + r - 1 = T_e$, and the entries are computed via (6). Usually we consider situations in which the agent applies its controls for r time steps, and then after a gap observes the state for s steps. That is, $s_1 = r_2 + 1 + \Delta T$, where $\Delta T \geqslant 0$ is the gap between the end of the control sequence and the start of the observations, as defined in Fig. 1. Analogously to Eq. (7), we define $\mathcal{H}_{r_1,r_2}^{s_1,s_2}(x_0)$ with the corresponding entries, $\frac{\partial \bar{x}_s}{\partial \eta_r}$.

Notice that traditional definitions of sensitivity of a dynamical system to its controls are blocks $\mathcal{F}_{r_1',r_2'}^{s_1',s_2'}$ in this overall sensitivity matrix, $\mathcal{F}_{r_1,r_2}^{s_1,s_2}$. For example, if $r_1' = r_2' = 0$, $\Delta T' = T_e - 1$, and $s_1' = T_e$, then $s_2' = T_e$, and the sensitivity matrix collapses to just the entries that measure the sensitivity of the current state to the controls during the immediately preceding time step, $\mathcal{F}_{0,0}^{T_e,T_e}(x_0) = \frac{\partial \bar{x}_{T_e}}{\partial a_0}$. This is also the blue block of the overall sensitivity matrix, (7). Other colored boxes in (7) will be explained later.

With the definitions above, in the linear-response regime, the effect of a sequence of (small) actions and perturbations on a sequence of states becomes

$$\delta x_{s_1}^{s_2} = \mathcal{F}_{r_1,r_2}^{s_1,s_2}(x_0)\, \delta a_{r_1}^{r_2} + \tilde{\eta}, \tag{8}$$

where δa and δx are the reverse-time-ordered vectors of small actions and the induced deviations of states (which themselves can be vectors).

Here $\tilde{\eta} = \mathcal{H}_{r_1,r_2}^{s_1,s_2}(x_0)\, \delta \eta_{r_1}^{r_2} + \eta_o$ models the effect of perturbation noise, $\delta\eta$, and the noise, η_o, of the subsequent observation of the state perturbation $\delta x_{s_1}^{s_2}$, which we assume is Gaussian. The choice of observation noise takes into account the imperfect observation of the outcome states by the agent. This observation noise effectively determines the resolution at which the end state is considered.

We note that the approximation is linear with respect to variations in the trajectory only; the calculation remains nonlinear with respect to state.

2. *Generalized empowerment*

Since the entire dynamics is now linear, cf. Eq. (8), we can consider formally the effects of arbitrary length sequences of actions on arbitrary length sequences of future states. In other words, we can define the *generalized empowerment*,

$$\mathcal{C}^{T_e,T_a,\Delta T}(x_0) := \max_{p(\vec{a}|x_0)} I\big(X_{T_a+\Delta T}^{T_e}; A_0^{T_a-1} \big| x_0\big). \tag{9}$$

Here, T_a denotes the number of time steps at which actions are performed, ΔT is the time gap between the action sequence and the beginning of the observation of the resulting states, and T_e is the last step in that observed sequence. That is, $\mathcal{C}^{T_e,T_a,\Delta T}$ measures the maximum mutual information contained in the state sequence, $X_{T_a+\Delta T}^{T_e}$, about the preceding action sequence, $A_0^{T_a-1}$, rather than in the final state only, X_{T_e}, like empowerment does; cf. Eq. (2).

We observe that computing the generalized empowerment in discretized time with an arbitrary discretization step and an arbitrary time horizon T_e reduces to a traditional calculation of the channel capacity of a linear Gaussian channel [21], though with a large number of dimensions reflecting both the duration

of the signal and the duration of the response. Specifically,

$$\mathcal{C}^{T_e,T_a,\Delta T}(x_0) = \max_{\substack{\sigma_i \geqslant 0 \\ \sum_i \sigma_i = P}} \frac{1}{2} \sum_{i=1}^{d_x} \ln(1 + \rho_i(x_0)\sigma_i). \tag{10}$$

Here $\rho_i(x_0)$ are the singular values of the appropriate submatrix $\mathcal{F}^{s_1',s_2'}_{r_1',r_2'}(x_0)$; for example, the traditional empowerment corresponds to the red-dashed submatrix in (7). Further, P is the *power* of the control signal Δa over the whole control period, and $\sigma_i \geqslant 0$ is that part of the overall power of the control signal that is associated with the ith singular value (called *channel power*). The channel power can be computed by the usual *water-filling* procedure [21]. Note that here we denote P as power, as per control-theoretic convention, but since we fix the time interval over which it is applied, the units of P are those of energy. As per our *weak control* assumption, we assume P to be suitably small.

With (10), calculation of any generalized empowerment becomes tractable, at least in principle. This also shows explicitly that the (generalized) empowerment is a function of the sensitivity matrix $\mathcal{F}$, and with it of quantities used to characterize dynamics, such as the Lyapunov exponents.

To compute $\mathcal{C}^{T_e,T_a,\Delta T}(x_0)$ efficiently for an arbitrary dynamical system (1) and arbitrary long time horizons and arbitrary small discretization steps, we start by discretizing the time and calculating the linear-response matrix $\mathcal{F}$. While in this paper we do this by analytical differentiation, numerical differentiation can be used whenever f is unknown. We then calculate the singular values of $\mathcal{F}$; this is straightforward on modern computers for dimensionalities of up to a few hundred. Finally, we apply the "water filling" procedure to find the set of channel powers σ_i to match the available total power P in (10), and from there we calculate the (generalized) empowerment value.

To determine the action of the agent, we finally use (8) to compute (10) and plug this into (4) to select the actual action of the agent. Because of the linear approximation assumption, the effect of the noise averages out through the expectation in (4). We can therefore just drop the expectation and the noise term in the algorithm below. In all our examples below, the agent will employ this approach.

ALGORITHM 1. In the numerical experiments, we use the simplest possible control law, greedy control, for maximizing the empowerment, which is summarized by the pseudocode below and implemented in [41].

Intrinsically-motivated control
Require: x, f, P, T_e, T_a, ΔT
1: **Repeat**
2: Calculate sensitivity gain, $\mathcal{F}(x)$ Eq. (7)
3: Calculate channel capacity, $\mathcal{C}(x)$ Eq. (10)
// Derive an optimal action, random process averages out in linear approximation:
4: $a^*(x) = \operatorname{argmax}_{a \in \mathcal{A}} \mathcal{C}(x + f(x)\Delta t' + g(x)a\Delta t')$ Eq. (4)
5: $x \leftarrow f(x, a^*(x))$ // Ascend empowerment
6: **until** 'convergence to a locally maximally-empowered state'

3. *Connecting generalized empowerment to related quantities*

Generalized empowerment with different durations of action and observation sequences is related to various quantities describing dynamical systems, including those defining intrinsic motivation [8,20,23,42]. For example, Causal Entropic Forcing (CEF) [20] is defined as actions that maximize the entropy of future trajectories of a system. With $T_a = 1$ and $\Delta T = 0$, $\mathcal{C}^{T_e,T_a,\Delta T}$ in (9) measures the immediate consequences of a single action on a trajectory with a fixed time horizon T_e. Maximizing $\mathcal{C}^{T_e,T_a,\Delta T}$ is then equivalent to choosing actions that maximize *susceptibility*, and not the entropy of trajectories with a given time horizon. In other words, one can interpret $\mathcal{C}^{T_e,1,0}$ as a "kicked," or agent-controllable, version of CEF, where just the first action can be selected by the agent at any time, and uncontrolled future variability is discarded in action planning (see Fig. 1 for an illustration). Such a kicked CEF corresponds to the green submatrix in (7).

Now consider the top right corner (blue) of (7) with $T_e = T_a = 1$, or, equivalently, $s_2' = s_2$ and $s_1' = s_2' - 1$. In the limit of a very long horizon, $s_2 \to \infty$, the appropriate submatrix of $\mathcal{F}$ becomes

$$\Lambda \equiv \lim_{s_2 \to \infty} \left(\left(\frac{\partial \bar{x}_{s_2}}{\partial a_{r_1}} \right) \left(\frac{\partial \bar{x}_{s_2}}{\partial a_{r_1}} \right)^{\dagger} \right)^{\frac{1}{s_2}}, \tag{11}$$

where $\dagger$ is the transpose, and $\frac{\partial \bar{x}_{s_2}}{\partial a_{r_1}}$ is given by (6). In the special case that the control gain is the identity, $g(x) = x$, the logarithm of the eigenvalues of Λ reduces to the usual characteristic Lyapunov exponents of the dynamical system [43]. However, once a more general control gain is applied, the action-controlled perturbation, a_{r_1} may be able to affect only a part of the state space. This means that Λ not only is a generalized empowerment with specific indices, but it is also a specialization of the concept of Lyapunov exponents to the controllable subspace. Thus we refer to the log-spectrum of Λ as the *control Lyapunov exponents*; cf. Fig. 1.

In summary, (9) and the linearization, (7), provide a unified view of various sensitivities of the dynamics to the controls, and hence on various versions of intrinsic motivation.

C. Intrinsic motivation in power-constrained agents

An agent controlling a system with unconstrained actions can trivially reach any state in a controllable dynamical system [44] by simply forcing their desired outcome without sophisticated control. Thus to render the setup interesting, we consider only power-constrained or *weak* agents. To show that empowerment maximization, in the linearized regime, is an efficient control principle, we use it to stabilize a family of inverted pendula (single pole, double pole, and cart-pole), which are simple, paradigmatic models of important phenomena, such as human walking [45].

Solutions for the stabilization problem are known. They require the accumulation of energy by swinging the pendulum back and forth into resonance without overshooting and then to keep the pendulum upright. When details of the system are not specified *a priori*, this solution needs to be learned by the agent. Finding such an indirect control policy by traditional reinforcement learning is nontrivial [3], since the increasing oscillations require a long time for the balancing to take

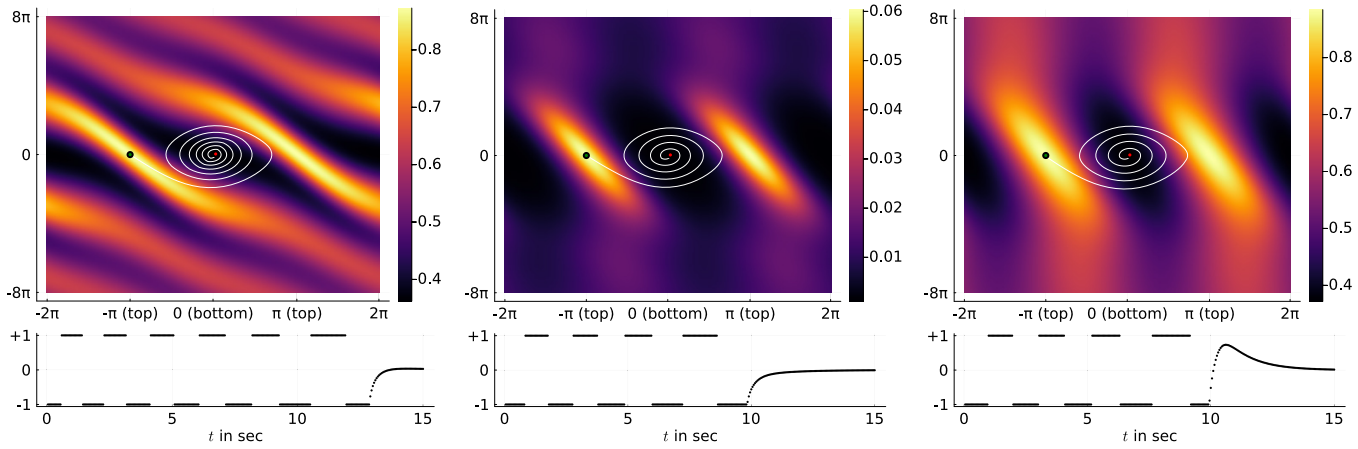

FIG. 2. Intrinsic motivation based control in the power-constrained regime. Top row: generalized empowerment landscapes in the linear approximation for empowerment (left), controlled Lyapunov exponent (middle), and kicked CEF (right) versions of the problem, plotted against θ (horizontal axis) and $\dot{\theta}$ (vertical axis), measured in rad and rad/s, respectively. The color bars indicate the empowerment values in bits. Black dots in each panel are the final state, and white lines are the trajectories of the pendulum, starting at the bottom denoted by the red dots. Bottom row: the control signals chosen from the generalized empowerment maximization as a function of time. Here the time horizon is $t_e = 0.5$ s.

place, and the acquisition of informative rewards indicating success is significantly delayed. As we will show, it is precisely in such situations that intrinsic motivation based on empowerment is especially useful, since it is determined from only comparatively local properties of the dynamics along the present trajectory and its potential future variations.

Here $x, f, P, T_e, T_a, \Delta T$ are the initial state, the system dynamics, the power of the control signal, and the time window parameters for state and action sequences, as explained in the paragraph after Eq. (9). The actions used to calculate empowerment (hypothetical futures) are stochastic. However, when actually taking the action (line 4 in the pseudocode), the specific deterministic action derived in Eq. (4) is chosen in a greedy fashion. The dynamics, f, is used as a forward model for the projected actions to calculate empowerment of the successor states; only after choosing the action that maximizes the empowerment of the successor state is it used, in line 4, to actually carry out the step.

Note that empowerment maximization provides the effective utility function, which is computed for the given dynamics of the system. No hand-crafted reward is required to generate the behavior. Instead empowerment defines the utility structure without knowing the dynamics as an input.

In the following, we demonstrate the proposed general formalism in dynamical control systems, where control, a, and perturbation, η, represent controllable and uncontrollable forces, affecting the state through the control gain $g(x)$.

1. Inverted pendulum

We start with a relatively simple task of swinging up and stabilizing an inverted pendulum without an external reward. With an angle of θ (in radians) from the upright vertical, the equations of motion of the pendulum are

$$\begin{pmatrix} d\theta(t) \\ d\dot{\theta}(t) \end{pmatrix} = \begin{pmatrix} \dot{\theta}(t)dt \\ \frac{g}{l}\sin(\theta(t))\,dt + \frac{da(t)}{ml^2} + \frac{dW(t)}{ml^2} \end{pmatrix}, \qquad (12)$$

where $\dot{\theta}$ is the angular velocity of the pendulum, $m = 1$ kg is its mass, $l = 1$ m is the length, $a(t)$ is the torque applied by the agent, $g = 9.8\,\mathrm{m/s^2}$ is the free fall acceleration, and $dW(t)$ is a Wiener process.

We apply a (stochastically chosen) control signal $a(t)$ for the duration T_e and observe the final state $\tilde{\theta} = \theta + \tilde{\eta}_{\mathrm{obs}}$, where $\tilde{\eta}_{\mathrm{obs}}$ is the standard Gaussian observation noise at the final state. Empowerment is then given by the maximally achievable mutual information between $a(t)$ and $\tilde{\theta}$ at a given power level for $a(t)$, i.e., the channel capacity between the two.

We now apply our empowerment-based control protocol, (4), to the inverted pendulum. We calculate the empowerment landscape by using the time-discretized version of Eqs. (1) and (12). For this, we map the deterministic part of the dynamics [f, g in (1)] onto discrete time as per (5). We then compute the channel capacity by applying (10) using the singular values from (8), where states are given by $(\theta, \dot{\theta}) \in \mathbb{R}^{d_x}$, and actions consist of applying a torque a. The landscapes for the original empowerment, the controlled Lyapunov exponent, and the kicked CEF versions of the problem, all with the time horizons of $t_e = 0.5$ s and the discretization $\Delta t = 10^{-3}$, are shown in Fig. 2. Then, from each state, we choose the control action to greedily optimize the generalized empowerment. The panels in the upper row in this figure also show trajectories obtained this way. The lower row shows time traces of the control signal derived from the generalized empowerment maximization. In all cases, initially, the agent drives the pendulum at the maximum allowable torque, which we set to be power-constrained to ± 1 Nm. Around 13, 10, and 10 s after the start (for the three versions of the empowerment, respectively), the pendulum accumulates enough energy to reach the vertical, and the agents reduce the torques to very small values, $a \ll 1$ Nm, which are now sufficient to keep the pendulum in the upright position and prevent it from falling. It is striking that the generalized empowerment landscapes and their induced trajectories are qualitatively similar to those that would be generated by an optimal value function for the stabilization task, derived by standard optimal control techniques

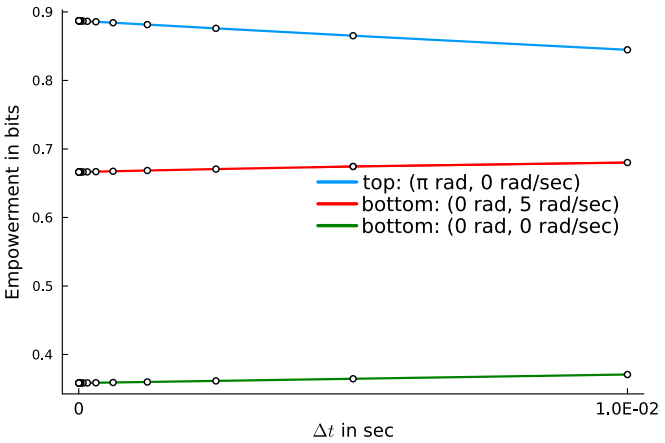

FIG. 3. Convergence of the method for $\Delta t \to 0$ and $t_e = 0.5$ s at three different states "top with zero velocity," "bottom with velocity equal to 5 rad/s," and "bottom with zero velocity" in blue, red, and green, respectively. As time resolution is refined twofold at every stage, one arrives at a well-defined value for the empowerment estimation as $\Delta t \to 0$. The numerical stability of this limit approximation is consistent throughout the landscape.

based on a reward specifically designed to achieve the top position [3].

In our analysis, we chose a particular discretization $\Delta t = 10^{-3}$ s, and we need to show that our results depend only weakly on this choice. For this, we repeat our analysis at different Δt. Figure 3 shows the dependence of the maximum value of the original empowerment (black dot in the left panel of Fig. 2) on Δt. To the extent that the estimate converges to a well-defined number linearly as $\Delta t \to 0$, the discrete time dynamics provides a consistent approximation to the continuous time dynamics.

2. Double pendulum

Now we show that the empowerment maximization formalism is capable of dealing with more challenging problems, such as a power-constrained control of a (potentially chaotic) double pendulum [16], Fig. 4, with equations of motion:

$$
\begin{aligned}
d\ddot{\theta}_1(t) &= -\frac{1}{d_1(t)}(d_2(t)\ddot{\theta}_2(t) + \phi_1(t)),\\
d\ddot{\theta}_2(t) &= \frac{1}{m_2\ell_{c_2}^2 + I_2 - \frac{d_2^2(t)}{d_1(t)}}\Big(da(t) + dW(t) + \frac{d_2^2(t)}{d_1(t)}\phi_1(t)\\
&\quad - m_2\ell_1\ell_{c_2}\dot{\theta}_1(t)^2\sin\theta_2(t) - \phi_2(t)\Big),
\end{aligned}
\tag{13}
$$

with

$$
\begin{aligned}
d_1(t) &= m_1\ell_{c_1}^2 + m_2\big[\ell_1^2 + \ell_{c_2}^2 + 2\ell_1\ell_{c_2}\cos\theta_2(t)\big] + I_1 + I_2,\\
d_2(t) &= m_2\big[\ell_{c_2}^2 + \ell_1\ell_{c_2}\cos\theta_2(t)\big] + I_2,\\
\phi_1(t) &= -m_2\ell_1\ell_{c_2}\dot{\theta}(t)^2\sin\theta_2(t) - 2m_2\ell_1\ell_{c_2}\dot{\theta}_2(t)\dot{\theta}_1(t)\\
&\quad \times \sin\theta_2(t) + (m_1\ell_{c_1} + m_2\ell_1)g\cos\theta_1(t) + \phi_2(t),\\
\phi_2(t) &= m_2\ell_{c_2}g\cos[\theta_1(t) + \theta_2(t)].
\end{aligned}
$$

We add Wiener noise, $dW(t)$, and permit the controller to apply a scalar control signal $|a(t)| \leqslant 1$, at the joint between the two links. In the equations of motion, $m_i = 1$ kg, $\ell_i = 1$ m, $\ell_{c_i} = 0.5\ell_i$, and I_i stand for the mass, the length, the length to center of mass, and the moment of inertia of the ith link, $i \in [1, 2]$, respectively. Figure 4 shows the landscape for the original empowerment for selected slices of the phase space. This landscape is more complex than for the single-pendulum. Nonetheless it retains the property that, following the local gradient in the state space directly, one ultimately reaches the state of the maximum empowerment, which is precisely where both links of the pendulum are balanced upright. The vertical position, however, is *a priori* not sufficient to guarantee the balancing since the control only applies torque at the joint linking the pendulum halves. That is, the controller cannot move the pendulum in arbitrary directions through the state space. Surprisingly, this concern notwithstanding, the algorithm still balances the pendulum; cf. Fig. 4.

3. Cart-pole

We have additionally verified that the empowerment maximization also balances an inverted pendulum on a moving cart; cf. Fig. 5. Here the control signal (force) is applied to the cart. Thus the pendulum is now affected only indirectly. The dynamics of this system is

$$
\begin{aligned}
d\ddot{x}(t) &= \frac{m\sin\theta(t)[\ell\dot{\theta}^2(t) + g\cos\theta(t)] + da(t) + dW(t)}{M + m\sin^2\theta(t)},\\
d\ddot{\theta}(t) &= -da(t)\cos\theta(t) - m\ell\dot{\theta}^2(t)\cos\theta(t)\sin\theta(t)\\
&\quad - (M+m)g\sin\theta(t),
\end{aligned}
\tag{14}
$$

where $x(t)$, $\theta(t)$, $m = 1$ kg, $M = 10$ kg, $\ell = 1$ m, g, $|a(t)| \leqslant 1$ are the x coordinate of the center of mass of the cart, the angle of the pole, the pole mass, the cart mass, the pole length, the free fall acceleration, and the force applied to the cart.

III. DISCUSSION

In this study, we focused on a class of intrinsic motivation models that mimic decision-making abilities of biological organisms in various situations without explicit reward signals. We used an information-theoretic formulation in which the controller starts with knowledge of the (stochastic) dynamical equations describing the agent and the environment, and then selects actions that "empower" the agent. That is, the controller improves its ability to affect the system in the future, as measured by the mutual information between the action sequence and the subsequent responses. This leads the system to the most sensitive points in the state space—quite generally, and without relying on the details of the dynamics being controlled—which we showed solves a problem known to be difficult for simple reinforcement learning algorithms: balancing inverted pendula. Depending on which subsets of the past actions and future responses are used to drive the intrinsic motivation, our approach interpolates between the original formulation of empowerment maximization, maximization of the "kicked" version of causal entropic forcing, and maximization of the "controlled" subset of the Lyapunov exponents of the agent-environment pair.

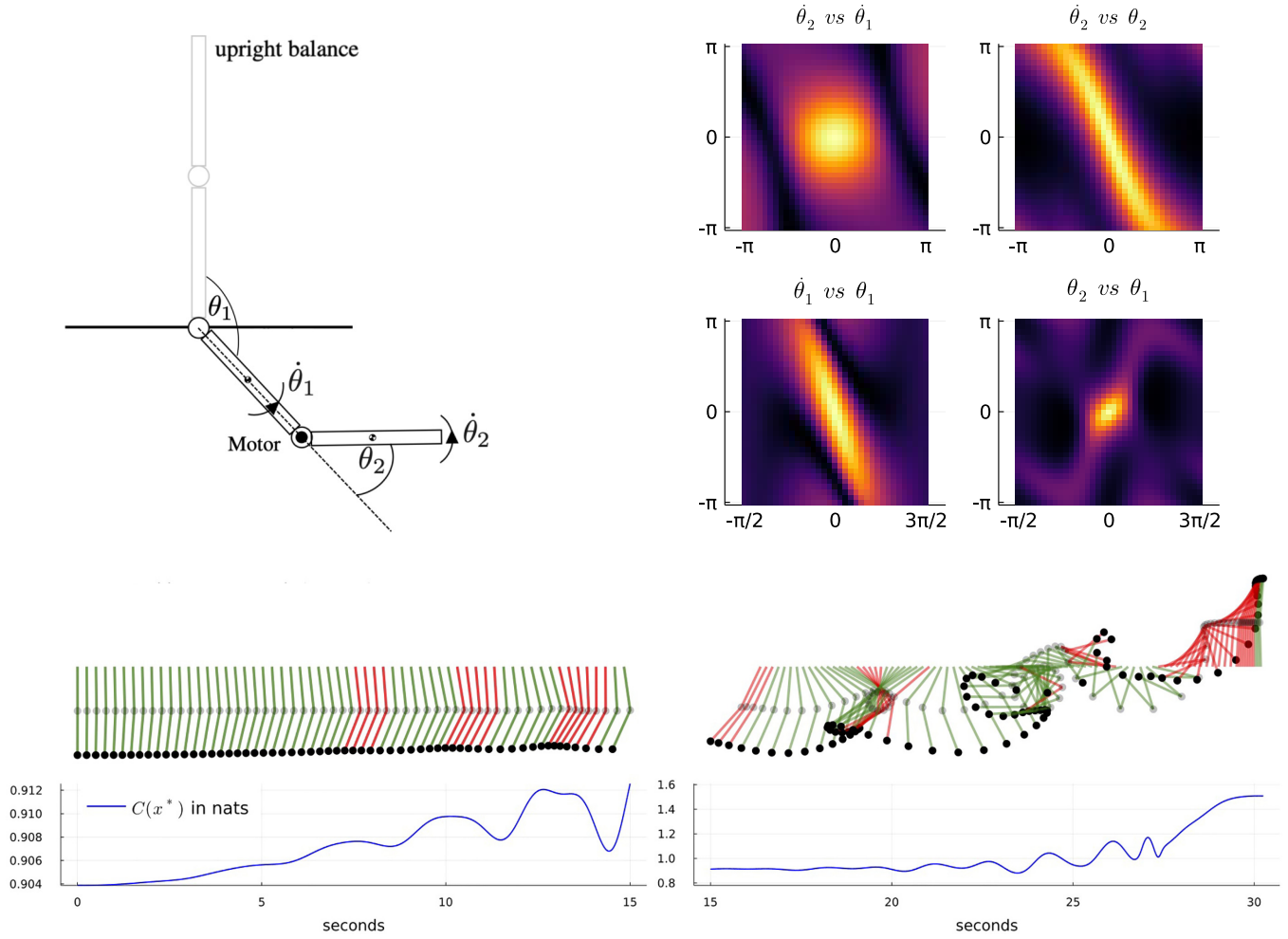

FIG. 4. Top left: Double pendulum with control torque on the joint between the links with dynamics given by (13). Top right: Slices through the empowerment landscape of a double pendulum. Each subplot shows a particular slice in the 4D landscape, when two other coordinates are zero. For example, the plot with axes $\dot{\theta}_2, \dot{\theta}_1$ is shown for $\theta_2 = 0$ rad and $\theta_1 = 0$ rad. Bottom: Traversing the state space of the double pendulum according to (4). The first and the second 15 s are shown with different scale for the instantaneous empowerment. The initial and the final positions are both links down and both links up, respectively. Torque is applied to the middle joint only. When the pendulum is depicted in green (red), it is absorbing (releasing) the energy from (against) the driving force.

This provides insight into which properties of the dynamical system are responsible for the behaviors produced by these different motivation functions.

Notably, there is an essential difference between empowerment maximization and optimal control approaches for the derivation of the optimal action policy. Empowerment-based

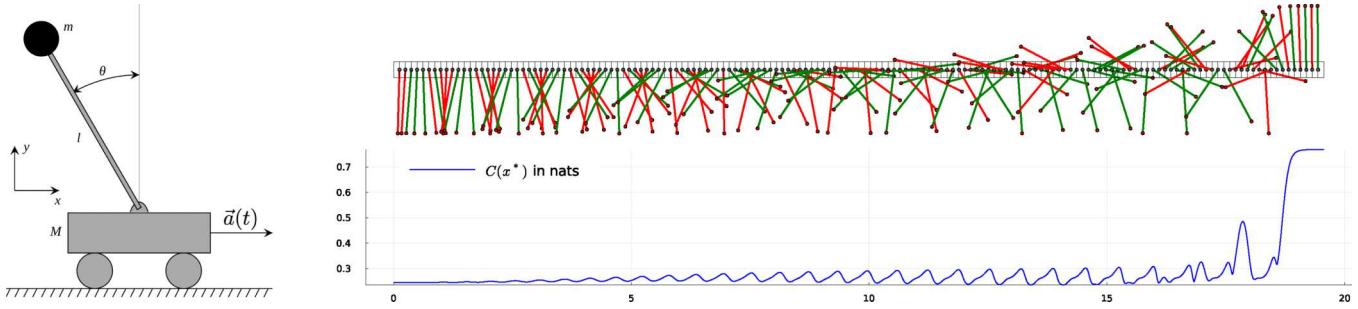

FIG. 5. Left: Cart-Pole system with control force, $\vec{a}(t)$, applied to the cart only, which moves on the rail (or on the edge of a table), allowing the pole to rotate in the x-y plane. Its dynamics is given by (14). Right: Traversing the state space of the pendulum on a cart according to empowerment maximization. The initial and the final state of the pole are down and up, respectively. The horizontal axis is time in seconds, $t \in [0, 20]$ s.

policy is derived from local properties of the dynamics, while optimal control policy is globally derived from an external reward by solving the Hamiltonian-Jacobi-Bellman equation [44]. Understanding a direct connection between these two alternatives requires further investigation.

One big challenge in using information-theoretic quantities is computing them, which can be difficult to do either analytically or from data. Our paper makes a significant contribution to solving this problem in the context of empowerment by providing an explicit algorithm for computing various versions of empowerment, for arbitrary lengths of pasts and futures, using the small noise/small control approximation to the dynamics, while still treating the dynamics as nonlinear. This is often the most interesting regime, modeling weak, power-constrained controllers. We point out that our small perturbation analysis does not assume the dynamics to be linear, and *a priori* it is not clear whether assuming small perturbations around a generic nonlinear solution would render the system tractable, which our algorithm achieves. To establish the robustness of our results, we studied the dependence of the empowerment estimate on the discretization step size, as the latter converges to zero. The estimate converges trivially, Fig. 3, and only depends minimally on the discretization step, including at the critical points of the dynamics.

Crucially, our algorithm is local, so that climbing up the empowerment gradient only requires estimation of the dynamics in the vicinity of the current state of the system. This should be possible in real control applications by using the data directly, possibly with the help of deep neural networks to approximate the relevant dynamical landscapes [46–48]. Therefore, knowing the exact form of the dynamical system, which could be a potential limitation of our approach, is not strictly required. This opens up opportunities for scaling our method to more complex scenarios.

Our work suggests that, in addition to the Lyapunov spectrum, defined via the trajectory divergence in time due to a small *arbitrary* perturbation, one may want to consider the *optimal* Lyapunov spectrum, where the initial perturbation is *optimally* aligned with the controllable directions in the dynamics. We defer a systematic study of optimal Lyapunov spectra to future work. In this context, one could also ask if maximization of empowerment and application of control in optimal directions might result in instabilities in the system's dynamics. Since empowerment optimization is usually applied in the case of limited resources, such as power, we do not expect such runaway solutions. However, formal analysis of this is left for the future.

While stabilization of pendula, including the double pendulum, are classic test cases for control-theoretic algorithms, our empowerment based approach needs substantial additional developments to become a general control strategy. First, it leads an agent only to very specific points in the state space, which optimize the sensitivity to control, and hence potentiate future actions. Second, the specific algorithm we used, greedy empowerment, is unlikely to result in an ability to control systems as complex as humanoid robots. Solution to both problems lies in combining empowerment optimization with problem-specific goals and with explicitly learning the underlying dynamical system in an RL-style model. We anticipate that empowerment maximization will be faster and more reliable within the RL paradigm, and it will dominate early steps of control strategies, effectively endowing RL approaches with exploratory possibilities not directly related to the eventual goal. In its turn, achieving specific RL goals will be easier from such high empowerment regions at later steps of control.

A potential extension of our analysis relates to social interactions. Interacting agents have their own intrinsic motivations and affect each other's ability to achieve their goals. Understanding how multiple agents interact, each trying to empower itself in the presence of others, and whether and when this leads to cooperation or conflict is a promising area for future research. Crucially, the ability to affect someone else's empowerment may provide insight into what distinguishes social interactions from purely physical interactions among nearby individuals.

ACKNOWLEDGMENTS

S.T. was supported in part by the NSF Grant No. 2246221 and Pazy Foundation ID 195-2020. I.N. was supported in part by the Simons Foundation Investigator award, the Simons-Emory Consortium on Motor Control, and NIH Grant No. 2R01NS084844. D.P. acknowledges partial support by the EC H2020-641321 socSMCs FET Proactive project and the Pazy Foundation ID 195-2020.

[1] P.-Y. Oudeyer and F. Kaplan, What is intrinsic motivation? A typology of computational approaches, Front. Neurorob. **1**, 6 (2007).

[2] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA, 2018).

[3] K. Doya, Reinforcement learning in continuous time and space, Neural Comput. **12**, 219 (2000).

[4] S. Mohamed and D. J. Rezende, Variational information maximisation for intrinsically motivated reinforcement learning, Adv. Neural Inf. Proc. Syst. **2**, 2125 (2015).

[5] K. Gregor, D. J. Rezende, and D. Wierstra, Variational intrinsic control, arXiv:1611.07507.

[6] K. Baumli, D. Warde-Farley, S. Hansen, and V. Mnih, Relative variational intrinsic control, Proc. AAAI Conf. Artificial Intell. **35**, 6732 (2021).

[7] T. Kwon, Variational intrinsic control revisited, in *International Conference on Learning Representations (ICLR)*, Vol. 5 (2021).

[8] A. Sharma, S. Gu, S. Levine, V. Kumar, and K. Hausman, Dynamics-aware unsupervised discovery of skills, arXiv:1907.01657.

[9] A. Sharma, M. Ahn, S. Levine, V. Kumar, K. Hausman, and S. Gu, Emergent real-world robotic skills via unsupervised off-policy reinforcement learning, RSS, Robotics: Science and Systems (2020).

[10] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, Diversity is all you need: Learning skills without a reward function, in *International Conference on Learning Representations (ICLR)* (2018).

[11] R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel, Vime: Variational information maximizing exploration, Adv. Neural Inf. Proc. Syst. **29**, 1117 (2016).

[12] J. Achiam, H. Edwards, D. Amodei, and P. Abbeel, Variational option discovery algorithms, arXiv:1807.10299.

[13] J. Choi, A. Sharma, H. Lee, S. Levine, and S. S. Gu, Variational empowerment as representation learning for goal-conditioned reinforcement learning, in *International Conference on Machine Learning* (2021), pp. 1953–1963.

[14] C. Salge, C. Glackin, and D. Polani, Empowerment–An introduction, in *Guided Self-Organization: Inception* (Springer, 2014), pp. 67–114.

[15] A. S. Klyubin, D. Polani, and C. L. Nehaniv, Empowerment: A universal agent-centric measure of control, in *IEEE Congress on Evolutionary Computation* (IEEE, Piscataway, NJ, 2005), pp. 128–135.

[16] T. Jung, D. Polani, and P. Stone, Empowerment for continuous agent—environment systems, Adapt. Behav. **19**, 16 (2011).

[17] A. S. Klyubin, D. Polani, and C. L. Nehaniv, Keep your options open: An information-based driving principle for sensorimotor systems, PLoS ONE **3**, e4018 (2008).

[18] R. Zhao, P. Abbeel, and S. Tiomkin, Efficient empowerment estimation for unsupervised stabilization, in *International Conference Learning Representations* (2020).

[19] R. Zhao, S. Tiomkin, and P. Abbeel, Dynamical system embedding for efficient intrinsically motivated artificial agents, in *Advances in Neural Information Processing Systems (NeurIPS), DeepRL* (2019).

[20] A. D. Wissner-Gross and C. E. Freer, Causal entropic forces, Phys. Rev. Lett. **110**, 168702 (2013).

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, 2012).

[22] A. S. Klyubin, D. Polani, and C. L. Nehaniv, Empowerment: A universal agent-centric measure of control, in *2005 IEEE Congress on Evolutionary Computation* (IEEE, Piscataway, NJ, 2005), Vol. 1, pp. 128–135.

[23] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, Diversity is all you need: Learning skills without a reward function, in *International Conference Learning Representations* (2019).

[24] Y. Du, S. Tiomkin, E. Kiciman, D. Polani, P. Abbeel, and A. D. Dragan, AvE: Assistance via empowerment, in *Neural Information Processing Systems, NeurIPS 2020* (2020).

[25] C. Salge and D. Polani, Empowerment as replacement for the three laws of robotics, Front. Rob. AI **4**, 25 (2017).

[26] W. Bialek, I. Nemenman, and N. Tishby, Predictability, complexity, and learning, Neural Comput. **13**, 2409 (2001).

[27] C. Holmes and I. Nemenman, Estimation of mutual information for real-valued data with error bars and controlled bias, Phys. Rev. E **100**, 022404 (2019).

[28] R. Zhao, K. Lu, P. Abbeel, and S. Tiomkin, Efficient empowerment estimation for unsupervised stabilization, in *International Conference on Learning Representations* (2021).

[29] A. Y. Ng, D. Harada, and S. Russell, Policy invariance under reward transformations: Theory and application to reward shaping.

[30] P. Dayan and G. E. Hinton, Feudal reinforcement learning, Adv. Neural Inf. Proc. Syst. 271 (1993).

[31] T. D. Kulkarni, K. Narasimhan, A. Saeedi, and J. Tenenbaum, Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation, Adv. Neural Inf. Proc. Syst., NeurIPS **29**, 3682 (2016).

[32] T. Anthony, D. Polani, and C. L. Nehaniv, General self-motivation and strategy identification: Case studies based on sokoban and pac-man, IEEE Trans. Comput. Intell. AI Games **6**, 1 (2014).

[33] Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros, Large-scale study of curiosity-driven learning, in *ICLR* (2019).

[34] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, Curiosity-driven exploration by self-supervised prediction, in *ICML* (2017).

[35] M. Karl, M. Soelch, P. Becker-Ehmck, D. Benbouzid, P. van der Smagt, and J. Bayer, Unsupervised real-time control through variational empowerment (2018).

[36] C. Salge, C. Glackin, and D. Polani, Approximation of empowerment in the continuous domain, Adv. Complex Syst. **16**, 1250079 (2013).

[37] C. Salge, C. Glackin, and D. Polani, Empowerment–An introduction, in *Guided Self-Organization: Inception* (Springer, 2014), pp. 67–114.

[38] N. C. Volpi, D. De Palma, D. Polani, and G. Indiveri, Computation of empowerment for an autonomous underwater vehicle, IFAC-PapersOnLine **49**, 81 (2016).

[39] H. J. Charlesworth and M. S. Turner, Intrinsically motivated collective motion, Proc. Natl. Acad. Sci. USA **116**, 15362 (2019).

[40] The perturbation gain, $\bar{h}$, allows for a more general representation than in Eq. (1). However, we stick to Eq. (1) by restricting $\bar{h}$ to Δt.

[41] https://github.com/stastio/im.git.

[42] J. Schmidhuber, Formal theory of creativity, fun, and intrinsic motivation (1990–2010), IEEE Trans. Auton. Ment. Dev. **2**, 230 (2010).

[43] H. D. Abarbanel, R. Brown, and M. B. Kennel, Local lyapunov exponents computed from observed data, J. Nonlin. Sci. **2**, 343 (1992).

[44] L. C. Evans, An introduction to mathematical optimal control theory, version 0.2 (1983).

[45] A. D. Kuo, The six determinants of gait and the inverted pendulum analogy: A dynamic walking perspective, Hum. Mov. Sci. **26**, 617 (2007).

[46] B. C. Daniels and I. Nemenman, Automated adaptive inference of phenomenological dynamical models, Nat. Commun. **6**, 8133 (2015).

[47] S. L. Brunton, J. L. Proctor, and J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Natl. Acad. Sci. USA **113**, 3932 (2016).

[48] B. Chen, K. Huang, S. Raghupathi, I. Chandratreya, Q. Du, and H. Lipson, Automated discovery of fundamental variables hidden in experimental data, Nat. Comput. Sci. **2**, 433 (2022).