Seeking entropy: Complex behavior from intrinsic motivation to occupy action-state path space

Jorge Ramírez-Ruiz[1], Dmytro Grytskyy[1] and Rubén Moreno-Bote[1,2]

[1]Center for Brain and Cognition, and Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain
[2]Serra Húnter Fellow Programme, Universitat Pompeu Fabra, Barcelona, Spain

Motivation

Most theories of behavior posit that agents tend to maximize some form of reward or utility. However, very often animals move with curiosity and seem to be motivated in a reward-free manner. We propose the Maximum Path Occupancy (MPO) principle, a reward-free objective: Maximizing occupancy of future paths of actions and states. According to this view, rewards are the means to occupy path space, not the goal per se; goal-directedness simply emerges as a rational way of searching for resources so that movement, understood amply, never ends.

Background

Our framework builds over an extensive literature on entropy-regularized reinforcement learning (RL), where, however, rewards serve still as the major drive of behavior. In intrinsic motivation approaches, exploration is promoted through explicit information-seeking objectives in a learning setting, such as surprise minimization and novelty seeking. Our MPO principle differs from these approaches in that agents will still move and occupy path space even if there is nothing to learn. Other entropic reward-free approaches focus instead on the coverage problem, maximizing the steady-state state entropy, and they do so typically to generate better policies in the exploitation phase when a well-defined task is to be solved. Finally, there is a class of reward-free objectives known as empowerment that focus instead on generating policies that maximize the agent's predictive power, and therefore generate qualitatively different behaviors in stochastic environments than our maximum path occupancy principle.

Main idea

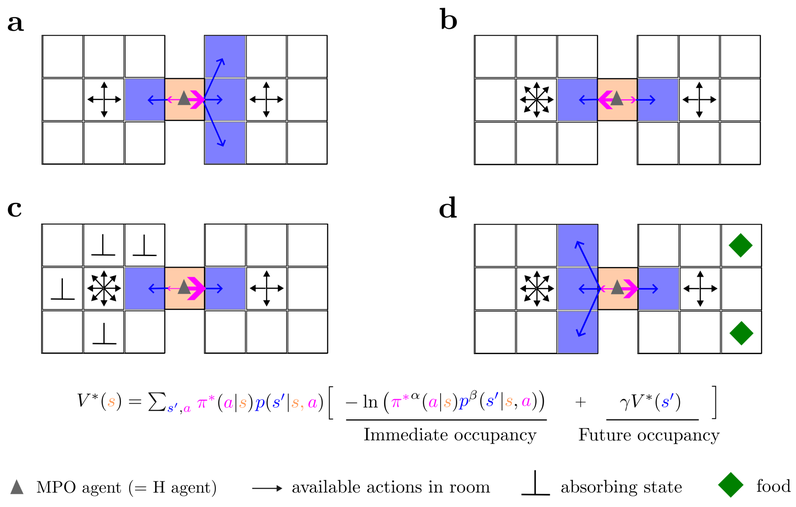

Figure 1: Entropy-seeking agents (H agents) achieve maximum path occupancy (MPO). (a) An H agent (grey triangle) in the middle of two rooms has the choice between going left or right. When the number of actions (black arrows) in each room are the same, the agent prefers going to the room with more state transitions (blue arrows indicate random transitions after choosing moving right or moving left actions, and pink arrow width indicates the probabilities of those actions). (b) When the states transitions are the same in the rooms, the H agent prefers the room with more available actions. (c) If there are many absorbing states in the room where many actions are available, the H agent avoids it. (d) Even if there are action and state-transition incentives (in the left room), an H agent might prefer a region of state space where it can reliably get food (right room), ensuring occupancy of future action-state paths.

Contributions

Focusing on finite state-action Markov Decision Processes (MDP),

1) We find that action-state path entropy is the only measure consistent with additivity and other intuitive properties of expected future action-state path occupancy.

2) We provide analytical expressions that relate the optimal policy and state-value function, and prove convergence of our value iteration algorithm.

3) Using discrete and continuous state tasks, we show that complex behaviors such as 'dancing', hide-and-seek and a basic form of altruistic behavior naturally result from the intrinsic motivation to occupy path space.

Theory

We study finite state and action MDPs. A policy $\pi(\cdot|s)$ defines a probability for each action at every state s in the MDP. At every state s, and given action a, the environment produces a distribution over successor states s' acording to $p(s'|s,a)$. Starting at $t = 0$ in state s_0 and following policy π, an agent experiences a path in action-state space $\tau \equiv (s_0, a_0, s_1, ..., a_t, s_{t+1}, ...)$.
Then, we find the intrinsic return from four properties of path occupancy, and find it to be

$$R(\tau) = -\sum_{t=0}^{\infty} \gamma^t \ln\left(\pi^{\alpha}(a_t|s_t)p^{\beta}(s_{t+1}|s_t,a_t)\right).$$

The value function of a given policy at a given state $V_\pi(s)$ is defined as the expected return when following the policy π. The optimal value function $V^*(s)$ follows the optimal Bellman equation, whose solution we find to be

$$V^*(s) = \alpha \ln Z(s) = \alpha \ln\left[\sum_a \exp\left(\alpha^{-1}\beta\mathcal{H}(S'|s,a) + \alpha^{-1}\gamma \sum_{s'} p(s'|s,a)V^*(s')\right)\right].$$

And the optimal policy uses this optimal value,

$$\pi^*(a|s) = \frac{1}{Z(s)} \exp\left(\alpha^{-1}\beta\mathcal{H}(S'|s,a) + \alpha^{-1}\gamma \sum_{s'} p(s'|s,a)V^*(s')\right).$$

Four-room grid world

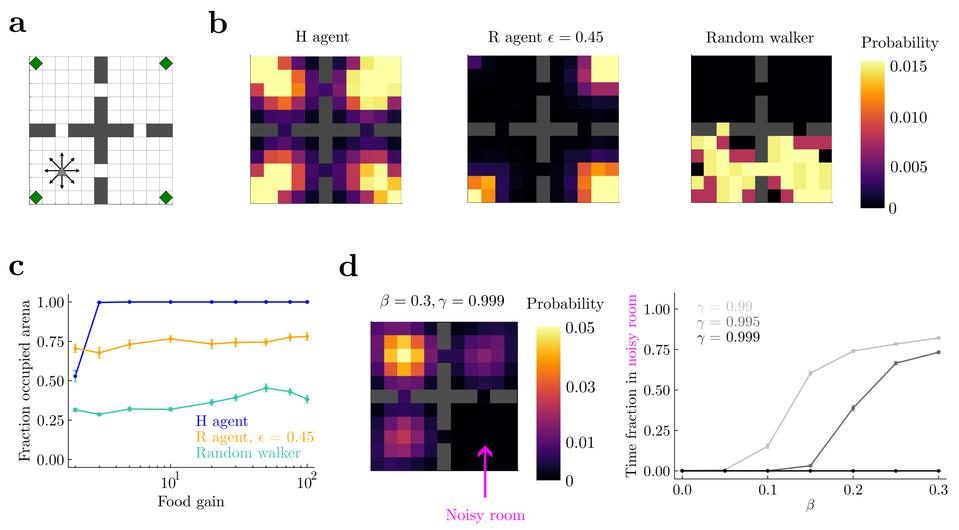

Figure 2: Maximizing future path occupancy leads to high occupancy of physical space. (a) Grid-world arena. The agents have nine available actions (arrows, and staying still) when alive (internal energy larger than zero) and away from walls. There are four rooms, each with a small food source in a corner (green diamonds). (b) Probability of visited spatial states for an entropy-seeking (H) agent, an ϵ-greedy reward (R) agent that survives as long as the H agent, and a random walker. Food gain = 10 units, maximum reservoir energy = 100, episodes of 5×10^4 time steps, and $(\alpha, \beta) = (1, 0)$ for the H agent. All agents are initialized in the middle of the lower left room. (c) Fraction of locations of the arena visited at least once per episode as a function of food gain. Error bars correspond to s.e.m over 50 episodes. (d) Noisy room problem. The bottom right room of the arena was noisy, such that agents in this room jump randomly to neighboring locations regardless of their actions. Food gain equals maximum reservoir energy = 100. Histogram of visited locations for an episode as long as in (b) for a H agent with $\beta = 0.3$ (left) and time fraction spent in the noisy room (right) show that H agents with $\beta > 0$ can either be attracted to the room or repelled depending on γ.

Predator-Prey

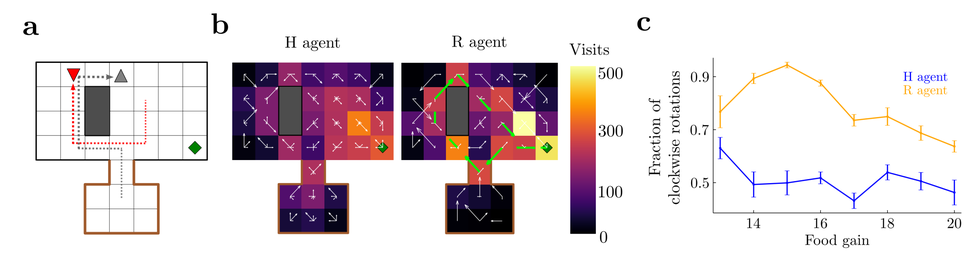

Figure 3: Complex hide-and-seek and escaping strategies in a prey-predator example. (a) Grid-world arena. The agent has nine available actions when alive and far from walls. There is a small food source in a corner (green diamond). A predator (red, down triangle) is attracted to the agent (gray, up triangle), such that when they are at the same location, the agent dies. The predator cannot enter the locations surrounded by the purple border. Arrows show a clockwise trajectory. (b) Histogram of visited spatial states across episodes for the H and R agents. The vector field at each location indicates probability of transition at each location. Green arrows on R agent show major motion directions associated with the its dominant clockwise rotation. (c) Fraction of clockwise rotations (as in panel (a)) to total rotations as a function of food gain, averaged over epochs of 500 timesteps. Error bars are s.e.m.

Cartpole

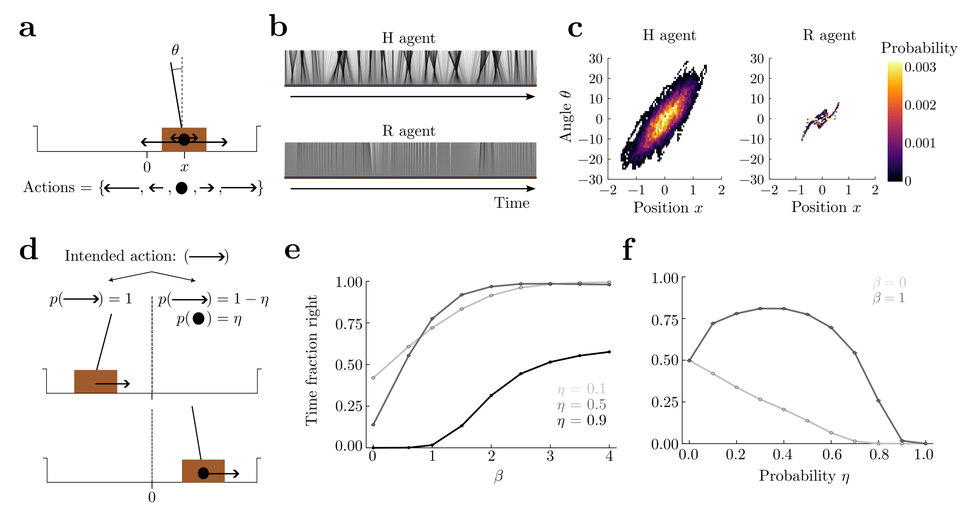

Figure 4: Dancing of an entropy-seeking cartpole. (a) The cart (brown rectangle) has a pole attached. The cartpole reaches an absorbing state if the magnitude of the angle θ exceeds 36 deg or its position reaches the borders. There are 5 available actions when alive: a big and a small force to either side (arrows on cartpole), and doing nothing (full circle). (b) Time-shifted snapshots of the pole in the reference frame of the cart as a function of time for the H (top) and R (bottom) agents. (c) Position and angle occupation for a 2×10^5 time step episode. (d) Here, the right half of the arena is stochastic, while the left remains deterministic. In the stochastic half, the intended state transition due to an applied action (force) succeeds with probability $1 - \eta$ (and thus zero force is applied with probability η). (e) Fraction of time spent on the right half of the arena increases as a function of β, regardless of the failure probability η. (f) The fraction has a non-monotonic behavior as a function of η when state entropy is important for the agent ($\beta = 1$), highlighting a stochastic resonance behavior. When the agents do not seek state entropy ($\beta = 0$) the fraction of time spent by the agent on the right decreases with the failure probability, and thus they avoid the stochastic right side. $\gamma = 0.99$ for panels (e,f).

Conclusions

- Often, the success of agents in nature is not measured by the amount of reward obtained, but by their ability to expand in state space and perform complex behaviors. Here we have proposed that a major goal of intelligence is to 'occupy path space'.
- We show that the intuitive notion of path occupancy is captured by future action-state path entropy, and we have proposed that behavior is driven by the maximization of this sole intrinsic goal –the MPO principle.
- In four examples we have shown that the MPO principle, along with the agent's constraints and dynamics, leads to complex behaviors that are not observed in other simple reward maximizing agents.
- Several steps remain to fully characterize MPO agents, which includes learning and partially observable environments.