The maximum occupancy principle: Complex behavior from intrinsic motivation to occupy action-state path space

Jorge Ramírez-Ruiz[1,3], Dmytro Grytskyy[1], Chiara Mastrogiuseppe[1], Yamen Habib[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
[3]Département de Neurosciences, Faculté de Medicine, Université de Montréal, Canada

arXiv

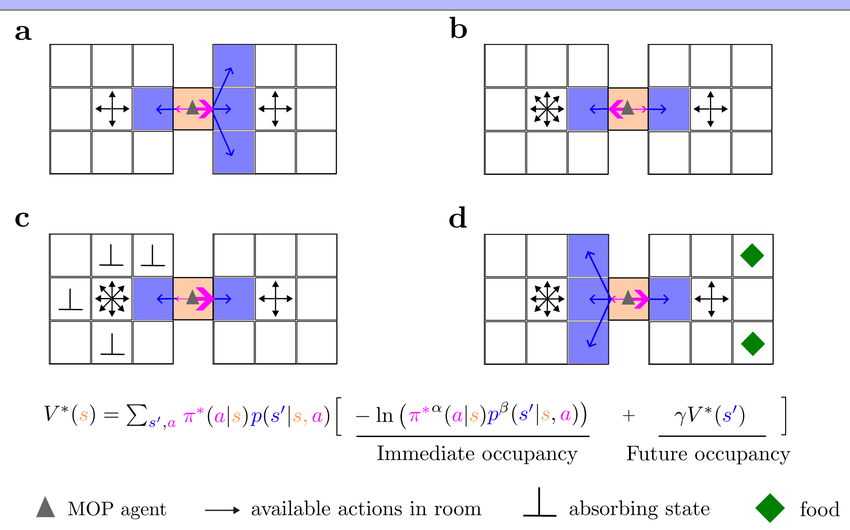
Motivation

- Most theories of behavior posit that agents tend to maximize some form of extrinsic reward or utility.
- However, very often animals move with curiosity and seem to be motivated in a reward-free manner.
- We propose the Maximum Occupancy Principle (MOP), a reward-free objective: Maximizing occupancy of future paths of actions and states.
- 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.

Related work

- In entropy-regularized reinforcement learning (RL), extrinsic rewards are their ultimate driver of goal-directed behavior [1,2].
- Surprise or prediction error minimization [3,4] and novelty seeking [5] are intrinsic motivations that change as a state of learning. MOP always pushes agents to occupy path space.
- Other entropic reward-free approaches focus on the coverage problem [6,7], maximizing the steady-state state entropy.
- Empowerment maximizes mutual information between states and actions [8] and does not push agents to actually fulfill a diverse and predictable set of states.

Main idea



MOP agents maximize path occupancy. (a) A MOP agent 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. (b) When the states transitions are the same in the rooms, the MOP agent prefers the room with more available actions. (c) The MOP agent avoids absorbing states, even when there are many immediate actions. (d) Even if there are action and state-transition incentives (in the left room), a MOP agent might prefer a region of state space where it can reliably get food (right room), ensuring occupancy of future action-state paths.

Theory

A policy $\pi(a|s)$ defines a probability for each action a at every state s in an MDP, and $p(s'|s,a)$ is the probability of transition to state s'. The intrinsic return from following an action-state path $\tau \equiv (s_0, a_0, s_1, ..., a_t, s_{t+1}, ...)$ comes from four *desiderata* for path occupancy, and turns out to be

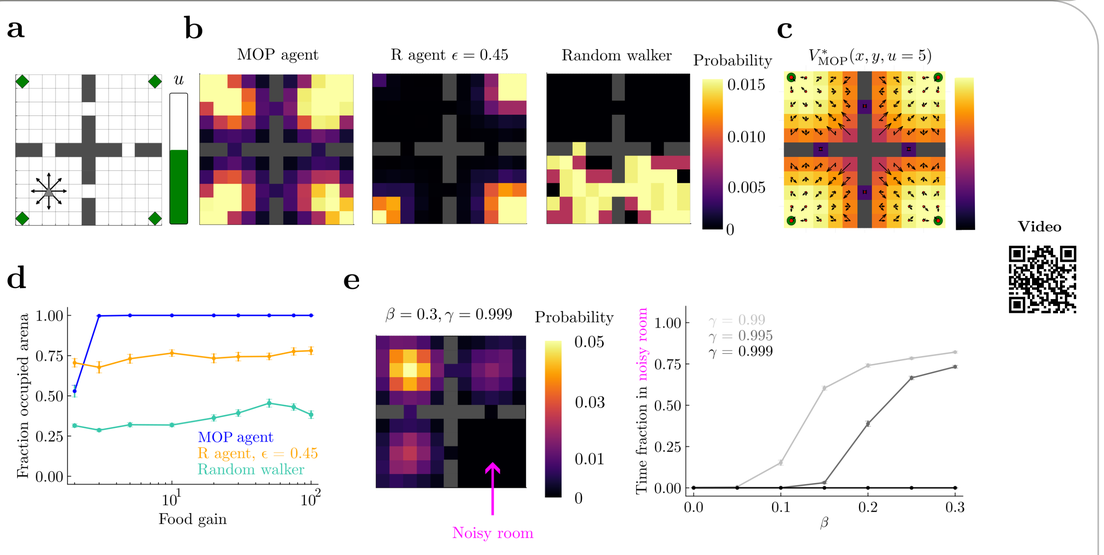
$$R(\tau) = -\sum_{t=0}^{\infty} \gamma^t \ln\left(\pi^{\alpha}(a_t|s_t)p^{\beta}(s_{t+1}|s_t,a_t)\right).$$

The optimal value function $V^*(s)$ follows the optimal Bellman equation, with the following solution,

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

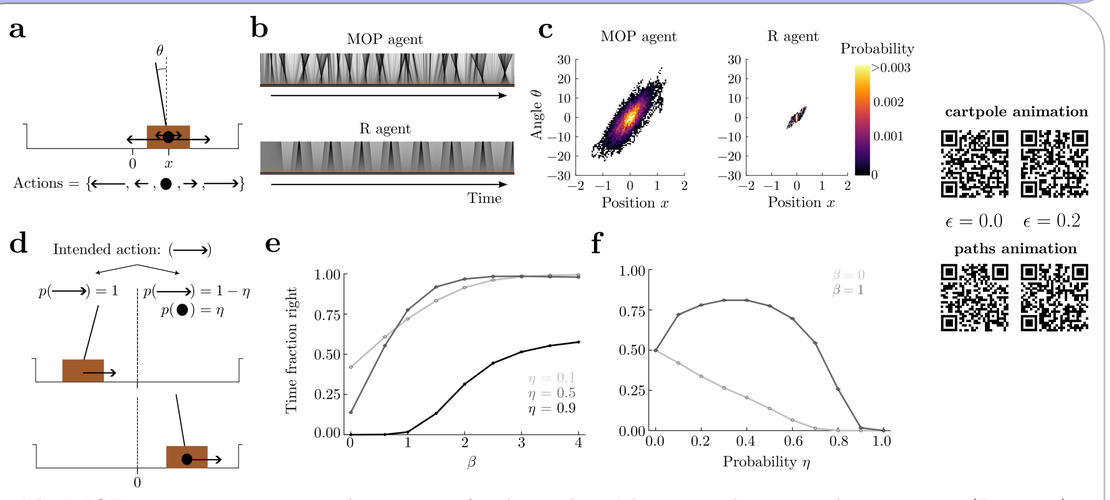
With optimal policy $\pi^*(a|s) = \frac{1}{Z(s)} \exp\left(\alpha^{-1}\beta\mathcal{H}(S'|s,a) + \alpha^{-1}\gamma \sum_{s'} p(s'|s,a)V^*(s')\right)$.

MOP agents cover state space when constraints allow



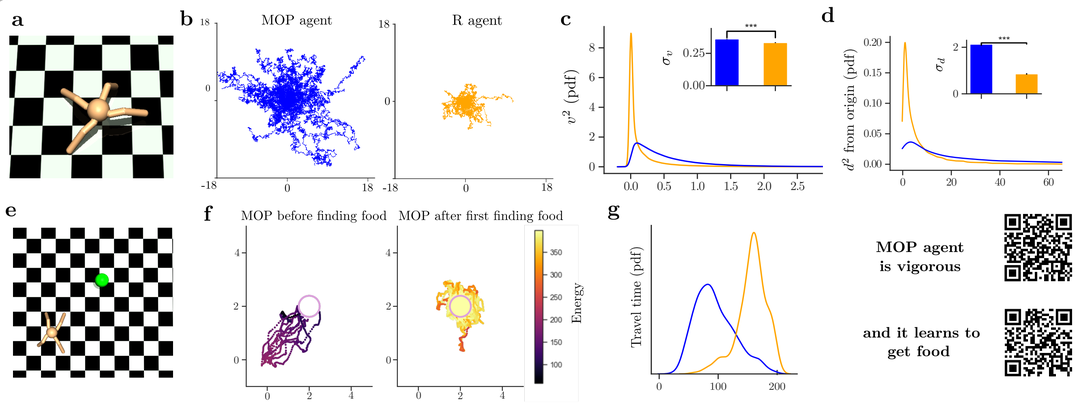
- The MOP agent enters the four rooms in the long term by storing energy u in its reservoir (b). ϵ-greedy reward maximizers (R agent) linger over the food sources.
- The MOP agent is also capable of deterministic behavior (c, only one action considered at corners).
- Getting stuck in stochastic regions where energy cannot be predictably obtained is avoided by sufficiently long-sighted MOP agents (e).

Stochastic dance in a cartpole



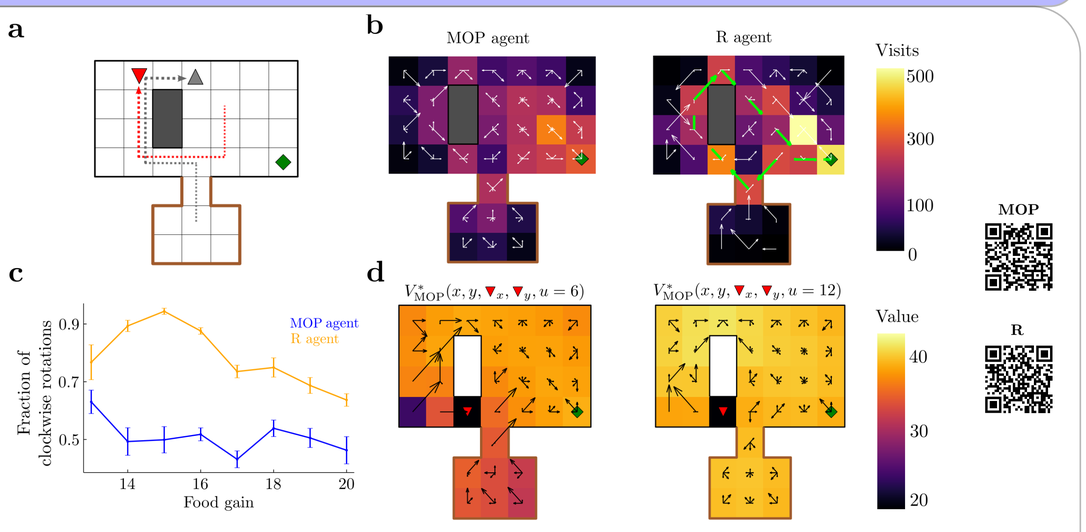
- The MOP agent occupies a wide variety of pole angles. The ϵ-greedy survival maximizer (R agent) balances the pole with very little behavioral variability.
- The MOP agent exhibits the most appropriate sort of variability for a given average lifetime.

Scaling to high dimensions



- MOP can be applied to control a high-dimensional quadruped. The MOP agent avoids falling (top row) and starving (bottom), while generating vigorous, variable behavior.

Diverse behavioral repertoires in a predator–prey scenario



- A MOP prey trades off escaping from a predator and getting food, can be deterministic, and it displays more diverse escaping strategies than an ϵ-greedy survival maximizer (R agent).

Conclusions

- We have proposed that a major feature of behavior is to occupy path space, captured by future action-state path entropy.
- We have shown that MOP, along with the agent's constraints and dynamics, leads to complex, *goal-directed* behaviors without extrinsic rewards.
- Several steps remain to fully characterize MOP agents, which includes the effect of online learning and partially observable environments.

References

[1] Todorov, E. PNAS, 2009
[2] Ziebart, B., 2010
[3] Friston, K. *et al.* PLOS ONE, 2009
[4] Pathak, D. *et al.* ICML, 2017
[5] Bellemare, M. *et al.* NeurIPS, 2016
[6] Hazan, E. *et al.* ICML, 2019
[7] Amin, S. *et al.* Methods in RL, 2021
[8] Klyubin, A. *et al.* PLOS ONE, 2008