CNeuro 2024, Lectures
Soft RL and Maximum Occupancy Principle (MOP)

Rubén

May 1, 2024

Lecture 1: Introduction to Soft Reinforcement Learning

Most theories of behavior posit that humans and other animals maximize some form of extrinsic reward. Reinforcement Learning (RL) has became a popular tool to model behavior in natural agents based on the principle of reward maximization. Not only is RL guiding research in neuroscience, but it is an instrumental tool to solve hard control problems in artificial intelligence. In the first lecture, I will introduce the basic notions of RL: reward, value function, Q-value, Bellman equation and temporal-difference (TD) error. Then, I will describe different ways the exploration-exploration dilemma is addressed in RL, a dilemma that aims at optimally balancing maximizing immediate rewards and exploring to improve learning and future policies. Then, I will describe how to improve learning in RL by using entropy regularization to the reward objective (soft RL) [Todorov, 2009, Peters et al., 2010], which smooths action selection and leads to efficient exploration in some cases. Next, I will describe how researchers have started to challenge the idea of reward-maximization by positing that intrinsic motivations largely drive behavior, such as empowerment [Klyubin et al., 2005] and causal entropic forces [Wissner-Gross and Freer, 2013]. After commenting on the limitations of those approaches, I will briefly introduce the maximum occupancy principle (MOP), based on the hypothesis that natural agents maximize future action-state path occupancy [Ramírez-Ruiz et al., 2022, Moreno-Bote and Ramirez-Ruiz, 2023]. Next lecture will take MOP as starting point.

Lecture 2: Behavior without rewards: the maximum occupancy principle (MOP)

Intrinsic motivation generates behaviors that do not necessarily lead to immediate reward, but help exploration and learning. I will show that agents having the sole goal of maximizing occupancy of future actions and states, that is, moving and exploring on the long term, are capable of complex, goal-directed behavior [Ramírez-Ruiz et al., 2022, Moreno-Bote and Ramirez-Ruiz, 2023]. This maximum occupancy principle (MOP) is related but different to other forms of reward-free model of behavior, such as empowerment and free energy. Action-state path entropy is the only measure consistent with additivity and other intuitive properties of expected future action-state path occupancy. Using discrete and continuous state tasks, I will finally show that 'dancing', hide-and-seek and a basic form of altruistic behavior naturally result from entropy seeking without external rewards. MOP agents can learn by themselves to survive by flexibly fabricating their own goals.

References

[Klyubin et al., 2005] Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2005). Empowerment: A universal agent-centric measure of control. In *2005 ieee congress on evolutionary computation*, volume 1, pages 128–135. IEEE.

[Moreno-Bote and Ramirez-Ruiz, 2023] Moreno-Bote, R. and Ramirez-Ruiz, J. (2023). Empowerment, free energy principle and maximum occupancy principle compared. In *NeurIPS 2023 workshop: Information-Theoretic Principles in Cognitive Systems*.

[Peters et al., 2010] Peters, J., Mulling, K., and Altun, Y. (2010). Relative entropy policy search. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

[Ramírez-Ruiz et al., 2022] Ramírez-Ruiz, J., Grytskyy, D., and Moreno-Bote, R. (2022). Seeking entropy: complex behavior from intrinsic motivation to occupy action-state path space. *arXiv preprint arXiv:2205.10316*.

[Todorov, 2009] Todorov, E. (2009). Efficient computation of optimal actions. *Proceedings of the national academy of sciences*, 106(28):11478–11483.

[Wissner-Gross and Freer, 2013] Wissner-Gross, A. D. and Freer, C. E. (2013). Causal entropic forces. *Physical review letters*, 110(16):168702.