The brain ascends reward gradients to learn in continuous motor spaces

Our brains learn complex motor skills, such as serving a tennis ball or tying shoelaces, with great efficiency compared to artificial agents. Such skills are mastered largely through a process of trial-and-error reinforcement learning (RL). Much of our understanding of how the brain performs RL comes from studies of decision-making tasks, like the multi-armed bandit task, that require subjects to choose between a few discrete options. However, motor learning requires the brain to learn in high-dimensional, continuous action spaces, where exhaustive exploration strategies are ineffective. How the brain solves this exploration challenge posed by motor learning is unknown. Here, we aim to understand the principles of efficient learning in continuous action spaces by uncovering the brain's strategies for solving complex motor tasks.

We distinguish between two classes of reinforcement learning algorithms: (a) value-based methods which derive an optimal policy by first learning a value function over the action space through global sampling, and (b) policy-based methods which directly update the parameters of an action policy using stochastic gradient ascent. Behavioural studies [1] have found that humans learn faster on reward landscapes with steeper gradients, and this has been proposed as evidence for policy-gradient learning. To test this, we simulated steep and shallow reward landscapes (Fig. 1) and measured the learning rates of value-based and policy-gradient algorithms in these environments. Our results indicate that value-based methods like Q-learning [2] and Thompson sampling learn faster on steeper landscapes as compared to shallow ones, similar to policy-gradient algorithms [3]. We conclude that modulation of learning speed by the steepness of a reward gradient cannot identify the RL strategy employed by a learning agent.

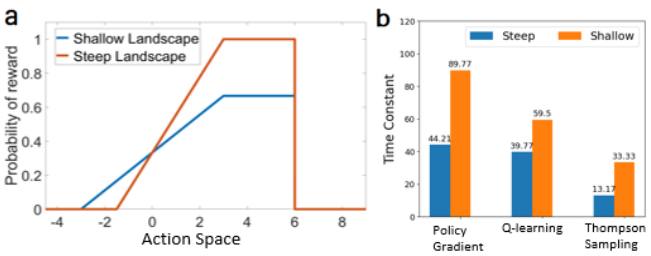

Figure 1: (a) Steep and shallow reward gradients. **(b)** Bar plot of time constants derived from fits to the learning curves of the three algorithms in different environments. Lower time constant indicates faster learning rate.

Instead, we devised a novel approach to infer the brain's trial-by-trial learning algorithm by analysing massive behavioural datasets acquired in a motor learning task. The principle here is that variability inherent to RL can be averaged out to reveal learning-related signals on the scale of individual trials. We used a large dataset spanning 3 million trials acquired from rats learning to adapt the angle at which they pressed a 2-D joystick to maximize reward [4]. We found that the rat's trial-by-trial behaviour was better fit by policy-gradient RL algorithms than value-based methods (Fig. 2). We also found that rats' future actions were biased in the direction of the reward gradient estimated from previous trials, consistent with the predictions of policy-gradient algorithms.

We conclude that the brain employs gradient-ascent algorithms that rely on local exploration when learning a complex motor skill. This strategy is optimized for learning in high-dimensional continuous action spaces. Our results have relevance for fields like robotics which also face the challenge of motor learning.

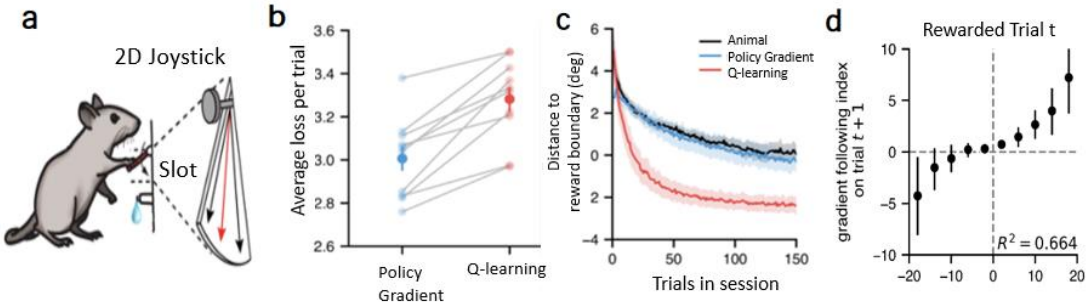

Figure 2: (a) Schematic of the motor learning task. **(b)** Average negative log likelihood loss per trial for optimal model fits of policy gradient and Q-learning algorithms to behavioural data (blue and red dots represent mean and error bars represent standard error of the mean, lighter dots represent individual animals) **(c)** Average learning curves indicating deviation from reward boundary for behavioural data and simulated algorithms. **(d)** Gradient following index on trial *t*+1 as a function of the gradient estimate on trial *t* for rewarded trials (dots represent means and error bars represent standard error of the mean, n = 10 rats)

[1] Cashaback et al 2019 https://doi.org/10.1371/journal.pcbi.1006839
[2] Reinforcement Learning: An Introduction (MIT Press, 2018)
[3] Degris et al 2012 https://doi.org/10.1109/ACC.2012.6315022
[4] Dhawale et al 2019 https://doi.org/10.1016/j.cub.2019.08.052