Model-based reinforcement learning with spiking neurons

Johannes Friedrich, Máté Lengyel

Abstract

Behavioural and neuroscientific data on reward-based decision making point to a fundamental distinction between habitual and goal-directed action selection. An increasingly explicit set of neuroscientific ideas has been established for habit formation, whereas goal-directed control has only recently started to attract researchers' attention. While using functional magnetic resonance imaging to address the involvement of brain areas in goal-directed control abounds, ideas on its algorithmic and neural underpinning are scarce. Here we present the first spiking neural network implementation for goal-directed control that selects actions optimally in the sense of maximising expected cumulative reward. Finding the optimal solution is commonly considered a difficult optimisation problem, yet we show that it can be accomplished by a remarkably simple neural network over a time scale of a hundred milliseconds. We provide a theoretical proof for the convergence of the neural dynamics to the optimal value function. We extend the applicability of the model from discrete to continuous state spaces using linear function approximation. Further, we also show how the environmental model can be learned using a local synaptic plasticity rule in the same network. After establishing the performance of the model on various benchmark tasks from the machine learning literature, we present a set of simulations reproducing behavioural as well as neurophysiological experimental data on tasks ranging from simple binary choice to sequential decision making. We also discuss the relationship between the proposed framework and other models of decision making.

Problem formulation. The generic goal of an agent is to choose actions optimally to maximise expected total future reward (= value) in a Markov Decision Process (MDP). The Bellman optimality equation establishes a recursive relationship for the values obtainable with the optimal policy, which selects the action with the highest value:

$$V^*(s) = \max_a \left(\langle r(s,a) \rangle + \gamma \sum_{s'} P(s'|s,a)\, V^*(s') \right)$$

Here, $r(s,a)$ is the immediate reward for choosing action a in state s with angular brackets denoting expectation, because reward delivery can be stochastic. With probability $P(s'|s,a)$ a transition towards next state s' occurs. $0 \leq \gamma \leq 1$ is the temporal discount factor. Model-free reinforcement learning (RL) relies on stored values, whereas model-based RL involves prospective planning and comparison of action outcomes based on a model of the environment ($P(s'|s,a)$ and $\langle r(s,a) \rangle$).

Approach. We introduce a spiking neural network for goal-directed decision making whose dynamics implement value iteration to solve the Bellman equation using biologically realistic neurons and synapses. For a pictorial introduction of the model and illustration of the neural dynamics, Fig. 1 shows a simple but non-trivial two-step task (which cannot be solved by open-loop control) with one stochastic transition and the way our network solves it successfully. Neurons code conjunctively for state-action (s,a) pairs and are modelled as stochastic spike-response-model neurons with the instantaneous firing rate $\phi(u)$ modelled as a linear rectified function of the membrane potential u. Neural activity in our network represents approximate values $\tilde{V}(s) := \sum_a \phi_{sa}$ that converge to the optimal values $V^*(s)$ as the network dynamics evolves (Fig. 1G). Based on the asymptotic activities, the optimal planning problem reduces to simply selecting the action for which the neuron with highest activity codes. Our design intuition for the neural network has been to interpret transition probabilities as well as immediate expected rewards as excitatory synaptic efficacies and to augment the network by mutual lateral inhibition to implement the nonlinear max-operation (Fig. 1C). For this choice of synaptic efficacies, the fixed point of the dynamics corresponds to the solution of the Bellman optimality equation. The dynamics provably always converges to this fixpoint for $\gamma < 1$, i.e. this fixed point is in fact globally stable. We also show that the network does not need to be pre-wired according to a known transition and reward model, but the connectivity can be learned through interaction with the environment if the model is unknown – as it is initially in biological settings. A local plasticity mechanism enables learning the correct weights based on reward-prediction and state-prediction errors. We were also able to generalise the model to continuous states by representing value functions $\tilde{V}(s)$ using a

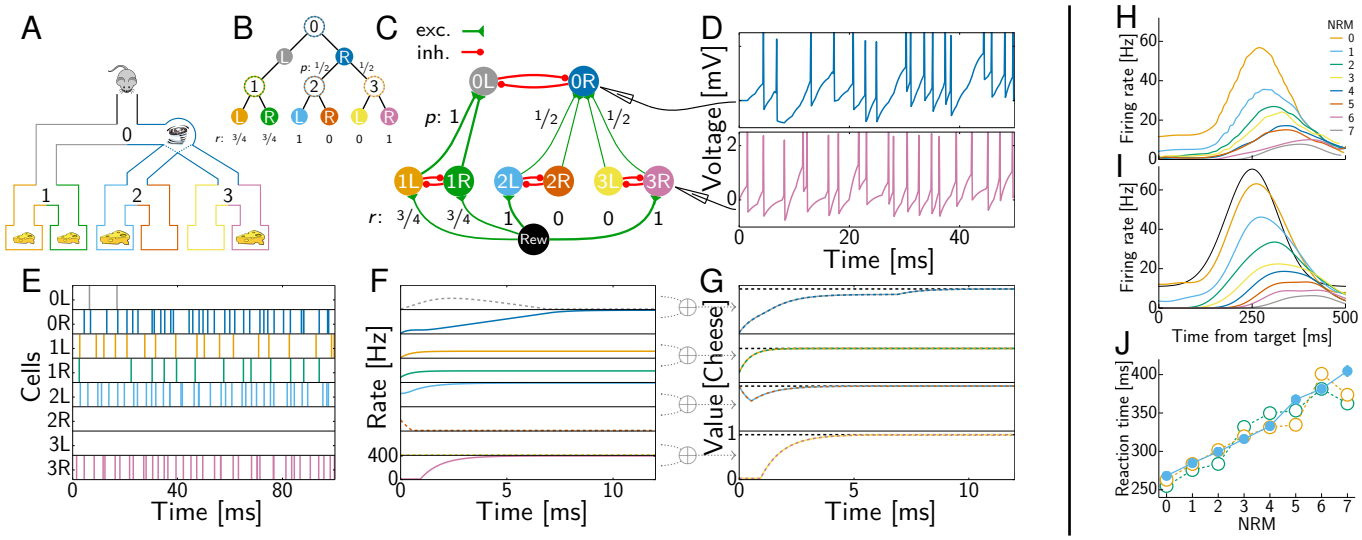

Figure 1: **The model solving a simple illustrative sequential decision making task: (A)** Task. **(B)** Decision graph. Numerical values indicate transition probabilities and rewards. Action nodes (coloured) are identified with neurons. **(C)** Corresponding neural network. One further neuron (black) provides constant activity input. **(D)** Voltage traces for two neurons from (C). **(E)** Spike trains of all (sa)-neurons. **(F)** Underlying firing rate. **(G)** Approximate values (sums from (F)) converge to the optimal values (black dashed lines). **The model reproducing neural and behavioural data in a sequential decision making task: (H)** Experimental data. (See text for details.) **(I)** Model prediction. **(J)** Experimental reaction times (green) highly correlate with the firing rate peak times from (H) (orange). The same holds for the peak times from the simulations in (I) (cyan). Experimental data reproduced from [3].

linear combination of basis functions $\tilde{V}_\theta(s) = \sum_i \theta_i \psi_i(s)$ without making any assumptions about their shapes.

Results. We verified the model's ability to cope with challenging problems by considering some well-known benchmark tasks (Blackjack, Maze, and Pendulum swing-up) from the machine learning literature. Convergence time depends on task difficulty, but already after merely 200 ms a nearly optimal policy has been found in all tasks. The derived plasticity rules are shown to indeed enable the learning of the environmental model. We next turned to tasks typically used in behavioural neuroscience experiments to test the presented theoretical framework against empirical data. For binary delayed-response choice tasks [1], in which monkeys chose between two juice offers, our model correctly predicted the sigmoidal psychometric curves for choice probabilities, as well as behavioural reaction times, with decisions taking longer when options are closely matched in value. It also correctly predicted the experimentally found neural response activities [1] which are a hallmark of our model: encoding of the value if the action which the neuron encodes is chosen, as well as suppression due to mutual inhibition if unchosen. The model did not only correctly predict time averaged neural activities, but further captured their time courses and their dependence on both offer values [2]. Crucially, in tasks involving multiple steps, researchers reported neurons that modulated their activity according to the number of remaining movements (NRM, Fig. 1H) [3]. They also reported behavioural reaction times, which we observed to be approximately equal to the peak times of the neural activity traces (Fig. 1J) [3]. Our model yielded similar activity profiles (Fig. 1I), in particular the observed decrease in firing rate and increased onset delay for an increasing number of remaining movements was a fundamental prediction of our model. Based on the time of the peak, we extracted the reaction times for our model, which also aligned well with the experimental data. Our model further postulates neural reactivation during execution of the action sequence to make the information about current state-action transitions available locally at the synapse thus enabling learning, i.e. plasticity. Indeed, this reactivation is a critical finding of another experimental study [4].

[1] Padoa-Schioppa & Assad (2006) *Nature* 441:2236.
[2] Padoa-Schioppa (2013) *Neuron* 80:132236.
[3] Sohn & Lee (2007) *J Neurosci* 27:1365566.
[4] Musiake et al. (2006) *Neuron* 50:631641.