Spike-based Reinforcement Learning in Continuous State and Action Space.

Eleni Vasilaki§ Robert Urbanczik‡ Walter Senn‡
Wulfram Gerstner§

§Laboratory of Computational Neuroscience, EPFL, 1015 Lausanne, Switzerland
‡Department of Physiology, University of Bern, Bühlplatz 5, 3012 Bern, Switzerland

April 8, 2009

1 Introduction

We propose a reward-modulated synaptic learning rule for spiking neurons based and demonstrate its performance on a learning task in continuous space inspired by the Morris Water maze. The synaptic update rule modifies the release probability of synpatic transmission and depends on the timing of presynaptic spike arrival, postsynaptic action potentials, as well as the membrane potential of the postsynaptic neuron. The rule is implemented in a population of spiking neurons using a network architecture that combines feedforward input with lateral connections. States are represented as spike trains of poisson neurons with overlapping receptive fields. Actions are represented by a population of hypothetical action cells whose activity is read out at theta frequency. We show that the model can learn the Morris watermaze task within 10-15 trials, consistent with experiments.

Our approach can be classified as a policy gradient method similar to the earlier work of [1, 2, 3], but goes beyond these earlier studies. The resulting synaptic update rules can be formulated as a differential equation in continuous time that has the form of a three-factor rule

$$\frac{dw_{ij}}{dt}(t) = \alpha(w_{ij})\; R\; e_{ij}(t) \tag{1}$$

$$e_{ij}(t) = \int_0^\infty \gamma(t-t')\, f_1(pre_j(t'))f_2(post_i(t'))\, dt' \tag{2}$$

The term e_{ij}, called eligibility trace, picks up the correlations between pre- and postsynaptic activity just as in a Hebbian learning rule and convolves these with a low-pass filter γ. However, the final weight change is implemented only in the presence of a reward signal $R(t)$. In contrast to earlier work of [1] but similar to [2, 4, 3] our approach takes into account spiking neurons with refractoriness and includes examples such as the standard integrate-and-fire model. In contrast to most earlier work [2, 3, 5], our learning rule is applied to a network of neurons that combines feed-forward input with lateral interactions. Our novel learning rule works significantly faster and more reliable than earlier policy gradient rules for spiking neurons [2, 3].

2 Results

We perform simulations of a model rat navigating in a square maze of 1 m^2, with a constant speed of 20cm/s. The rat performs a number of trials, with each trial consisting of an attempt to find the goal within a time limit of 90 seconds. At the beginning of each trial, the rat is placed near one of the walls of the maze. Actions are chosen at theta frequency (every 200ms). The rewarded position (target) is at a random position near the central region of the maze and remains fixed at the same position within a set of trials whereas the initial position of the rat varies, as in the experimental paradigm [6, 7, 8]. Positive reward is only given if the rat reaches its target and negative reward if it hits the wall. Given the rat's location, the direction of the next move is decided by the population vector of the action cells. We use the population vector of the synaptic strength of the feedforward connections from a given place cell to visualize the direction of motion starting at that location. The combination of vectors gives a flow map, corresponding to the navigation map of the rat, see Figure 1 . We find that already after 10 trials, a rough strategy for the Morris watermaze task. The performance of the rat is measured by the time it takes to reach the target, corresponding to the escape latency in the experimental literature [6, 7, 8], see Figure 2. The performance is similar to that seen in experimental data [6] and previous models [9, 10].

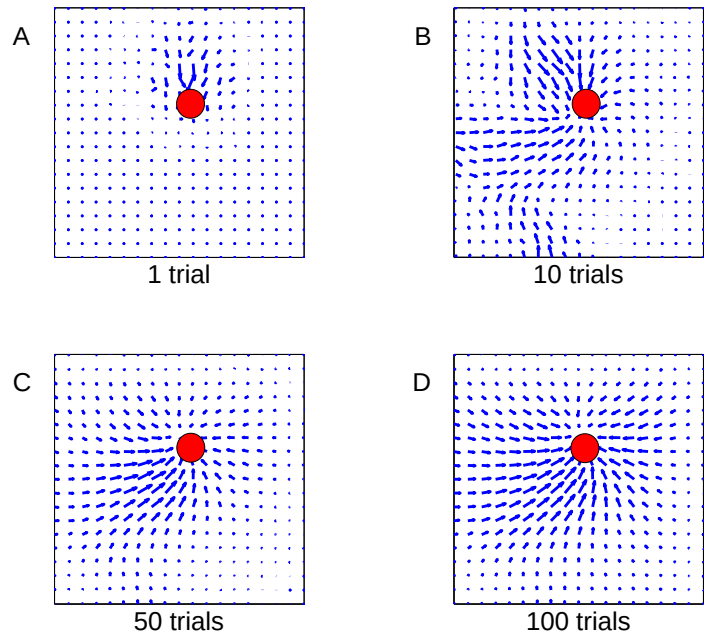

Figure 1: Navigation map of the rat after A:1, B:10, C:50 and D:100 trials visualized in the water maze by a set of direction vectors. At each grid point (defined by the center of a place cell j) in the graph, we plot the stochastic release probability q_{ij} for fixed j in the form of a population vector denoting the direction the animal would most likely take at this location. The circle marks the position of the hidden platform.

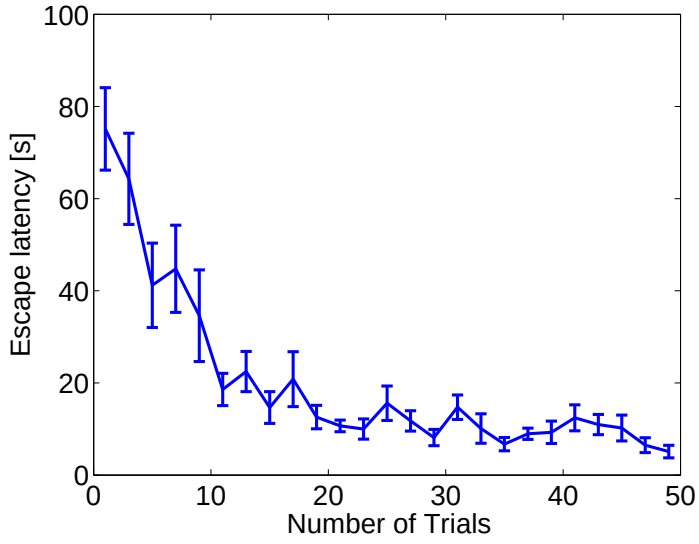

Figure 2: Water Maze task, average performance for 10 sets of trials. The number of steps it takes to reach the platform decreases with the number of trials, indicating that the simulated rat is learning the location of the platform. Error bars show standard error. Place cells are located every 5 cm, with a gaussian receptive field of σ=8 cm, and maximum firing rate 120Hz. Release probabilities are low and upper bounded by the values 0.15 and 1.

References

[1] X. Xie and S. Seung. Learning in neural networks by reinforcement of irregular spiking. *Phys. Rev. E*, 69:41909, 2004.

[2] J.-P. Pfister, T. Toyoizumi, D. Barber, and W. Gerstner. Optimal spike-timing dependent plasticity for precise action potential firing in supervised learning. *Neural Computation*, 18:1309–1339, 2006.

[3] R. V. Florian. Reinforcement learning through modulation of spike-timing-dependent synaptic plasticity. *Neural Computation*, 19:1468–1502, 2007.

[4] E.M. Izhikevich. Solving the distal reward problem through linkage of stdp and dopamine signaling. *Cerebral Cortex*, 17:2443–2452, 2007.

[5] Robert Legenstein, Dejan Pecevski, and Wolfgang Maass. Theoretical analysis of learning with reward-modulated spike-timing-dependent plasticity. In J.C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*. MIT Press, Cambridge, MA, 2008.

[6] R.G.M. Morris, P. Garrard, J.N.P. Rawlins, and J. O'Keefe. Place navigation impaired in rats with hippocampal lesions. *Nature*, 297:681–683, 1982.

[7] R.G.M. Morris, E. I. Moser, G. Riedel, S. J. Martin, J. Sandin, M. Day, and C. O'Carrol. Elements of a neurobiological theory of the hippocampus: the role of activity-dependent synaptic plasticity in memory. *Phil. Trans. R. Soc. Lond B: Biological Sciences*, 358:773 – 786, 2003.

[8] R. Morris. Theories of hippocampal function. In *The hippocampus book*, pages 581–713. Oxford university press, 2007.

[9] D. Sheynikhovich, R. Chavarriaga, T. Strösslin, and W. Gerstner. Spatial representation and navigation in a bio-inspired robot. In *Biomimetic Neural Learning for Intelligent Robots: Intelligent Systems, Cognitive Robotics, and Neuroscience*, pages 245–264, 2005.

[10] David Foster, Richard Morris, and Peter Dayan. Models of hippocampally dependent navigation using the temporal difference learning rule. *Hippocampus*, 10:1–16, 2000.