

On the Biological Plausibility of Reinforcement Learning by Policy Search

Leonid Peshkin (pasha@ai.mit.edu) MIT AI Lab., 200 Technology. Sq., Cambridge, MA 02139

Virginia Savova (savova@wjh.harvard.edu) Cognitive Science, Johns Hopkins University, Baltimore, MD 21218

The principal question what kind of learning could occur in biological neural networks is of considerable importance to both machine learning and neuroscience. Our work offers an analysis of the bio-plausibility of learning algorithms with an emphasis on recently developed reinforcement learning (RL) methods for cooperative agents. We take a learning mechanism to be biologically plausible if it could be realized within the temporal, spatial and material constraints acting on a living organism.

RL [Sutton and Barto'98, Kaelbling et al.'96]¹ is an extension of supervised learning for the case when supervision is infeasible, but feedback signal is still present. There is considerable body of work on bio-plausibility [Dayan & Abbott'01; Rumelhart & McClelland'86] with respect to classical methods, extending dynamic programming in Markov Decision Processes (e.g. actor-critic algorithms [Konda & Tsitsiklis'99], and value search methods [Sutton & Barto'98]). However, a recently developed class of so-called policy search methods applicable to a broad class of realistic scenarios received no attention in light of bio-plausibility (with exception of perhaps [Bartlett & Baxter'99] who relate RL to Hebbian synaptic updates).

Initially introduced as a model of learning in a single agent, policy search by gradient descent [Williams 1992, Marbach & Tsitsiklis'99] was developed for the case of multi-agent learning [e.g. Peshkin'01] and has very attractive mathematical properties, e.g. achieving locally optimal performance. The general idea is that parameters encoding the behavior are adjusted towards the direction of the gradient of cumulative reward signal estimated empirically rather than calculated analytically. A remarkable property of these algorithms is that no coordination is necessary for agents to cooperate, as long as they receive the same reward (see Bartlett & Baxter'99, Peshkin'00). The neurons in a neural net can be viewed as cooperating agents in some cognitive control task.

The learning schema requires only simple computations (differentiation, integration) based on locally available quantities (cumulative reward) and could be in principle implemented in the neural network, except that the reward signal must be identical in all parts of the system. This requirement could be problematic for the common biological scenario, where reward is delivered into the neural net in the form of a mediator secreted by some unevenly distributed sources. The nature of diffusion suggests that the concentration of the mediator is smaller for more distant locations.

We point out that the physical distribution of the reward signal would cause an uneven delay in different parts of the neural net. We present a hypothetical solution to this problem via alternation of functional and learning epochs implemented by signal transduction pathways from synaptic activity to gene expression in the nucleus as an adaptive response (e.g. [Curtis & Finkbeiner'99]). We go on to discuss the biological plausibility of reward distribution via dopamine. Dopamine has been put forth as a possible reward signal in other reinforcement learning algorithms linked to biology (e.g. [Kakade & Dayan'00]). One important fact about the distribution of dopamine in the brain is that it is fairly uniform because the dopamine innervations of the basal ganglia and the cortical regions are denser than the size of an average nerve cell body [Schultz'98]. This circumstance significantly eases the implementation of the policy search algorithm into a biological system. However, the biological role of dopamine depends on the threshold activation of dopamine receptors. This effectively means the update neurons perform does not depend linearly on dopamine concentration. We invite a discussion with the neuroscience community regarding candidate mechanisms for reward and regulatory/expression cycles in the neurons related to synaptic adjustments, as well as evidence of "wake/sleep" interchanging epochs [Hinton G., Dayan P., Frey B. and N. Radford'95].

Selected references:

- P.L. Bartlett and J. Baxter (1999). Hebbian Synaptic Modifications in Spiking Neurons that Learn. *Technical Report, School of Information Sciences and Engineering, Australian National University*
- J. Curtis and S. Finkbeiner (1999). Sending Signals From the Synapse to the Nucleus: Possible Roles for CaMK, Ras/ERK, and SAPK Pathways in the Regulation of Synaptic Plasticity and Neuronal Growth, *Journal of Neuroscience Research* 58:88–95 Hebb, D. O (1949). *The Organization of Behavior*, New York
- P. Dayan and L. Abbott, (2001) *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (MIT Press, Cambridge MA).
- S. Kakade and P. Dayan (2000). Dopamine Bonuses. *Technical Report*, Gatsby Comp. Neuroscience Unit, London.
- L. Peshkin (2001) Reinforcement Learning via Policy Search. PhD dissertation. Brown University
- R. Sutton and A. Barto (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA

¹ Complete list of references and this abstract are available at <http://www.ai.mit.edu/~pasha/papers.html>