Bernstein Conference 2019

Deep reinforcement learning in a time-continuous model

Akos Ferenc Kungl [1], Dominik Dold [1], Oskar Riedler [2], Walter Senn [3], Mihai Alexandru Petrovici [1, 3]

1. Departement for Physics and Astronomy, Kirchhoff Institute for Physics, Heidelberg University, Im Neuenheimer Feld 227, 69120, Germany
2. Heidelberg University, Heidelberg, Germany
3. Institute for Physiology, University of Bern, Bühlplatz 5, 3012 Bern, Switzerland

Inspired by the recent success of deep learning [1], several models emerged trying to explain how the brain might realize plasticity rules reaching similar performances as deep learning [2-5]. However, all of these models consider only supervised and unsupervised learning, where an external teacher is needed to produce an error signal that guides plasticity.

In this work, we introduce a model of reinforcement learning based on the principle of Neuronal Least Action (R-NLA). We extend previous works on time-continuous error backpropagation in cortical microcircuits [4, 6] to achieve a biologically plausible model implementing deep reinforcement learning.

In R-NLA the neurosynaptic dynamics is derived from the energy function using the variational principle. In the resulting dynamics the phase-advanced firing of the neurons effectively undoes the network delay introduced by finite membrane time-constants. Errors are introduced to the network by nudging, and they are propagated to deeper layers via cortical microcircuits. Instead of having an explicit teacher, the output neurons, which represent the actions, form a soft winner-take-all network (Fig A). This winner-take-all structure evokes a nudging on the soma of the output neurons, which is subsequently backpropagated through the network. A reward prediction error $\delta = R - \langle R \rangle$ modulates the plasticity multiplicatively as a formally deduced global reward-specific neuromodulator [7]. By construction, the learning rule approximates the policy gradient of the mean expected reward.

Using a simple pattern recognition problem as a toy example, we show that R-NLA can learn classification tasks in the reinforcement learning framework with similar performance as an equivalent deep reinforcement learning model (Fig B). Further, we show that it is robust against time-delayed rewards, even if the reward delay is not constant but randomly distributed (Fig C).

R-NLA constitutes a time-continuous implementation of biologically plausible deep reinforcement learning, robust to delayed reward. The self-teaching soft winner-take-all mechanism removes the necessity of an explicit teacher and the proposed learning rule solves the problem of synaptic consolidation. The model can be extended to an actor-critic model, where a second (deep) critic network learns the state-value function.

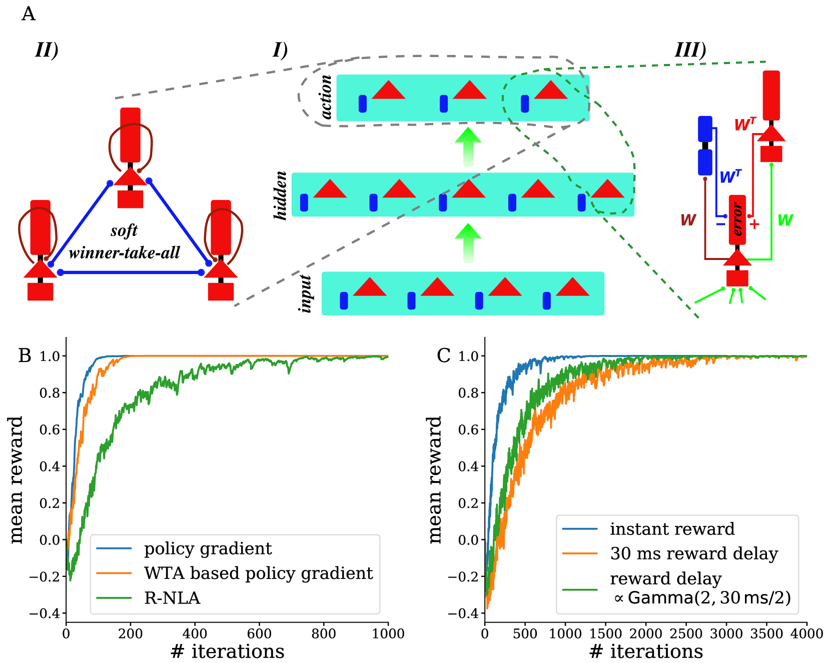

Figure 1: A-I) Network schematics. A-II) Soft winner-take-all network in the output layer. A-III) Microcircuit for error backpropagation. B) Comparison to classical reinforcement learning methods. C) Robustness with respect to reward delay.

Acknowledgements

The work leading to these results has received funding from the European Union grant agreements No 720270 and 785907 (Human Brain Project, HBP). We owe particular gratitude to the sustained support of our research by the Manfred Stärk Foundation.

References

1. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature, 521(7553):436, 2015. (https://www.nature.com/articles/nature14539), 10.1038/nature14539 (http://dx.doi.org/10.1038/nature14539)
2. Randall C O'Reilly. Biologically plausible error-driven learning using lo- cal activation differences: The generalized recirculation algorithm. Neural Computation, 8(5):895–938, 1996. (https://dl.acm.org/citation.cfm?id=1362149), 10.1162/neco.1996.8.5.895 (http://dx.doi.org/10.1162/neco.1996.8.5.895)
3. James CR Whittington and Rafal Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. Neural Computation, 29(5):1229–1262, 2017. (https://www.mitpressjournals.org/doi/10.1162/NECO_a_00949), 10.1162/NECO_a_00949 (http://dx.doi.org/10.1162/NECO_a_00949)
4. Joao Sacramento, Rui Ponte Costa, Yoshua Bengio, and Walter Senn. Den- dritic cortical microcircuits approximate the backpropagation algorithm. In Advances in Neural Information Processing Systems, pages 8721–8732, 2018. (https://papers.nips.cc/paper/8089-dendritic-cortical-microcircuits-approximate-the-backpropagation-algorithm.pdf)
5. James CR Whittington and Rafal Bogacz. Theories of error back- propagation in the brain. Trends in Cognitive Sciences, 2019. (https://www.cell.com/trends/cognitive-sciences/fulltext/S1364-6613(19)30012-9), 10.1016/j.tics.2018.12.005 (http://dx.doi.org/10.1016/j.tics.2018.12.005)

6. Dominik Dold, Akos F Kungl, Joao Sacramento, Mihai A Petrovici, Kaspar Schindler, Jonathan Binas, Yoshua Bengio, and Walter Senn. Lagrangian dynamics of dendritic microcircuits enables real-time backpropagation of errors. In Cosyne Conference, 2019. (http://www.kip.uni-heidelberg.de/Veroeffentlichungen/details.php?id=3855)
7. Nicolas Fremaux and Wulfram Gerstner. Neuromodulated spike-timing- dependent plasticity, and theory of three-factor learning rules. Frontiers in Neural Circuits, 9:85, 2016. (https://www.frontiersin.org/articles/10.3389/fncir.2015.00085/full), 10.3389/fncir.2015.00085 (http://dx.doi.org/10.3389/fncir.2015.00085)

Citation: Kungl AF, Dold D, Riedler O, Senn W, Petrovici MA (2019) Deep reinforcement learning in a time-continuous model. Bernstein Conference 2019. doi: 10.12751/nncn.bc2019.0168 (http://doi.org/10.12751/nncn.bc2019.0168)