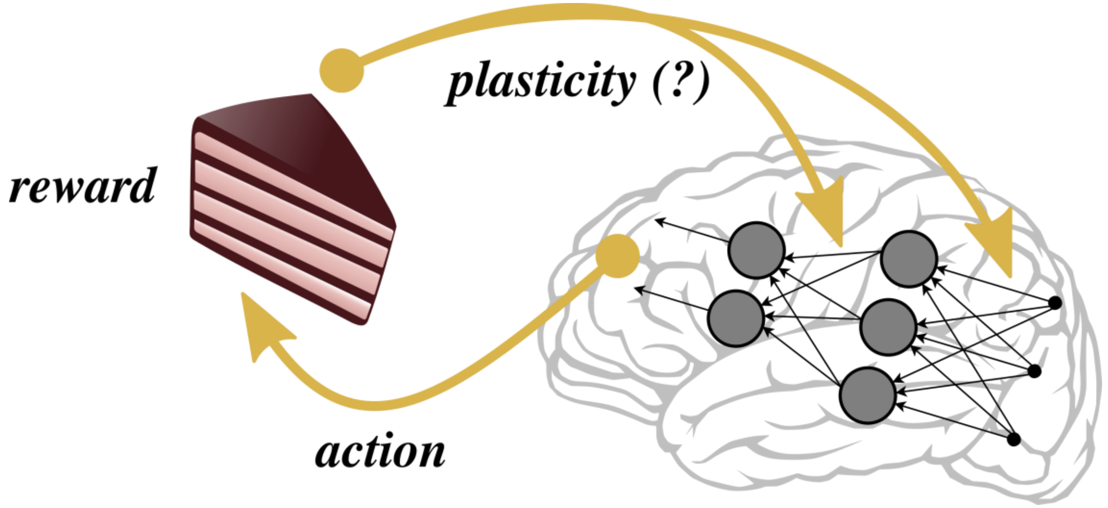
Deep reinforcement learning for time-continuous substrates

NICE Conference 2021 - online

Akos F. Kungl[1], Dominik Dold[1], Oskar Riedler[2], Walter Senn[3], Mihai A Petrovici[3]

[1]Kirchhoff Institute for Physics, Heidelberg, Germany; [2]Heidelberg University, Heidelberg, Germany; [3]Department of Physiology, Bern, Switzerland

Motivation: How is error propagated over several processing stages?



Observations:

1. From machine learning we know the benefits of learning over several non-linear layers, and we can train such models with backpropagation.
2. In the brain, we also observe information processing over several layers of neurons. How do the synapses calculate their contribution to the overall error? How is this error generated?

Challenges: Constraints of realistic models

1. Locality of plasticity: The update of the synapses should only explicitly depend on locally available parameters.
2. Time-continuous system: The brain observes time-continuous dynamics. Discretizing time or putting time on hold should be preferably avoided.

Our contribution:

A framework of deep supervised and reinforcement learning that respects the time-continuous dynamics and the locality.

Sacramento et al. (2018), Dold and Kungl et al. (2019)
Kungl (2020), Kungl et al (in preparation)

Least action principle for neural networks

what the neuron does

what the neuron's input would imply

$$E(\boldsymbol{u}) = \sum_i \frac{1}{2} \|\boldsymbol{u}_i - \boldsymbol{W}_i \bar{\boldsymbol{r}}_{i-1}\|^2 \; + Cost$$

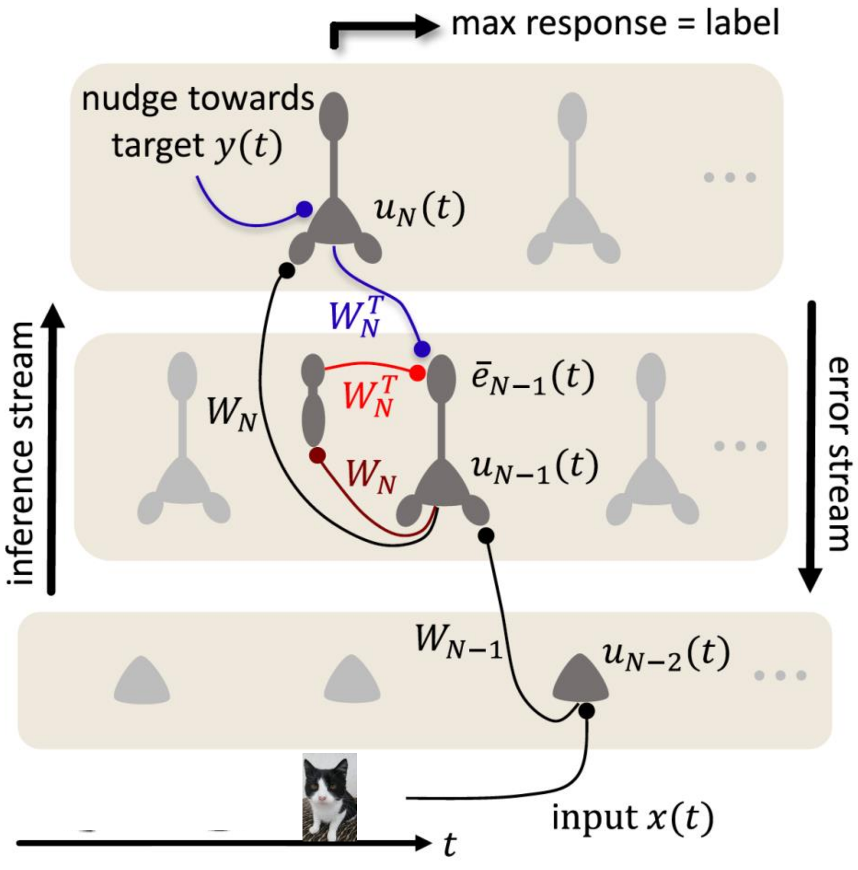
$$\boldsymbol{u} = \tilde{\boldsymbol{u}} - \tau \dot{\tilde{\boldsymbol{u}}} \longrightarrow \quad L(\tilde{\boldsymbol{u}}, \dot{\tilde{\boldsymbol{u}}})$$

$$\frac{\partial L}{\partial \tilde{\boldsymbol{u}}_i} - \frac{d}{dt}\frac{\partial L}{\partial \dot{\tilde{\boldsymbol{u}}}_i} = 0 \qquad \dot{\boldsymbol{W}}_i = -\eta \frac{\partial E}{\partial \boldsymbol{W}_i}$$

$\tau \dot{\boldsymbol{u}}_i = \boldsymbol{W}_i \boldsymbol{r}_{i-1} - \boldsymbol{u}_i + \boldsymbol{e}_i \;\; \rightarrow$ neuron dynamics!

$\bar{\boldsymbol{e}}_i = \bar{\boldsymbol{r}}_i' \odot [\boldsymbol{W}_{i+1}^T \boldsymbol{u}_{i+1} - \boldsymbol{W}_{i+1}^T \boldsymbol{W}_{i+1} \bar{\boldsymbol{r}}_i] \;\; \rightarrow$ communication of error!

$\dot{\boldsymbol{W}}_i = \eta(\boldsymbol{u}_i - \boldsymbol{W}_i \bar{\boldsymbol{r}}_{i-1}) \bar{\boldsymbol{r}}_{i-1}^T \;\; \rightarrow$ local learning rule!



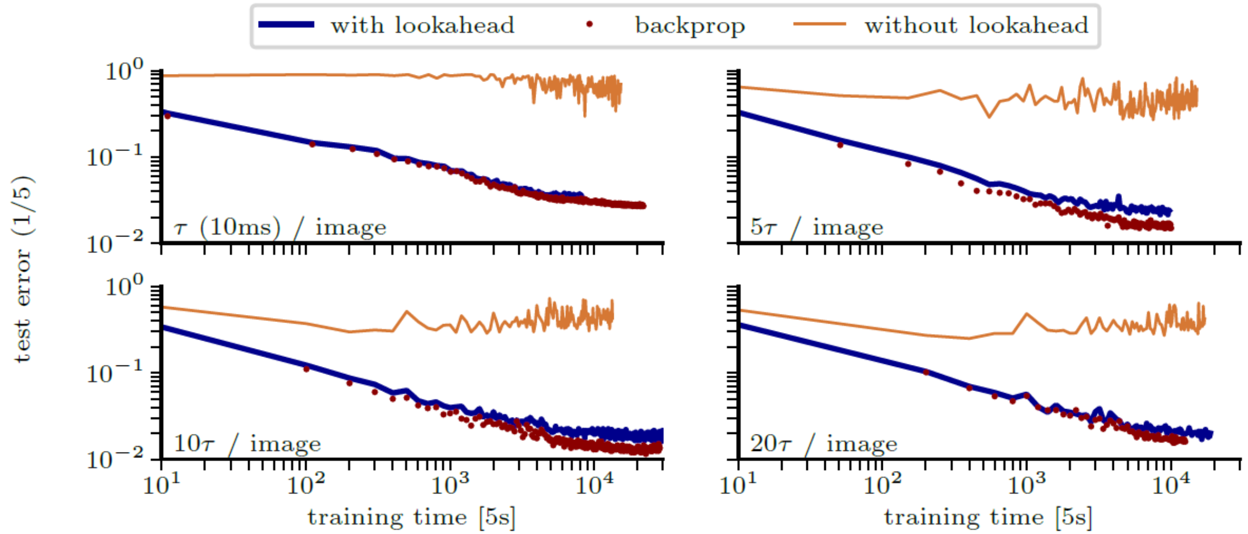
Sacramento et al. (2018), Dold and Kungl et al. (2019)
Kungl (2020), Kungl et al (in preparation)

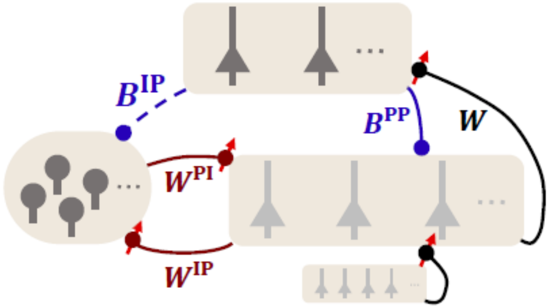
Results with supervised learning replicate backpropagation

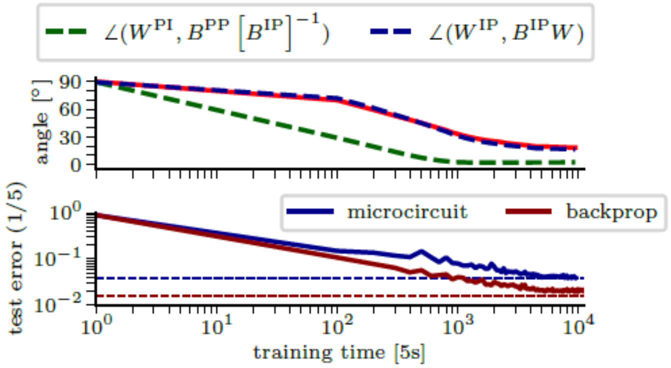
The results on the MNIST dataset show that **our model replicates backpropagation** as measured on learning per iteration.

The **look-ahead capability of the neurons is an essential feature** required for backpropagation. Without it the forward and the backward pass experience a delay compared to each other and the learning breaks.

With further changes in the architecture, **the symmetry of distant weights can be reduced completely**, and the model is still capable of learning.







Dold (2020), Senn, Dold et al. (in preparation)

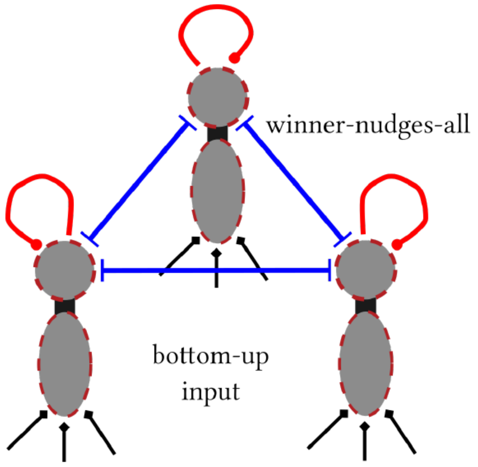
Reinforcement Learning by approximating policy gradient learning

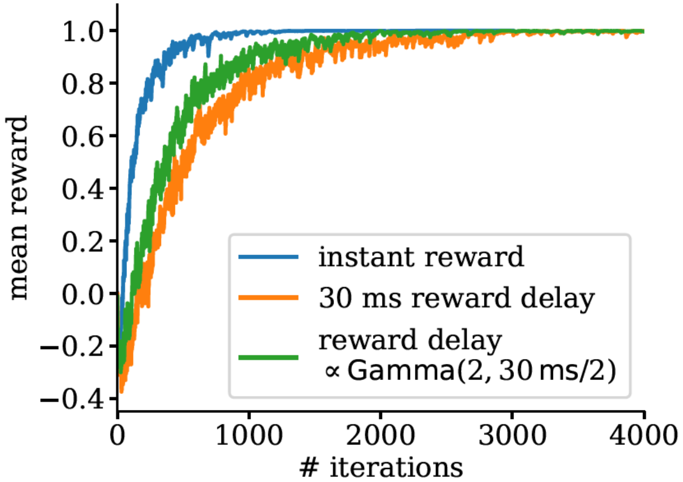
Local error term
(spatial credit assignment)

$$\dot{\boldsymbol{W}}_i(t) \sim (R - \bar{R}) \int_{-\infty}^{t} (\boldsymbol{u}_i - \boldsymbol{W}_i \bar{\boldsymbol{r}}_{i-1}) \bar{\boldsymbol{r}}_{i-1}^T \, e^{-\frac{t-\hat{t}}{\tau}} d\hat{t}$$

Reward feedback
(global signal)

Relate cause and reward in time
(temporal credit assignment)





The winner nudges all structure creates an error signal from the current choice of the network.

The resulting error represents a **Hill climbing** on the mean expected reward.

This error is propagated back to the layers closer to the input.

An incoming reward reinforces or penalizes the suggested weight changes.

Kungl (2020), Kungl et al. (in preparation)

Limitations and Outlook

For more see Walter Senn's talk CET 17:25, Thursday 18th March

Limitations:

1. The model does not include spikes and synaptic delays.
2. The look-ahead mechanism has experimental indication but is not modelled explicitly.

Outlook:

1. The supervised learning can be canonically extended to improve the vanilla backpropagation.
2. The reinforcement learning could be extended to an actor-critic framework.
3. Neuromorphic hardware could greatly benefit from the continuous learning capabilities.

Köndgen (2008), Fremaux (2013)

References

Joao Sacramento, Rui Ponte Costa, Yoshua Bengio, and Walter Senn. *Dendritic cortical microcircuits approximate the backpropagation algorithm*. In Advances in Neural Information Processing Systems, pages 8721–8732, 2018.

Kungl, Ákos Ferenc. *Robust learning algorithms for spiking and rate-based neural networks*. Diss. 2020.

Dold, Dominik. *Harnessing function from form: towards bio-inspired artificial intelligence in neuronal substrates*. Diss. 2020.

Walter Senn, Dominik Dold, Akos F. Kungl, Benjamin Ellenberger, Yoshua Bengio, Joao Sacramento, Jakob Jordan, and Mihai A. Petrovici. *Least action principle for real-time dendritic errorpropagation across cortical microcircuits*, (in preparation)

Harold Kondgen, Caroline Geisler, Stefano Fusi, Xiao-Jing Wang, Hans-Rudolf Luscher, and Michele Giugliano. *The dynamical response properties of neocortical neurons to temporally modulated noisy inputs in vitro.* Cerebral cortex, 18 (9):2086–2097, 2008.

Nicolas Fremaux, Henning Sprekeler, and Wulfram Gerstner. *Reinforcement Learning Using a Continuous Time Actor-Critic Framework with Spiking Neurons*. PLoS Computational Biology, 9(4):e1003024, 4 2013. ISSN 1553-7358. doi:10.1371/journal.pcbi.1003024.