Deep reinforcement learning in a time-continuous model

A. F. Kungl[1,2], D. Dold[1,2], O. Riedler[3], M. A. Petrovici[1,2], W. Senn[2]

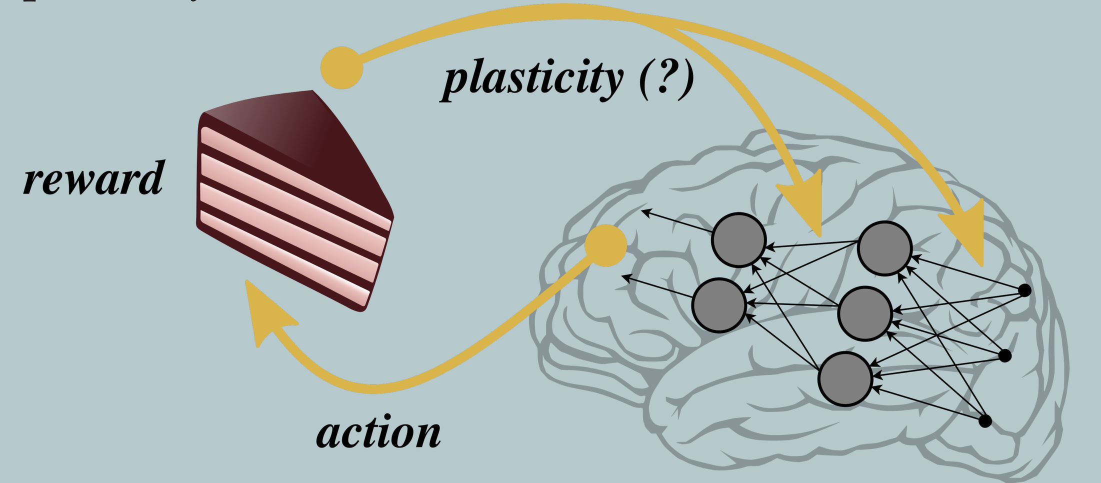
[1]Heidelberg University, Kirchhoff-Institute for Physics; [2]University of Bern, Institute for Physiology; [3]Heidelberg University

The Manfred Stärk Foundation

Electronic Vision(s)

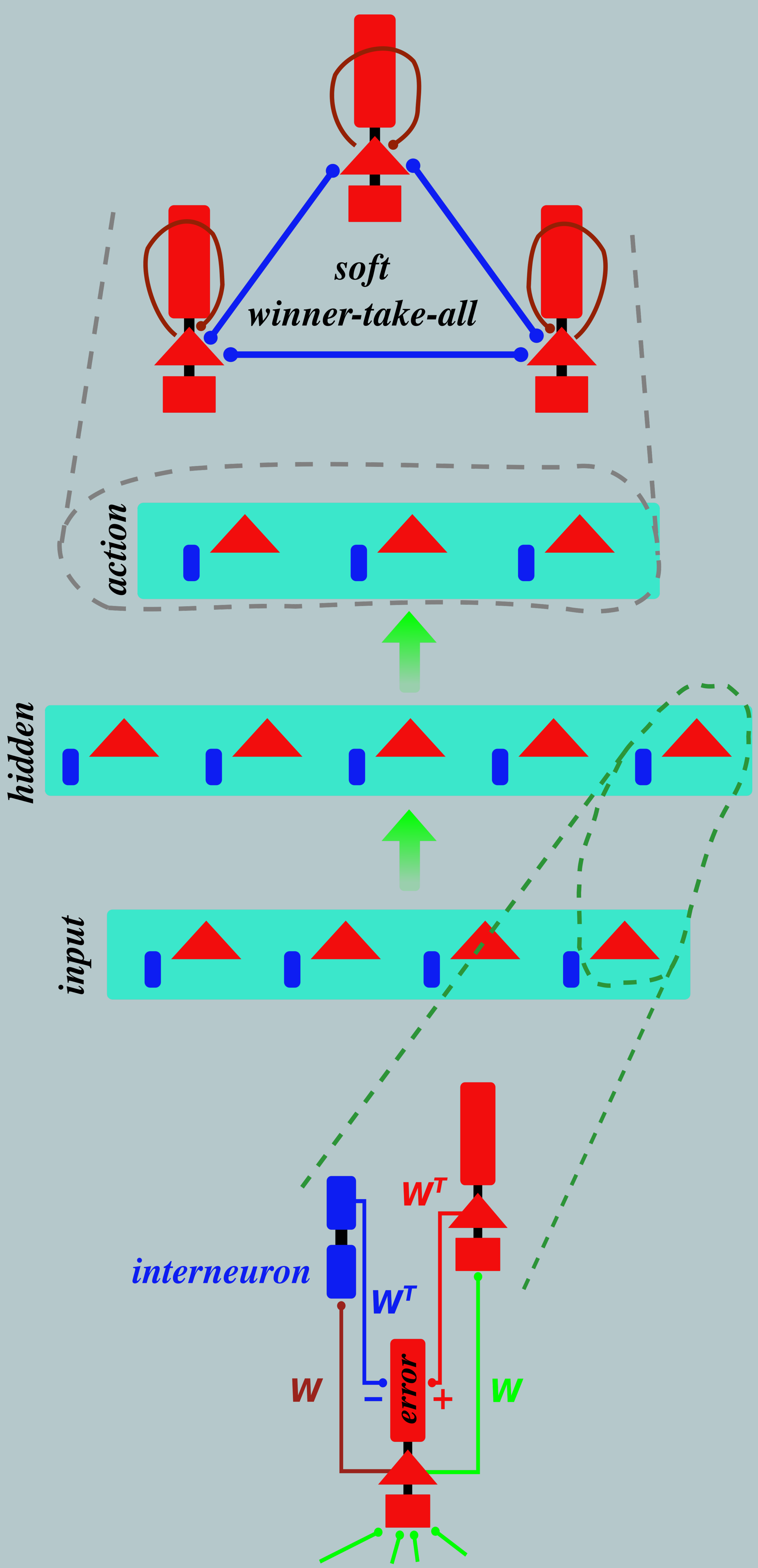
1 Objectives

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]. 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.



We introduce a model of **reinforcement learning with the principle of Neuronal Least Action (R-NLA)**. We extend previous works on **time-continuous error backpropagation in cortical microcircuits** [3, 4] to achieve a biologically plausible model implementing deep reinforcement learning.

2 A time-continuous deep learning model



The neuro-synaptic dynamics is derived with the **Lagrange formalism** using the Lagrange function

$$L = \underbrace{\frac{1}{2}\sum_i \|u_i - W\phi(u)_i\|^2}_{energy} + \underbrace{\beta M \int \phi(u_i) du_i}_{cost}$$

leading to **neuron dynamics resembling a leaky integrator**:

$$\begin{aligned}
\tau \dot{u}_i &= W_i r_{i-1} - u_i + e_i \\
r_i &= \bar{r}_i + \tau \dot{\bar{r}}_i \\
e_i &= \bar{e}_i + \tau \dot{\bar{e}}_i \\
\bar{e}_i &= \bar{r}'_i \odot W_i^T \left(u_i - W_{i+1}\bar{r}_i\right) \\
\bar{e}_N &= \beta M \bar{r}_N \\
M &= \begin{pmatrix} 1 & -\frac{1}{N-1} & \cdots & -\frac{1}{N-1} \\ -\frac{1}{N-1} & 1 & \cdots & -\frac{1}{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{N-1} & -\frac{1}{N-1} & \cdots & 1 \end{pmatrix}
\end{aligned}$$

In the last layer the **error vector** $\bar{e}_N = \beta M \bar{r}_N$ approximates the error of policy gradient, which is then propagated to the deeper layers via **stereotypical microcircuits** $\bar{e}_i = W_{i+1}^T \bar{e}_{i+1}$. The learning rule combines the notion of local error correction, eligibility traces and reward, forming a **three-factor learning rule** [5]:

$$\dot{W} = \frac{\eta}{\tau_{\mathrm{elig}}} \left(R - \langle R \rangle\right) \int_{-\infty}^{t} \kappa(\hat{t}) \exp\left(-\frac{t-\hat{t}}{\tau_{\mathrm{elig}}}\right) \mathrm{d}\hat{t}$$

$$\kappa(t) = \underbrace{\left(u_i - W_i \phi(u_{i-1})\right)}_{\propto \bar{e}_i} \phi(u_{i-1})^T$$

Ad hoc **surprise-based homeostasis** supports the learning:

$$\dot{W}_{\mathrm{hom}} = \frac{\eta_{\mathrm{hom}}}{\tau_{\mathrm{elig}}} \left|R - \langle R \rangle\right| \int_{-\infty}^{t} \kappa_{\mathrm{hom}}(\hat{t}) \exp\left(-\frac{t-\hat{t}}{\tau_{\mathrm{elig}}}\right) \mathrm{d}\hat{t}$$

$$\kappa_{\mathrm{hom}}(t) = \left(u_{\mathrm{trg}} - u_i\right) \phi(u_{i-1})^T$$

3 Intuition behind the soft WTA

The error vector arising from the self-nudging and reward modulation is similar to a supervised error.

target			00010
supervised		e_N^{super}	− − − + −
soft WTA	00010 $\longmapsto$	e_N^{wna}	− − − + −
	10000 $\longmapsto$	e_N^{wna}	− + + + +

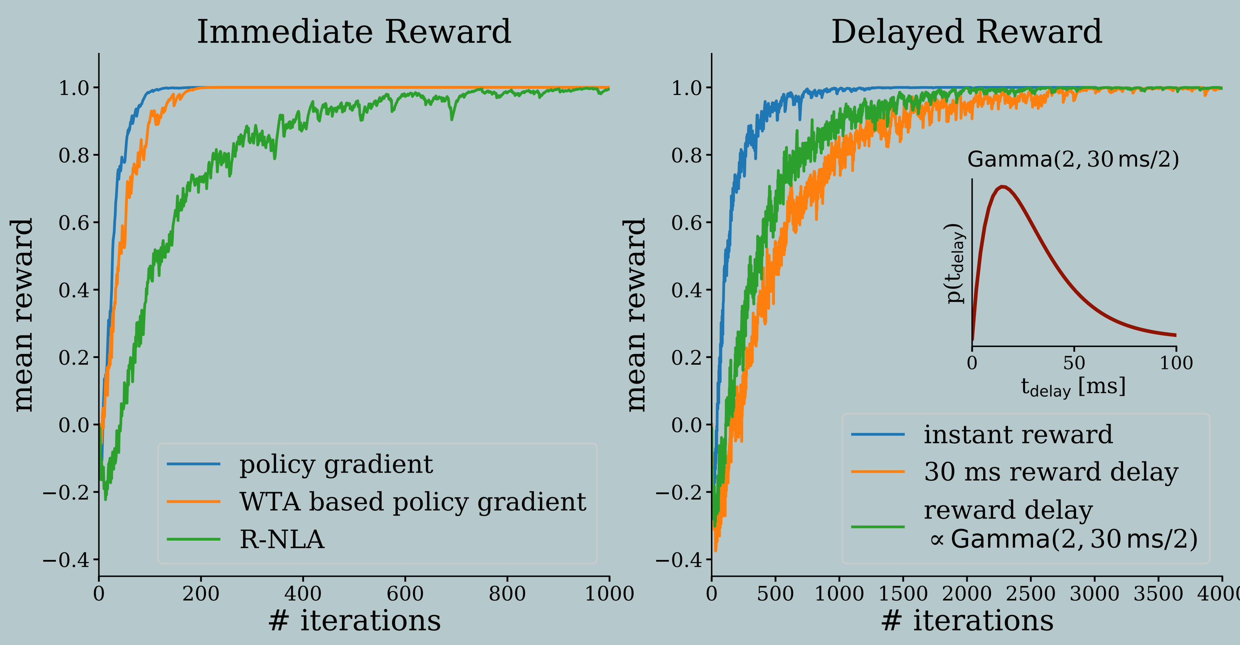
It can be shown that the **self-nudging error points approximately in the same direction as the policy gradient error**.

4 Learning is robust against delayed reward

The time-continuous model performs similarly well as a vanilla policy gradient algorithm or an artificial neural network based implementation.

The time-continuous model is **robust against delay** in the reward even if the reward is delayed:

- by one iteration
- by a random delay $t_{\mathrm{delay}} \sim \mathrm{Gamma}\left(2, \frac{\tau_{\mathrm{iteration}}}{2}\right)$



5 Conclusion

R-NLA unites several desired features

- time-continuous network dynamics
- No phases in the learning
- backpropagation based on local plasticity
- Self-teaching in the action layer (also compatible with node perturbation)
- Robustness against delay in the reward

Check out related research on poster T7 from Dominik Dold et. al.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.

[2] J. C. Whittington and R. Bogacz, "An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity," *Neural Computation*, vol. 29, no. 5, pp. 1229–1262, 2017.

[3] J. Sacramento, R. P. Costa, Y. Bengio, and W. Senn, "Dendritic cortical microcircuits approximate the backpropagation algorithm," in *Advances in Neural Information Processing Systems*, pp. 8721–8732, 2018.

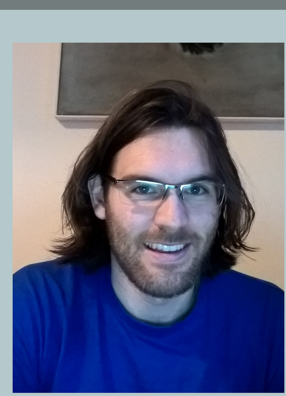
[4] D. Dold, A. F. Kungl, J. Sacramento, M. A. Petrovici, K. Schindler, J. Binas, Y. Bengio, and W. Senn, "Lagrangian dynamics of dendritic microcircuits enables real-time backpropagation of errors," in *Cosyne Conference*, 2019.

[5] N. Frémaux and W. Gerstner, "Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules," *Frontiers in Neural Circuits*, vol. 9, p. 85, 2016.

Funding and Contact

poster

Funded by the Manfred Stärk Foundation and the European Union Horizon 2020 (720270 and 785907, HBP). Calculations performed on UBELIX, the HPC cluster of the University of Bern.



Akos F. Kungl
PhD Student
Kirchhoff Institute for Physics
Heidelberg
Fields of interest: *functional models, neuromorphic computing, plasticity*