Deep reinforcement learning for time-continuous substrates

Akos F. Kungl*
Dominik Dold*
Kirchhoff-Institute for Physics
Heidelberg, Germany

Oskar Riedler
Heidelberg University
Heidelberg, Germany

Walter Senn
Mihai A. Petrovici
Department of Physiology
Bern, Switzerland

ABSTRACT

To achieve their goal of realizing fast and energy-efficient learning, neuromorphic systems require computationally powerful models that obey the constraints imposed by a physical implementation of neural network structure and dynamics, such as the inevitability of relaxation times or the locality of plasticity. In this work, we provide a first-principles derivation of a mechanistic model for cortical computation based on the premise of "neuronal least action". The resulting time-continuous neuron and synapse dynamics realize gradient-descent learning through error backpropagation both in supervised and in reinforcement learning scenarios. In particular, the derived equations of motion reproduce well-established microscopic phenomena such as neuronal leaky integration of afferent signals, while enabling synaptic learning using only locally available information. Our principled framework can thus serve as a starting point for hardware-focused models of highly efficient time-continuous learning.

CCS CONCEPTS

• **Computing methodologies → Modeling and simulation**; **Neural networks**; **Learning paradigms**.

KEYWORDS

neural networks, biological deep learning, prospective coding, physical & mechanistic models, supervised learning, reinforcement learning

ACM Reference Format:
Akos F. Kungl, Dominik Dold, Oskar Riedler, Walter Senn, and Mihai A. Petrovici. 2020. Deep reinforcement learning for time-continuous substrates. In *Heidelberg '20: Neuro-Inspired Computational Elements Workshop, March 17–20, 2020, Heidelberg, Germany*. TBD, 3 pages. https://doi.org/TBD

1 INTRODUCTION

Neuromorphic engineering promises fast and energy-efficient hardware realizations of neural networks that utilize novel computing paradigms inspired by the brain [1, 6, 8, 16, 18]. A special interest lies in systems that are capable of learning from continuous data streams and can therefore adjust to changes in the environment [2, 5]. However, the brain's solution for the credit assignment problem remains elusive, mainly due to the locality of information in physical systems [21]. The error backpropagation algorithm [12], for example, explicitly violates this principle of locality. While several efforts have been made to reconcile this paradigm with local synaptic plasticity, existing solutions either require a separation of time scales, with neuronal dynamics occurring much faster than or separately from synaptic weight changes [14, 20], or with multi-phased learning rules [7, 13, 15].

In this work, we introduce a new model that derives a time-continuous version of error backpropagation from a least-action principle [3] that can be used in supervised and reinforcement learning scenarios. More specifically, we show how learning can happen without separate phases and both with and without an external supervisor. The derived model is compatible with cortical structure and dynamics, suggesting that it can also be ported to brain-inspired neuromorphic systems, which often inherit many physical constraints from their biological archetype.

*Both authors contributed equally to this work.

NICE 2020, 17–20 March, 2020, Heidelberg, Germany

ACM ISBN TBD…$TBD
https://doi.org/TBD

2 RESULTS

2.1 Neuronal least action principle (NLA)

For simplicity, we derive the neuronal dynamics for a feedforward network with N layers, but the approach can be used for arbitrary network topologies. We start by defining the Lagrangian

$$\mathcal{L}(\tilde{\boldsymbol{u}}, \dot{\tilde{\boldsymbol{u}}}, \boldsymbol{W}) = \frac{1}{2} \sum_{k}^{N} \|f(\tilde{\boldsymbol{u}}_k, \dot{\tilde{\boldsymbol{u}}}_k) - \boldsymbol{W}_k \bar{\boldsymbol{r}}_{k-1}\|^2 + \beta C \ , \tag{1}$$

with $\tilde{\boldsymbol{u}}$ implicitly defined by $\boldsymbol{u}_k = f(\tilde{\boldsymbol{u}}_k, \dot{\tilde{\boldsymbol{u}}}_k) = \tilde{\boldsymbol{u}}_k - \tau\dot{\tilde{\boldsymbol{u}}}_k$. Here, $\boldsymbol{u}_k$ is the vector containing the membrane potentials of neurons in layer k, τ the respective membrane time constant and $\boldsymbol{W}_k$ the synaptic connections projecting into layer k. The stationary rate $\bar{\boldsymbol{r}}_{k-1} = \varphi(\boldsymbol{u}_{k-1})$ is given by the activation function φ. The Euclidean-norm cost function $C = \frac{1}{2}\|\boldsymbol{u}_N - \boldsymbol{y}_N\|^2$, which compares the label layer activity to some target $\boldsymbol{y}_N$, enters the Lagrangian with a scaling factor β.

Neuronal dynamics are derived from a least-action principle, i.e, from the requirement of the action being stationary: $\delta \int \mathcal{L} \mathrm{d}t \overset{!}{=} 0$. The solution is provided by the Euler-Lagrange equations $\left(\frac{\partial}{\partial \tilde{\boldsymbol{u}}} - \frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial}{\partial \dot{\tilde{\boldsymbol{u}}}}\right)\mathcal{L} = 0$ and yields

$$\tau \dot{\boldsymbol{u}}_k = -\boldsymbol{u}_k + \boldsymbol{W}_k \boldsymbol{r}_{k-1} + \boldsymbol{e}_k \ , \tag{2}$$

wherein leaky integrator dynamics are easily recognizable. Two components in this equation are particularly relevant for the realization of phase-free error backpropagation. First, the neuronal activity $\boldsymbol{r}_{k-1} = \bar{\boldsymbol{r}}_{k-1} + \tau\dot{\bar{\boldsymbol{r}}}_{k-1}$ is a high-pass, nonlinearly filtered version of the respective membrane potential that effectively undoes the low-pass filtering induced by the leaky-integrator membrane dynamics

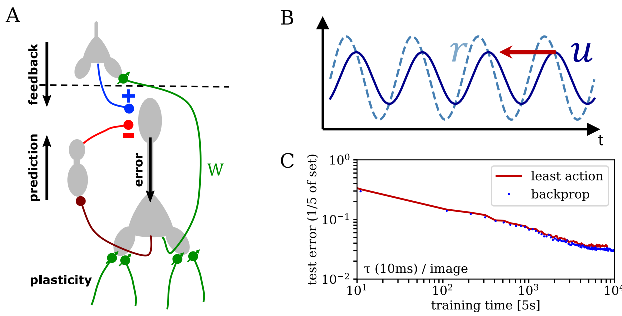

Figure 1: Neuronal least action principle. A) Cortical microcircuit used for error calculation and representation. Backpropagated errors are locally calculated by substracting bottom-up prediction from top-down feedback. B) The neuronal activity r is a non-linearly advanced version of the respective membrane potential that effectively undoes low-pass filtering caused by leaky integration. C) A network of 786 - 800 - 10 neurons learns MNIST from a continuous movie, where each digit is only shown briefly (time scale of the membrane time constant τ).

(fig. 1B). Biological neurons can, for example, implement this mechanism by non-linear sodium channel dynamics that depend both on $\boldsymbol{u}$ and $\dot{\boldsymbol{u}}$ [11]. This time advancement implements prospective coding on infinitesimal timescales, enabling neurons to forward-propagate their own future state, effectively guaranteeing instantaneous propagation of inputs that are smooth in time up to the label layer. Second, the error term $\boldsymbol{e}_k = \bar{\boldsymbol{e}}_k + \tau \dot{\bar{\boldsymbol{e}}}_k$ is a similarly time-advanced layerwise prediction error $\bar{\boldsymbol{e}}_k = \bar{\boldsymbol{r}}'_k \odot \boldsymbol{W}^{\mathrm{T}}_{k+1}(\boldsymbol{u}_{k+1} - \boldsymbol{W}_{k+1}\bar{\boldsymbol{r}}_k)$, which can be realized with a stereotypical microcircuit using pyramidal neurons and interneurons [3, 14], see fig. 1A. Defining plasticity as gradient descent on the Lagrangian, we obtain a biologically plausible and, in particular, fully local plasticity rule [19] that acts on the backpropagated error signal:

$$\dot{\boldsymbol{W}}_k \propto -\nabla_{\boldsymbol{W}} \mathcal{L} = (\boldsymbol{u}_k - \boldsymbol{W}_k \bar{\boldsymbol{r}}_{k-1})\bar{\boldsymbol{r}}^T_{k-1} \, . \tag{3}$$

The above equations specify a full model of real-time error backpropagation in cortical circuits, where plasticity can be shown to perform gradient descent on the cost function at every point in time [3]. We demonstrate this in a scenario where the network is exposed to a continuous stream of MNIST digits, reaching competitive classification results while the network dynamics are never close to stationarity during learning (fig. 1C).

2.2 Learning without a teaching signal

To remove the teaching signal $\boldsymbol{y}_N$, we combine the NLA principle with a reinforcement learning paradigm [17]. We associate K output neurons to actions and the input with the current state of the environment. The challenge lies in finding a mechanism that can give rise to a meaningful error signal for learning, while still harmonizing with error transport in the NLA framework. To this end, we extend the original NLA framework with lateral interactions in the last layer resembling soft winner-take-all structures. We postulate the cost function

$$C_{\mathrm{RL}} = \boldsymbol{M} \int_{-\infty}^{u} \bar{\boldsymbol{r}}(\hat{u})\, \mathrm{d}\hat{u} \, , \tag{4}$$

with the lateral interaction matrix $m_{ii} = 1$, $m_{ij} = -\frac{1}{K-1}$, which leads to recurrent dynamics in the output layer of the network: $\tau \dot{\boldsymbol{u}}_N = \boldsymbol{W}_N \boldsymbol{r}_{N-1} - \boldsymbol{u}_N + \beta \boldsymbol{M} \boldsymbol{r}_N$. The error in the last layer $\boldsymbol{e}_N =$

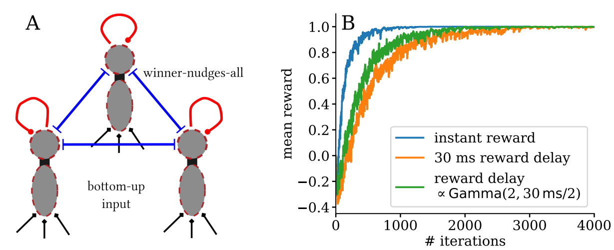

Figure 2: Reinforcement learning in the NLA framework. A) Lateral somato-somatic interaction with self-excitation and mutual inhibition gives rise to an error nudging approximating policy gradient. B) The network successfully learns on a small time-continuous classification problem based on 3 MNIST digits. Learning is robust in the presence of delayed rewards, even if the reward delay is stochastic.

$\beta \boldsymbol{M} \boldsymbol{r}_N$ nudges the chosen (winner) action positively and all other actions negatively (fig. 2A). The microcircuit structure (fig. 1A) then propagates the errors to the lower layers. Using an eligibility trace and plasticity modulation via the reward prediction error $[R(t) - \langle R \rangle]$, which represents a single global signal, we form a three-factor learning rule [4]:

$$\dot{\boldsymbol{W}}_k \propto \frac{1}{\tau_{\mathrm{elig}}} \left(R(t) - \langle R \rangle\right) \int_{-\infty}^{t} \boldsymbol{\kappa}_k(\hat{t}) \exp\left(-\frac{t-\hat{t}}{\tau_{\mathrm{elig}}}\right) \mathrm{d}\hat{t} \tag{5}$$

with $\boldsymbol{\kappa}_k(t) = (\boldsymbol{u}_k - \boldsymbol{W}_k \bar{\boldsymbol{r}}_{k-1})\bar{\boldsymbol{r}}^T_{k-1}$. Using $\boldsymbol{e}_N$ as a target error realizes hill-climbing on the mean expected reward as can be shown by comparison to direct policy gradient [22]. Such a network successfully learns with both immediate and delayed rewards from a continuous stream of inputs in a classification scenario (fig. 2).

3 SUMMARY

We show how real-time error backpropagation with and without an external supervisor can be implemented in a biologically plausible architecture. Our normative framework creates a bridge from the simple, but powerful least-action principle to the detailed morphology and physiology of a cortical circuitry model. An essential feature of this model is that the key requirements for its functionality can be realized by mechanisms available to both brain and brain-inspired hardware. Both forward and backward information streams, required for computing inference and errors, happen simultaneously in the network, relying on time-continuous and local dynamics. By utilizing prospective coding implemented through look-ahead neuronal responses, the framework avoids a separation of timescales or of dynamical phases. Unlike other recently developed algorithms that train deep networks with so-called synthetic gradients [9, 10], our framework backpropagates the true error generated at the output layer at all times.

ACKNOWLEDGMENTS

This work has received funding from the European Union under grant agreements 720270, 785907 (HBP) and the Manfred Stark Foundation. Calculations were performed on UBELIX (University of Bern HPC) and bwHPC (state BaWu HPC, supported by the DFG through grant no INST 39/963-1 FUGG).

REFERENCES

[1] Filipp Akopyan, Jun Sawada, Andrew Cassidy, Rodrigo Alvarez-Icaza, John Arthur, Paul Merolla, Nabil Imam, Yutaka Nakamura, Pallab Datta, Gi-Joon Nam, et al. 2015. Truenorth: Design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 34, 10 (2015), 1537–1557.

[2] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. 2018. Loihi: A neuromorphic manycore processor with on-chip learning. *IEEE Micro* 38, 1 (2018), 82–99.

[3] Dominik Dold, Akos Ferenc Kungl, João Sacramento, Mihai Alexandru Petrovici, Kaspar Schindler, Jonathan Binas, Yoshua Bengio, and Walter Senn. 2019. Lagrangian dynamics of dendritic microcircuits enables real-time backpropagation of errors. In *Cosyne Abstracts 2019, Lisbon, Portugal*.

[4] Nicolas Frémaux and Wulfram Gerstner. 2016. Neuromodulated spike-timing-dependent plasticity, and theory of three-factor learning rules. *Frontiers in neural circuits* 9 (2016), 85.

[5] Simon Friedmann, Johannes Schemmel, Andreas Grübl, Andreas Hartel, Matthias Hock, and Karlheinz Meier. 2016. Demonstrating hybrid learning in a flexible neuromorphic hardware system. *IEEE transactions on biomedical circuits and systems* 11, 1 (2016), 128–142.

[6] Steve Furber. 2016. Large-scale neuromorphic computing systems. *Journal of neural engineering* 13, 5 (2016), 051001.

[7] Jordan Guerguiev, Timothy P Lillicrap, and Blake A Richards. 2017. Towards deep learning with segregated dendrites. *ELife* 6 (2017), e22901.

[8] Giacomo Indiveri, Bernabé Linares-Barranco, Tara Julia Hamilton, André Van Schaik, Ralph Etienne-Cummings, Tobi Delbruck, Shih-Chii Liu, Piotr Dudek, Philipp Häfliger, Sylvie Renaud, et al. 2011. Neuromorphic silicon neuron circuits. *Frontiers in neuroscience* 5 (2011), 73.

[9] Max Jaderberg, Wojciech Marian Czarnecki, Simon Osindero, Oriol Vinyals, Alex Graves, David Silver, and Koray Kavukcuoglu. 2017. Decoupled neural interfaces using synthetic gradients. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1627–1635.

[10] Jacques Kaiser, Hesham Mostafa, and Emre Neftci. 2018. Synaptic plasticity dynamics for deep continuous local learning. *arXiv preprint arXiv:1811.10766* (2018).

[11] Harold Köndgen, Caroline Geisler, Stefano Fusi, Jing Wang, and Hans-Rudolf Lüscher. 2008. The Dynamical Response Properties of Neocortical Neurons to Temporally Modulated Noisy Inputs In Vitro. *Cerebral Cortex* September (2008). https://doi.org/10.1093/cercor/bhm235

[12] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[13] Isabella Pozzi, Sander Bohté, and Pieter Roelfsema. 2018. A biologically plausible learning rule for deep learning in the brain. *arXiv preprint arXiv:1811.01768* (2018).

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

[15] Benjamin Scellier and Yoshua Bengio. 2017. Equilibrium propagation: Bridging the gap between energy-based models and backpropagation. *Frontiers in computational neuroscience* 11 (2017), 24.

[16] Sebastian Schmitt, Johann Klähn, Guillaume Bellec, Andreas Grübl, Maurice Guettler, Andreas Hartel, Stephan Hartmann, Dan Husmann, Kai Husmann, Sebastian Jeltsch, et al. 2017. Neuromorphic hardware in the loop: Training a deep spiking network on the brainscales wafer-scale system. In *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2227–2234.

[17] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[18] Chetan Singh Thakur Thakur, Jamal Molin, Gert Cauwenberghs, Giacomo Indiveri, Kundan Kumar, Ning Qiao, Johannes Schemmel, Runchun Mark Wang, Elisabetta Chicca, Jennifer Olson Hasler, et al. 2018. Large-scale neuromorphic spiking array processors: A quest to mimic the brain. *Frontiers in neuroscience* 12 (2018), 891.

[19] Robert Urbanczik and Walter Senn. 2014. Learning by the dendritic prediction of somatic spiking. *Neuron* 81, 3 (2014), 521–528.

[20] James CR Whittington and Rafal Bogacz. 2017. An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity. *Neural computation* 29, 5 (2017), 1229–1262.

[21] James CR Whittington and Rafal Bogacz. 2019. Theories of error back-propagation in the brain. *Trends in cognitive sciences* (2019).

[22] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3-4 (1992), 229–256.