Gradient-based methods for spiking physical systems

J. Göltz[⋆1,2], **S. Billaudelle**[⋆1], **L. Kriener**[⋆2], **L. Blessing**[1], **C. Pehle**[1], **E. Müller**[1], **J. Schemmel**[1], **M. A. Petrovici**[2]

[⋆] *equal contributions* [1] *Kirchhoff-Insitute for Physics, Heidelberg University* [2] *Department of Physiology, University of Bern*

Summary Recent efforts have fostered significant progress towards deep learning in spiking networks, both theoretical and in silico. Here, we discuss several different approaches, including a tentative comparison of the results on BrainScaleS-2, and hint towards future such comparative studies.

Introduction Physical computation directly exploits the intrinsic dynamics of a given substrate to efficiently process and propagate information. In contrast to numerical computers, physical computers implicitly obey the dynamics required by certain models of information processing (e.g., neuronal integration) rather than calculating them explicitly by arithmetically manipulating binary representations thereof. Neuromorphic computers represent a prominent class of physical systems, drawing inspiration from the nervous system by mimicking the dynamics of neurons and synapses. They typically involve a massively parallel and time-continuous implementation of neuro-synaptic dynamics as well as an asynchronous event-based propagation of signals to efficiently emulate spiking neural networks (SNNs).

Physical systems are typically "programmed" by tuning their internal dynamics, such as time constants or coupling strengths. In our case, we employ gradient-based optimization schemes to adapt the emulated SNNs to a given task. Here, we discuss multiple approaches which all have been demonstrated on the mixed-signal neuromorphic system BrainScaleS-2 [6]. With this demonstration they have proven to address both the problems arising from the event-based characteristics of SNNs in general and the analog nature of the substrate in particular.

In-the-loop training of physical systems With self-adapting, local neuromorphic learning still in its infancy, we sometimes take inspiration from machine learning when training our neuromorphic devices, in particular from gradient-based optimization. To adopt these for physical computation, we need *differentiable estimates* of the system-internal dynamics. In our case of a time-continuous spiking neuromorphic system, this *model* has to capture the propagation and weighting of spikes as well as the neuronal dynamics. This could be realized as a collection of closed-form expressions for spike times or through a computation graph describing the propagation of stimuli through the network. An exact match between model and system dynamics is, however, often unattainable, and in a trade-off between model fidelity and complexity, higher-order effects are typically neglected. Instead, a continuous synchronization between model and system can mitigate the propagation of incorrect estimates. For this purpose, gradients

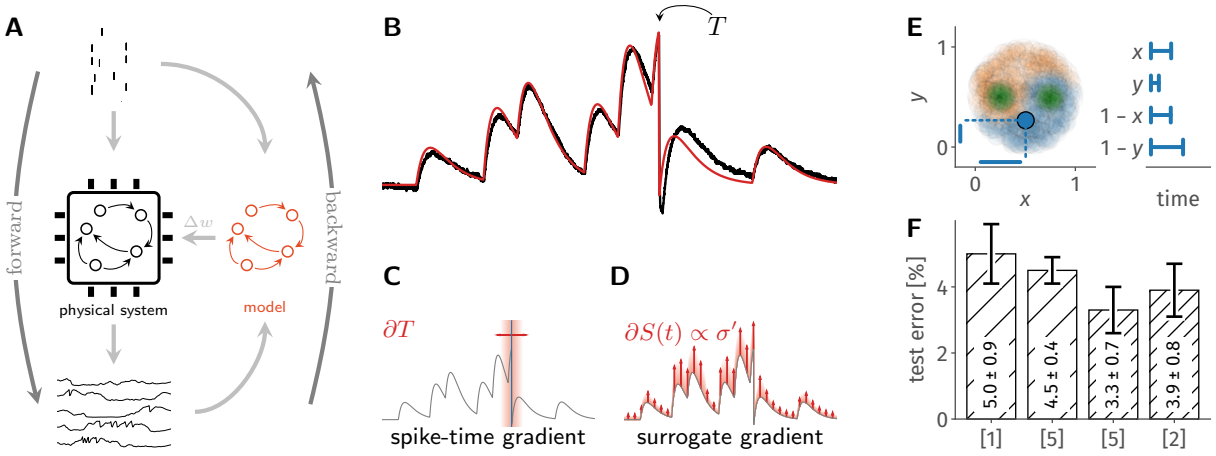

Figure 1: **A** In-the-loop approaches evaluate the physical system in the forward pass based on input spikes (raster plot, top). Combined with recorded observables (e.g., voltages and spikes, bottom), a model of the internal dynamics is then used to estimate gradients for weight update calculations. **B** Example data recorded from BrainScaleS-2 (black), as well as a simulated model trace replicating the core dynamics of the silicon neuron (red). **C** Spike-time gradients [1, 2] calculate the spike times' derivatives ∂T w.r.t. changes in presynaptic spike times or afferent weights. This quantity is only defined at the spike times and thus extremely sparse in time. **D** Surrogate gradients [3] estimate how a parameter change would affect the 'likeliness' of a spike at each point in time by replacing the 'hard' threshold θ by the smooth surrogate σ. **E** The Yin-Yang data set [4] and a corresponding temporal encoding of a randomly chosen sample. **F** Comparison of the three methods [1, 2, new results for 5]. Sparse and dense hatching indicate time-to-first-spike and voltage-based output decoding, respectively.

are estimated based on *measurements of observables*, including spike times or even membrane potential traces. Generally speaking, precise knowledge of the internal state can often be traded against model fidelity and vice versa.

Gradient calculation While the time-continuous and highly non-linear nature of SNNs impedes a straight-forward gradient descent, there exist multiple approaches for the estimation of gradients. These can be coarsely divided into methods either involving exact derivatives or resorting to inherent approximations of the spike triggering threshold.

Exact, sparse gradients can be obtained for a leaky integrate-and-fire (LIF) neuron's output spike times with respect to both input weights and presynaptic spike times. Analytical expressions for the gradients can – under certain assumptions – be derived based on differentiable expression for the spike time T as a function of only the input spikes and weights $T(\{t_i\}, \{w_i\})$ [1]. The derivatives of this function allow assignment of credit through multiple layers, and consequently gradient descent in deep networks. At the expense of an analytical solution but relaxing some assumptions, gradients can also be computed by relying on a backward evolution of adjoint dynamics [7]. Both approaches give exact relations on how to change weights to shift the *existing* spikes to reduce the loss function.

Surrogate gradients [3] offer an alternative approach by considering the output spike train $S(t) = \sum_i \delta(t - T_i)$ of a neuron, where the individual spike times are given by T_i. In order to estimate useful gradients not only at the individual output spike times, $\partial S(t)$ can be *approximated* with the help of a surrogate derivative $\sigma'(v(t))$ based on the membrane potential $v(t)$. In contrast to exact, spike-time-based approaches, this method also assigns gradients at times where no spikes occur. While leading to a potential memory and computation overhead, this approach allows explicit awareness of the creation or deletion of spikes and can hence innately train networks from a quiescent state by specifically recruit neuronal activity where required.

Results The three approaches outlined above have all been demonstrated on the BrainScaleS-2 system. Figure 1F shows respective results obtained for the Yin-Yang dataset [4]. All three methods are able to successfully solve the task and yield comparable classification accuracies. Classifiers based on a time-to-first-spike output seem to incur a slight performance penalty in comparison to voltage-based outputs. Further, approaches based on exact spike-time gradients appear to yield accuracies slightly below the ones reached by surrogate gradient methods. However, inhomogeneous choices of system- and hyperparameters spoil a direct comparison across publications. In addition to the presented results, BrainScaleS-2 has been trained on a variety of other datasets, with both feedforward and recurrent network topologies [1, 5]. The respective training methods were shown to be robust against parameter noise that was artificially induced to mimic fixed-pattern deviations not uncommon in physical systems.

Discussion The capability to train highly complex, nonlinear dynamical systems, not just in idealized simulations, but also in practice, represents a fundamental prerequisite for the real-world deployment of neuromorphic devices. Here, we have reviewed three recently demonstrated methods and their results on BrainScaleS-2 [1, 2, 5]. This demonstration is a testament to the maturity of this system in particular as well as to the progress of physical computation in general. It also paves the way for important future studies including comparisons on a variety of performance metrics such as convergence speed during training, data efficiency, robustness to noise and parameter changes.

Acknowledgements This work received funding from the EU research and innovation funding H2020 945539 (HBP SGA3), DFG project EXC 2181/1390900948 (STRUCTURES), and the Manfred Stärk Foundation.

References

[1] J. Göltz, L. Kriener, et al. "Fast and energy-efficient neuromorphic deep learning with first-spike times". In: *Nature Machine Intelligence* 3.9 (2021).

[2] C. Pehle, L. Blessing, et al. "Event-based Backpropagation for Analog Neuromorphic Hardware". In: *preprint arXiv:2302.07141* (2023).

[3] E. O. Neftci, H. Mostafa, and F. Zenke. "Surrogate gradient learning in spiking neural networks". In: *IEEE Signal Processing Magazine* 36.6 (2019).

[4] L. Kriener, J. Göltz, and M. A. Petrovici. "The Yin-Yang Dataset". In: *Neuro-Inspired Computational Elements Conference*. NICE 2022. Association for Computing Machinery, 2022.

[5] B. Cramer, S. Billaudelle, et al. "Surrogate gradients for analog neuromorphic computing". In: *Proceedings of the National Academy of Sciences* 119.4 (2022).

[6] C. Pehle, S. Billaudelle, et al. "The BrainScaleS-2 Accelerated Neuromorphic System with Hybrid Plasticity". In: *Front. Neurosci.* 16 (2022).

[7] T. C. Wunderlich and C. Pehle. "Event-based backpropagation can compute exact gradients for spiking neural networks". In: *Scientific Reports* 11.1 (2021).