TeDFA-δ: Temporal integration in deep spiking networks trained with feedback alignment improves policy learning

Jorin Overwiening (joverwiening@fas.harvard.edu)
Center for Brain Sciences, Harvard University
Cambridge, MA US

M Ganesh Kumar (mganeshkumar@seas.harvard.edu)
SEAS, Harvard University
Cambridge, MA US

Haim Sompolinsky (hsompolinsky@mcb.harvard.edu)
Edmond and Lily Safra Center for Brain Sciences, Hebrew University of Jerusalem
Jerusalem, Israel
and
Center for Brain Sciences, Harvard University
Cambridge, MA US

Abstract

Limitations in deep spiking reinforcement learning models hinder our understanding of how biological systems learn control policies. We address this by developing a biologically plausible deep reinforcement learning agent (TeDFA-δ) that combines spiking neurons with local Tempotron learning and global Direct Feedback Alignment and Temporal Difference error optimization. Despite using a suboptimal learning rule, TeDFA-δ outperforms backpropagation-trained MLPs on cartpole, acrobot, and dynamic bandit tasks. This improvement stems from temporal integration of states in spiking neurons rather than the learning algorithm itself, based on ablation studies. The network develops structured spatiotemporal representations where policy and value information coexist, with optimal performance at intermediate membrane time constants ($\tau \ll T$). Our results demonstrate that biological systems may compensate for imperfect credit assignment through temporal dynamics, suggesting neural representations outweigh learning rule optimality for control tasks. This framework enables new studies of biological learning while advancing neuromorphic computing.

Keywords: deep spiking networks; reinforcement learning

Introduction

Reinforcement learning (RL) in biological systems involves trial-and-error learning, where the ventral striatum computes value estimates, the dorsal striatum mediates stimulus-response associations, and dopaminergic neurons signal reward prediction errors (Kumar, Bordelon, Zavatone-Veth, & Pehlevan, 2024). While artificial RL algorithms often rely on rate-coded neurons, biological neurons communicate via spikes, but deep spiking neural networks (SNNs) remain limited in solving complex tasks (Kumar, Tan, Libedinsky, Yen, & Tan, 2022) due to the lack of biologically plausible credit assignment mechanisms (Neftci, Mostafa, & Zenke, 2019). Here, we develop a deep SNN with local Tempotron learning (Shi et al., 2021; Gütig & Sompolinsky, 2006) trained with a biologically plausible deep learning algorithm: Direct Feedback Alignment (DFA) (Nøkland, 2016) modulated by temporal difference error (Kumar et al., 2024). We demonstrate that it learns useful representations to solve complex control tasks, outperforming multi-layered perceptrons (MLPs) trained via backpropagation. We hypothesize that spiking neurons' temporal integration compensates for DFA's noisy error signals and also helps stabilize online temporal difference error, enabling efficient learning. We propose this framework as a basis for understanding how biological systems learn control faster than artificial ones.

Methodology

Spiking Network Model

We use a multi-layer Tempotron spiking neuron model (Gütig & Sompolinsky, 2006), where each neuron's membrane potential $V(t) = \sum_i \omega_i \phi_i(t - t_i)$ integrates postsynaptic potentials (PSPs) from incoming spikes $\phi_i(t) = \sum_{t_i} \left(e^{-(t-t_i)/\tau} - e^{-(t-t_i)/\tau_s} \right)$ with time constants τ and τ_s, and t_i incoming spike times at synapse i. A spike is emitted when $V(t) \geq V_\theta$. We extend this to deep networks using Direct Feedback Alignment (DFA) (Nøkland, 2016; Shi et al., 2021) through the proposed TempotronDFA (TeDFA) (Overwiening & Sompolinsky, 2025), where weight updates combine global error signals e_k^L with local temporal activity ϕ using fixed random feedback matrices F_{jk}^l:

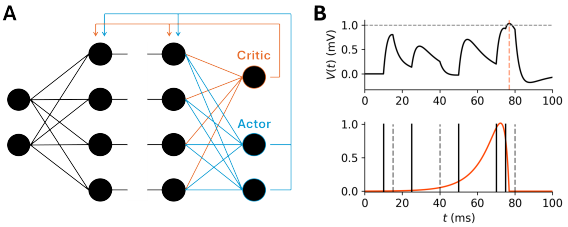

Figure 1: **A** The Tempotron actor critic model. The feedback paths from actor and critic neurons go to hidden layer neurons via the TeDFA-δ learning rule. **B** The Tempotron neuron with a sample trace (top) and the corresponding Tempotron learning rule (bottom, orange).

$$\Delta\omega_{ji}^l \propto -(F_{jk}^l \cdot e_k^L) \odot \phi_{ij}(t_{\max}^j - t_i^j) \tag{1}$$

Policy and Value Learning Error

To learn to estimate value and a suitable policy, we optimize the Temporal Difference (TD) error following (Kumar et al., 2024), but replace backpropagation with biologically plausible feedback:

$$Fe_t = F^\pi e_t^\pi + F^v e_t^v = F^\pi(\tilde{g}_t \cdot \delta_t) + F^v \delta_t \tag{2}$$

where $\tilde{g}_t$ is a normalized one-hot action vector, and F^π, F^v are separate fixed random feedback matrices for actor and critic pathways (Fig. 1A). This enables online credit assignment for spiking networks through fixed random projections, similar to dopaminergic pathways in the brain.

Implementation Details

Simulations use time windows ($T = 50$ ms) matching environment dynamics, with $\tau = 10$, $\tau_s = \tau/3$. Input states drive the first hidden layer via linearly increasing currents $V(t) = \sum_i \omega_i x_i \cdot t$, where x is the state value for this environment time step, while outputs (actions and value) are read from maximum membrane potentials $V_{\max}$ per window. All tasks use 2 hidden layers and 128 neurons per hidden layer for all models. Learning rates were 0.001 for MLP with BP and TeDFA, but a lower learning rate of 0.0001 was required for MLP with DFA.

Results

Cartpole Task Performance We first evaluate our model (TeDFA-δ) on the challenging cartpole task, where small pol-

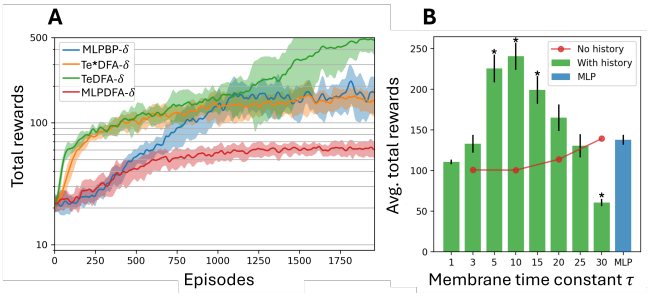

Figure 2: **A** The Tempotron model learns faster and outperforms a perceptron model on the online cartpole task. Shown are the total rewards averaged over 20 runs with standard deviations. TeDFA-δ is the full model with history, Te*DFA-δ is without state history. **B** TeDFA-δ total rewards averaged over 10 runs (after 1000 episodes) on cartpole for different values of τ, with and without history. * denotes significance with MLP performance ($p < 0.05$). Shaded areas and error bars are 95% CI.

icy changes can lead to failure and online learning is difficult due to long trial durations (500 steps). TeDFA-δ achieves near-perfect performance after 1500 episodes, outperforming both MLP with backpropagation (BP) and a memory-less variant (Te*DFA-δ) with membrane potential resets after each step (Fig. 2A). The 10x faster convergence of both spiking models suggests temporal integration is crucial - confirmed by the performance drop when using larger time constants ($\tau \sim T$) that make the model behave more like an MLP (Fig. 2B).

Learned Representations Analysis of hidden layer activity reveals TeDFA-δ develops structured representations where policy and value information coexist in feature space (Fig. 3). The model learns circular representations of cartpole states and prepares future actions through membrane potential history. Trajectories often begin/end in appropriate subspaces, suggesting memory of past states enhances performance beyond MLP's capabilities which uses instantaneous state representation. PC projection is done into a global spatial autocorrelation space, so that the distance to the origin is a metric of explained feature-variance, which is reached after some time in each world time window (see Fig. 3B).

Generalization to Other Tasks In the acrobot task, TeDFA-δ matches MLP-BP performance but learns faster (Fig. 4A). For the 10-armed bandit task, TeDFA-δ shows superior adaptability when reward distributions change every 1000 trials, maintaining performance while MLP-BP struggles to relearn (Fig. 4B). The memory-less Te*DFA-δ performs similarly to MLP, confirming that temporal integration enables more robust learning across changing conditions.

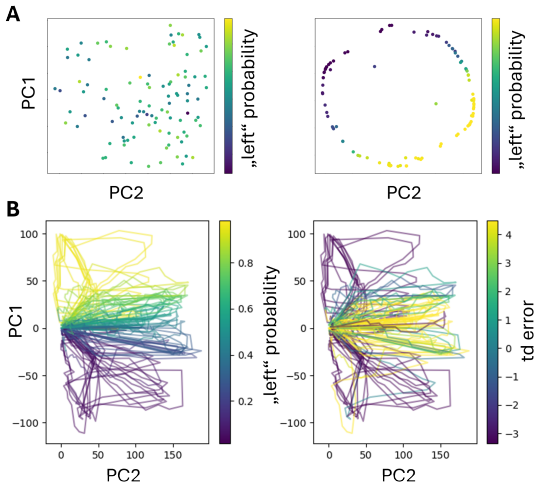

Figure 3: The Tempotron model learns useful policy and value representation. **A** Features for 100 sample inputs before (left) and after (right) training. **B** Trajectory of the model for one full sample episode in cartpole after 100 episodes of training.

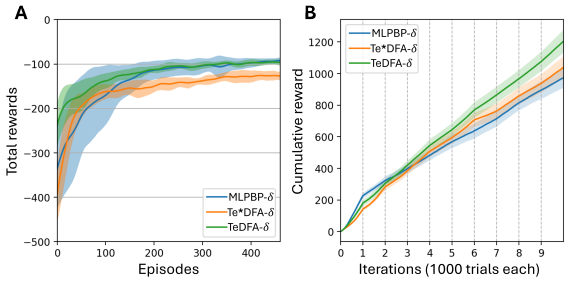

Figure 4: **A** Performances on the acrobot task. **B** Performances on the continuous k-armed bandit task. Averaged accumulated rewards over 100 runs with each run of 10 iterations with different randomly sampled reward probabilities with 1000 trials per iteration. Shaded areas are 95% CI.

Conclusion

We present a biologically plausible deep spiking model for online reinforcement learning that, despite DFA's constraints, outperforms backpropagation-trained MLPs through temporal integration and improved representation learning. This suggests biological networks may compensate for weaker local signals with temporal dynamics, enabling faster and more efficient learning. The model enables advances in challenging control tasks and energy-efficient neuromorphic hardware (T. Wang et al., 2025). Future work will explore additional tasks, like continuous control and more challenging problems, and a comparison to other models with history or memory aspects. Additionally, we will investigate scaling laws for this and similar models (K. Wang, Javali, Bortkiewicz, Trzciński, & Eysenbach, 2025) and assess biological alignment of learned representations to make testable predictions for neuromodulators (Kumar, Manoogian, Qian, Pehlevan, & Rhoads, 2025).

Acknowledgments

We acknowledge the support of the Swartz Foundation, the John A. Paulson School of Engineering and Applied Sciences at Harvard University, and the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University.

References

Gütig, R., & Sompolinsky, H. (2006). The tempotron: a neuron that learns spike timing–based decisions. *Nat Neurosci*, *9*, 420-428. doi: https://doi.org/10.1038/nn1643

Kumar, M. G., Bordelon, B., Zavatone-Veth, J. A., & Pehlevan, C. (2024). A model of place field reorganization during reward maximization. *bioRxiv*. doi: 10.1101/2024.12.12.627755

Kumar, M. G., Manoogian, A., Qian, B., Pehlevan, C., & Rhoads, S. A. (2025). Neurocomputational underpinnings of suboptimal beliefs in recurrent neural network-based agents. *bioRxiv*, 2025–03.

Kumar, M. G., Tan, C., Libedinsky, C., Yen, S.-C., & Tan, A. Y. (2022). A nonlinear hidden layer enables actor–critic agents to learn multiple paired association navigation. *Cerebral Cortex*, *32*(18), 3917–3936.

Neftci, E. O., Mostafa, H., & Zenke, F. (2019). Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, *36*(6), 51–63.

Nøkland, A. (2016). *Direct feedback alignment provides learning in deep neural networks.* Retrieved from `https://arxiv.org/abs/1609.01596`

Overwiening, J., & Sompolinsky, H. (2025). *Tempotron learning in deep neural networks.* (in preparation)

Shi, C., Wang, T., He, J., Zhang, J., Liu, L., & Wu, N. (2021). Deeptempo: A hardware-friendly direct feedback alignment multi-layer tempotron learning rule for deep spiking neural networks. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *68*(5), 1581-1585. doi: 10.1109/TCSII.2021.3063784

Wang, K., Javali, I., Bortkiewicz, M., Trzciński, T., & Eysenbach, B. (2025). *1000 layer networks for self-supervised rl: Scaling depth can enable new goal-reaching capabilities.* Retrieved from `https://arxiv.org/abs/2503.14858`

Wang, T., Tian, M., Wang, H., Zhong, Z., He, J., Tang, F., ... Shi, C. (2025). Morphbungee: A 65-nm 7.2-mm2 27-µj/image digital edge neuromorphic chip with on-chip 802-frame/s multi-layer spiking neural network learning. *IEEE Transactions on Biomedical Circuits and Systems*, *19*(1), 209-225. doi: 10.1109/TBCAS.2024.3412908