Temporal difference modulated spiking actor learning

Yunes Tihomirov[1], Roman Rybka[2], Alexey Serenko[2], Alexander Sboev[2]

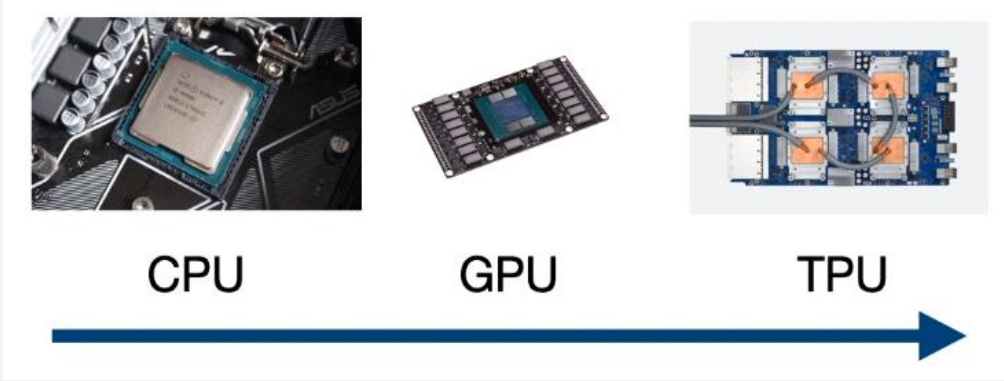
[1]National Research University Higher School of Economics, Moscow

[2]National Research Centre "Kurchatov Institute“, Moscow

Problem and relevance

- The problem is the development of an energy-efficient method for dynamic control in a time-varying environment for autonomous devices.
- One possible solution is to use neuromorphic devices together with spiking neural networks implemented on them.

von Neumann architecture



Neuromorphic computations

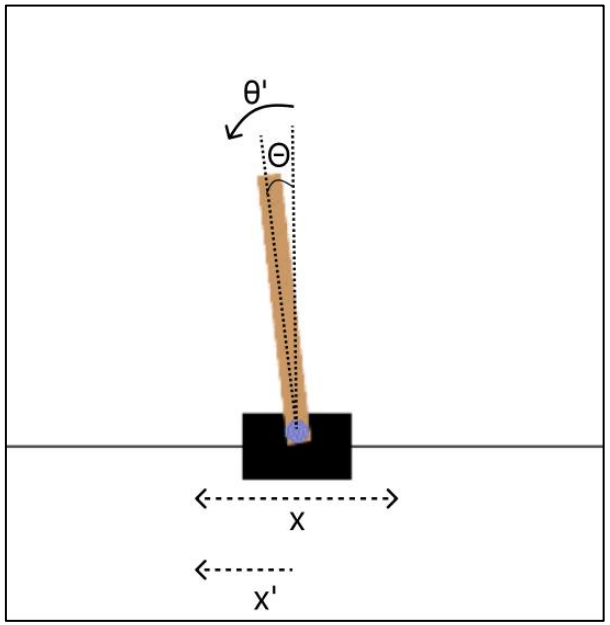
Objective and tasks

- The objective of this work is to develop a method for solving reinforcement learning problems based on local plasticity learning rules.
- The tasks are to select the spiking neural network architecture, the learning rules for the spiking neural network, the benchmarks and metrics for comparing solution quality.

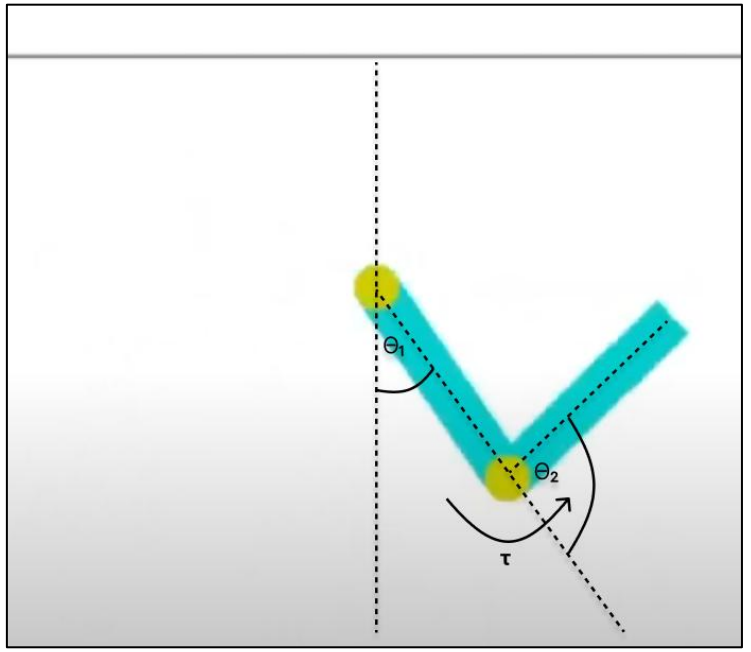
Related works

Approach	Features
Actor-critic with FBTDSTDP learning rule (Chung et al., 2020 [1])	Spiking actor is trained using feedback-modulated TD-STDP, spiking critic is trained using TD-STDP
Multilayer spiking neural network (Chevtchenko et al., 2023 [2])	Two-layer FEAST clustering of input states with temporal difference error modulation in a spiking actor-critic
R-STDP with formal critic (Aenugu et al., 2020 [3])	Critic is not spiking and is trained using TD(λ); spiking actor is trained using RSTDP
Actor-critic with TD-LTP learning rule (Fremaux et al., 2013 [4])	Actor and critic are spiking with TD-LTP learning rule

Reinforcement learning tasks

CartPole		
Goal	Balance the pole for 200 time steps by applying a unit horizontal force to the cart at each step	
Environment state	1. cart coordinate x 2. cart velocity $\dot{x}$ 3. Angle of the pole w.r.t. vertical θ 4. Angle velocity of the pole θ̇	
Reward	+1 for every step before the balance breaking	Fig. 1: Cartpole system

Reinforcement learning tasks

Acrobot		
Goal	Reach the top bar in under 500 time steps by applying, at each step, either a unit torque or zero torque to the joint between the links	
Environment state	1. angle Θ_1 2. angle Θ_2 3. Rate of change of angle Θ_1 4. Rate of change of angle Θ_2	
Reward	-1 for every step before reaching the upper bar и +10 in the end of successful episode	Fig. 2: Acrobot system

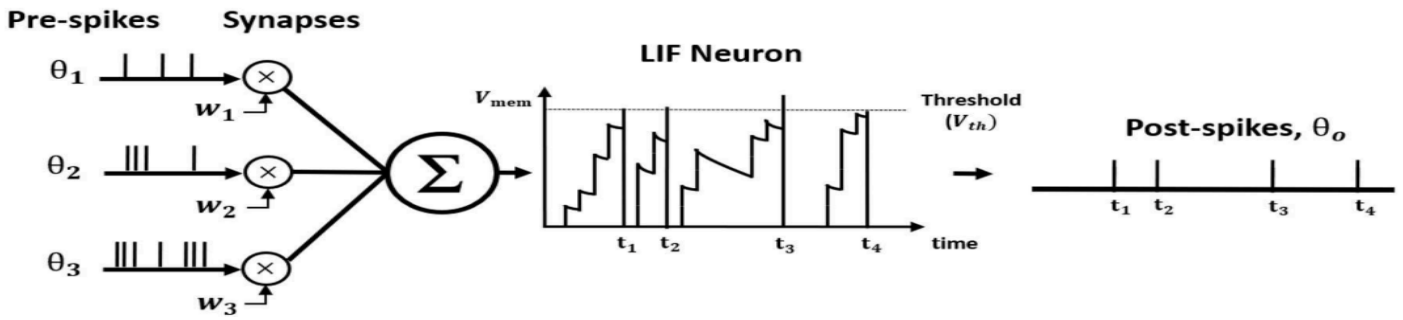


Fig. 3. Illustration of LIF neuron model

Neuron voltage dynamic for the **LIF (Leaky Integrate-and-Fire)** neuron model**:**

$$\boldsymbol{\tau \frac{du}{dt} = -(u(t) - u_{rest}) + RI(t)}$$

$$\text{if } u(t) = \theta \rightarrow \text{spike and reset}$$

Input impulses are stored in the membrane potential u, exponentially decaying with the rate τ. When the potential reaches the threshold θ, neuron emits spike and potential is set to the u_{rest}.

Neuron model

STDP - learning rule for the hidden layer of actor

STDP (Spike-timing dependent plasticity)

$$\begin{cases} \Delta w^{+} = A^{+} \cdot e^{-\frac{|\Delta t|}{\tau_{+}}} \\ \Delta w^{-} = -A^{-} \cdot e^{-\frac{|\Delta t|}{\tau_{-}}} \end{cases}$$

where $\boldsymbol{\Delta} t = t_i^f - t_j^f$ is the time difference between post- and presynaptic spikes, $\boldsymbol{\tau}_+$ and $\boldsymbol{\tau}_-$ are the time constants, $A^{\pm}(w)$ are the constant amplification and depression amplitudes.

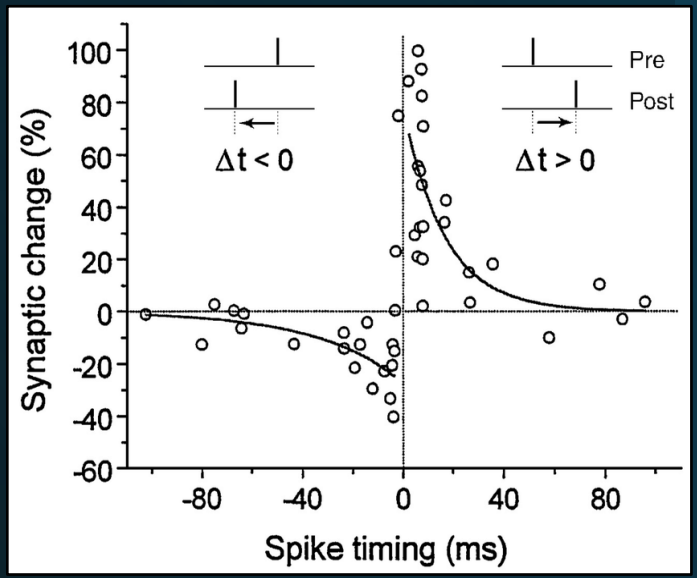


Fig.4. Illustration of spike timing dependent plasticity time windows

TD-STDP – learning rule for the output layer of actor

TD-STDP (STDP modulated by temporal difference error) learning rule is the following:

$$\frac{\boldsymbol{dw}}{\boldsymbol{dt}} \sim \boldsymbol{\delta_{TD}} \cdot \boldsymbol{e(pre, post)}$$

$$\dot{e} = -\frac{e}{\tau_e} + STDP(pre, post),$$

where $e(pre, post)$ is the eligibility trace keeping the history of spike activity, τ_e is the time constant of eligibility trace decay, pre and post are the spike trains of pre- and postsynaptic neurons respectively

Temporal difference error:

$$\boldsymbol{\delta_{TD}(t+1) = R_{t+1} + \gamma V(s_{t+1}) - V(s_t)},$$

where R_{t+1} is the reward from the environment during the transition from state s_t to s_{t+1}, γ is the discount factor, $V(s_t)$ is the value of state s_t

Network architecture

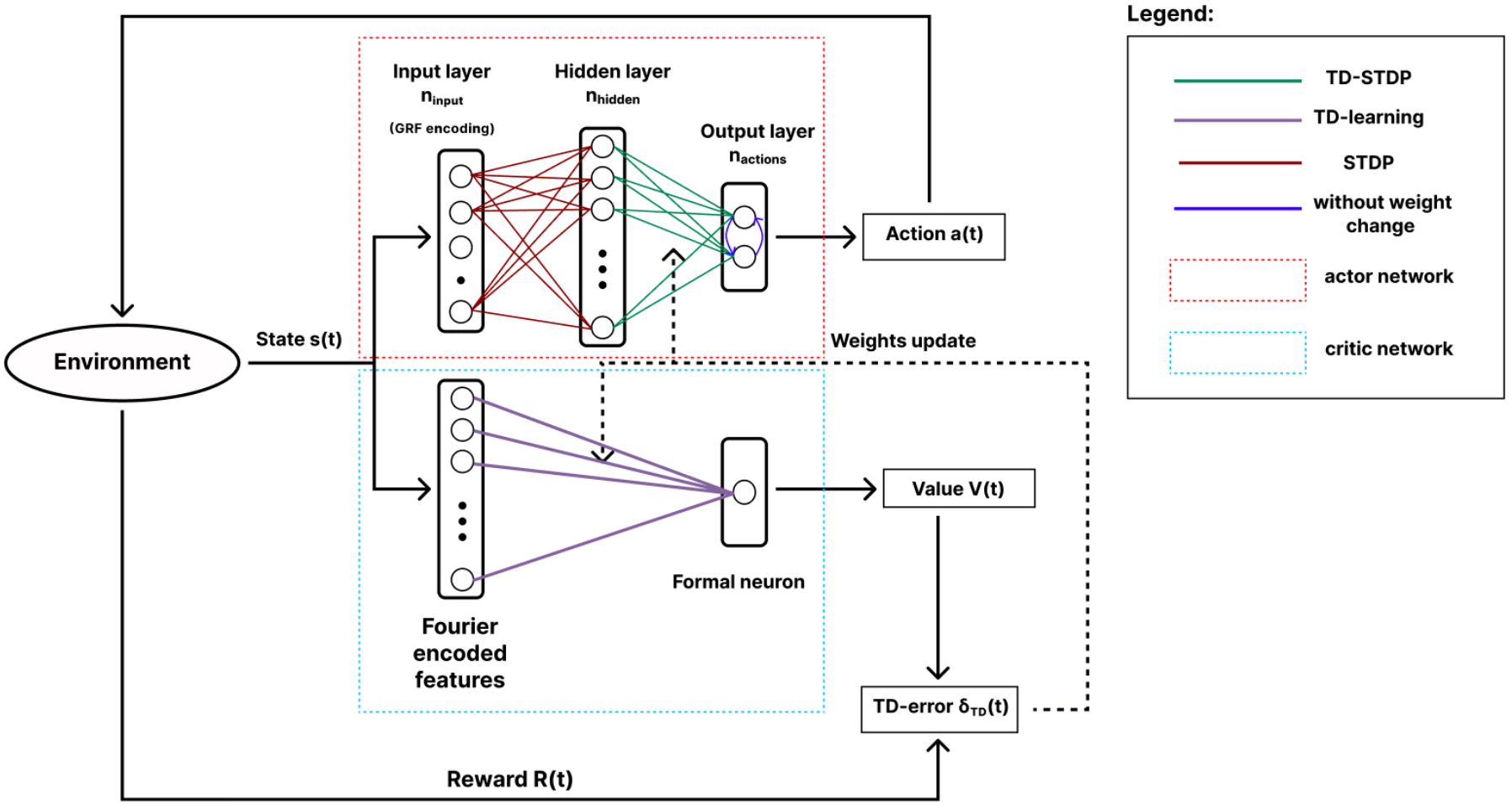


Fig. 5. Neural network diagram.

CartPole results

Fig. 6. Reward during training episodes of solving Cartpole-v0 task along with the running reward averaged over 50 episodes

Acrobot results

Fig. 7. Reward during training episodes of solving Acrobot-v1 task along with the running reward averaged over 10 episodes. Mean test reward is -301.

Results comparison

Approach	Description	Average test reward for Acrobot task	Number of training iterations for Cartpole task
Two-layer spiking actor with TD-STDP learning rule (current approach)	The first layer of actor is trained using STDP, the second - using TD-STDP, critic is not-spiking and is trained using TD-learning	**-301**	**715**
Actor-critic with FBTDSTDP learning rule (Chung et al., 2020 [1])	Spiking actor is trained using feedback-modulated TD-STDP, spiking critic is trained using TD-STDP	-	169.5
Multilayer spiking neural network (Chevtchenko et al., 2023 [2])	Two-layer FEAST clustering of input states with temporal-difference-error modulation in a spiking actor–critic	-128	400*
R-STDP with formal critic (Aenugu et al., 2020 [3])	Critic is not spiking and is trained using TD(λ); spiking actor is trained using RSTDP	-168	2000
Actor-critic with TD-LTP learning rule (Fremaux et al., 2013 [4])	Actor and critic are spiking with TD-LTP learning rule	-150	3500

* - approximate estimation of number of iterations from the plot in the paper

Conclusion

The novelty of this work lies in the development of a two-layer spiking actor trained using a three-factor TD-STDP rule, in which the global temporal-difference error modulates local weight updates. The proposed method was experimentally validated on the Acrobot and CartPole control environments, achieving goal attainment in Acrobot within 301 steps and convergence to a stable solution on CartPole after 715 training episodes – performance that is on par with other spiking neural network–based approaches.

Future work

- A key direction for further research is to train a spiking critic based on the outputs of a hidden layer trained with STDP. This will enable approach to solving a range of reinforcement learning tasks and reduce the number of synaptic connections that need to be trained.

References

1. Stephen Chung and Robert Thijs Kozma. Reinforcement learning with feedback-modulated td-stdp. ArXiv, abs/2008.13044, 2020.
2. Sergio F. Chevtchenko, Yeshwanth Bethi, Teresa B. Ludermir, and Saeed Afshar. A neuromorphic architecture for reinforcement learning from real-valued observations.2024.
3. Aenugu, S. et al. Reinforcement learning with a network of spiking agents. In NeurIPS20I9, 2019.
4. Fremaux, N. et al. Reinforcement learning using a continuous time actor-critic framework with spiking neurons. PLoS Comp. Biol., 2013.
5. Ji, X. et al. Reinforcement learning in memristive spiking neural networks through modulation of resume. In AIP Conf. Proc., 2019.
6. Wang, Z. et al. Reinforcement learning with analogue memristor arrays. Nature Electronics, 2, 115–124, 2019.
7. Vlasov, D. et al. Memristor-based spiking neural network with online reinforcement learning. Neural Networks 166, 512–523, 2023.

Thank you for your attention!