# SpikingGamma: Surrogate-Gradient Free and Temporally Precise Online Training of Spiking Neural Networks with Smoothed Delays

Roel Koopman(1), Sebastian Otte(2), and Sander Bohté(1)

(1) Machine Learning Group, CWI, (2) Institute of Robotics and Cognitive Systems, University of Lübeck

Neuromorphic hardware implementations of Spiking Neural Networks (SNNs) promise energy efficient, low-latency AI through sparse, event-driven computation. Yet, training SNNs under fine temporal discretization remains a major challenge, hindering both low-latency responsiveness and the mapping of software-trained SNNs to efficient hardware. In current approaches, spiking neurons are modeled as self-recurrent units, embedded into recurrent networks to maintain state over time, and trained with BPTT or RTRL variants based on surrogate gradients. These methods scale poorly with temporal resolution, while online approximations often exhibit instability for long sequences and tend to fail at capturing temporal patterns precisely. To address these limitations, we develop spiking neurons with internal recursive memory structures that we combine with sigma-delta spike-coding. We show that this SpikingGamma model supports direct error backpropagation without surrogate gradients, can learn fine temporal patterns with minimal spiking in an online manner, and scale feedforward SNNs to complex tasks and benchmarks with competitive accuracy, all while being insensitive to the temporal resolution of the model. Our approach offers both an alternative to current recurrent SNNs trained with surrogate gradients, and a direct route for mapping SNNs to neuromorphic hardware.

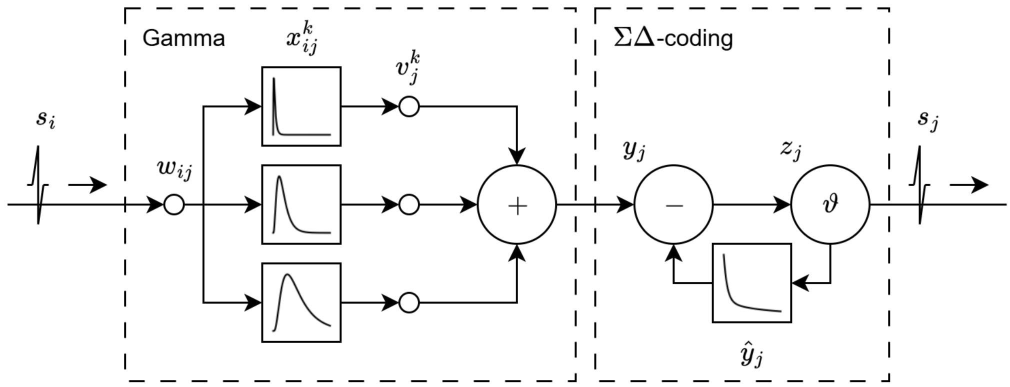


Figure 1: Overview of the neural processing model. At the synapse from neuron $i$ to $j$, incoming spikes generate weighted currents (with weight $w_{ij}$) that evolve over multiple timescales $k$. Within the neuron, the resulting synaptic responses ($x_{ij}^k$) are weighted according to their timescales (with weight $v_j^k$) and summed to produce a continuous neuronal signal ($y_j$). This signal is then converted back into spikes through ΣΔ-coding, allowing downstream synapses to reconstruct an estimate of the original signal ($\hat{y}_j$). This estimate enables gradients to flow directly from $\hat{y}_j$ to $y_j$, bypassing the spikes during the backward pass.