Convergence of Direct SNN Training with Surrogate Gradients

Zelin Wei[1] · **Tianxing Man**[2] · **Wanli Shi**[1] · **Bin Gu**[1]*

[1] School of Artificial Intelligence, Jilin University

[2] College of Computer Science, Inner Mongolia University

* Corresponding author

NEURAL INFORMATION PROCESSING SYSTEMS

Key message

Direct training of spiking neural networks (SNNs) uses hard spikes forward and surrogate derivatives backward, which biases the update.
We prove a finite-step bound on the hard-spike loss relative to a reference point and show how threshold proximity controls two error terms.
The same threshold statistic lets practitioners monitor hard–smooth mismatch directly from hard-forward membrane states.

01 Training Mismatch

Same derivative rule, different forward trajectories

Shared local derivative $g_\delta = h'_\delta = \mathbb{E}_z[G^2]$

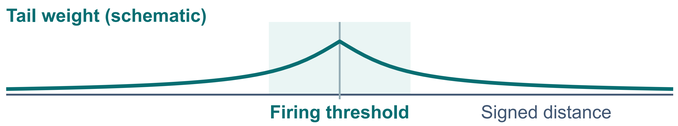
SG and normalized LocalZO match in mean with independent perturbations.

Hard spikes $h(u) = \mathbf{1}\{u > 0\}$	→ **Hard-path backward** g_δ	→ **Mean update** μ_k
Smoothed spikes $h_\delta(u)$	→ **Smooth-path backward** g_δ	→ **Reference gradient** $\nabla F_\delta(W_k)$

Trajectory-induced bias $b_k = \mu_k - \nabla F_\delta(W_k)$

Backpropagating along different forward paths induces the bias.

FIG. 2 Near-threshold states carry more weight



$$\mathcal{T}_\delta(W) = \mathbb{E}_\xi\left[\sum_{i=1}^{M} \Psi\left(\frac{|u_i^{\mathrm{hard}}(W;\xi)|}{\delta}\right)\right]$$

$\Psi(a) = \mathbb{P}(|\tilde{z}| \geq a)$ is the smoothing tail probability.

Conditions for Theorem 1

Smoothness of the analytical objective, bounded variance, and regular graph and loss maps. Propagation requires contraction or finite-horizon causal regularity. Restricted PL holds on the region containing the iterates.

02 Convergence Guarantee

FIG. 1 Convergence roadmap

Hard-forward training $\mu_k = \mathbb{E}[\hat{g}_k \mid \mathcal{F}_k]$

Smoothed reference $\nabla F_\delta(W_k)$

Threshold-tail $\mathcal{T}_\delta(k)$ — Near-threshold mass $\Psi(|u^{\mathrm{hard}}|/\delta)$

Trajectory-induced bias $b_k = \mu_k - \nabla F_\delta(W_k)$ $\quad \mathbb{E}\|b_k\|^2 \leq \Gamma^2 \Lambda_\delta^2 \mathcal{T}_\delta(k)$

Biased SGD + restricted PL — Transfer

Hard-spike objective: reference-point guarantee — Theorem 1

$$\mathbb{E}[F(W_R)] - F(\bar{W}) \leq \mathcal{E}_{\mathrm{opt}} + \mathcal{E}_{\mathrm{var}} + \mathcal{E}_{\mathrm{bias}} + \mathcal{E}_{\mathrm{gap}} + \mathcal{E}_{\mathrm{ref}}$$

Theorem 1 · Finite-step bound for hard spikes

$\mathbb{E}[F(W_R)] - F(\bar{W}) \leq$	Relative to the reference point	
$\frac{2\Delta_\delta}{\mu_{\mathcal{U}}\eta K}$	**Optimization** Finite-step descent	
$+ \frac{L_\delta \eta}{\mu_{\mathcal{U}}}\left(\sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m}\right)$	**Stochastic variance** Batch noise + LocalZO noise	
$+ \frac{3\Gamma^2\Lambda_\delta^2}{2\mu_{\mathcal{U}}} \bar{\mathcal{T}}_K$	**Gradient bias** Linear in the average tail	$\mathcal{E}_{\mathrm{bias}}$
$+ L_s \Lambda_\delta \sqrt{\bar{\mathcal{T}}_K}$	**Objective gap** Square root of the average tail	$\mathcal{E}_{\mathrm{gap}}$
$+ \left\lvert F(\bar{W}) - F_\delta(\bar{W}) \right\rvert$	**Reference smoothing** Hard–smooth gap at the reference	$\mathcal{E}_{\mathrm{ref}}$

$$\eta = \Theta(K^{-1/2}) \implies \mathcal{E}_{\mathrm{opt}} + \mathcal{E}_{\mathrm{var}} = O(K^{-1/2})$$

The three residuals can persist at fixed smoothing width.

$\Delta_\delta = F_\delta(W_0) - F_\delta(\bar{W})$ $\quad \bar{\mathcal{T}}_K = \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\mathcal{T}_\delta(W_k)]$

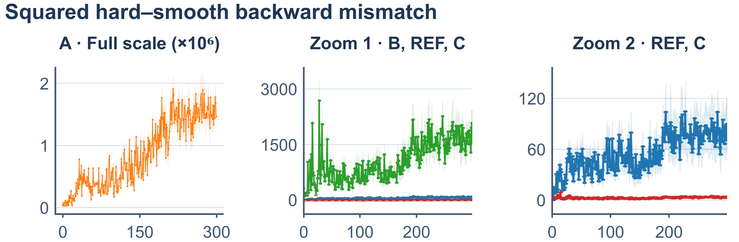
$W_k \in \mathcal{U}$ a. s., $\quad \bar{W} \in \arg\min_{W \in \mathcal{U}} F_\delta(W)$ $\quad R \sim \mathrm{Unif}\{0, \ldots, K-1\}$, $\quad \eta \leq (4L_\delta)^{-1}$

03 Experimental Evidence

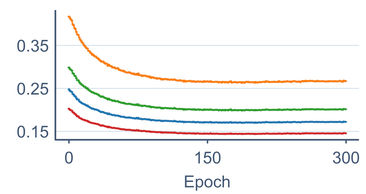
FIG. 3 Threshold proximity and gradient bias $\mathcal{E}_{\mathrm{bias}}$

Sequential CIFAR-10 · SCNN · Matched checkpoints and batches

Squared hard–smooth backward mismatch



Threshold-tail statistic D(k)



A > B > REF > C

Typical ordering in both measures.
A/B/C shift distances to threshold.
REF is the unmodified trajectory.

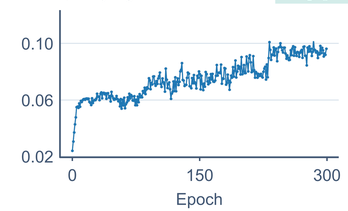
A · B · REF · C

Mean ± 1 SD over 3 seeds

Absolute hard–smooth loss gap

FIG. 4 Objective gap

Fixed δ, varying checkpoints $\mathcal{E}_{\mathrm{gap}}$

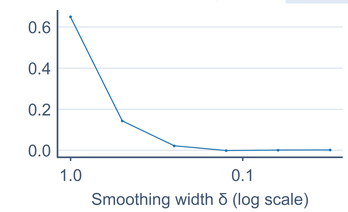


≈0.10 late in the plotted run

Same examples for both losses.

FIG. 5 Reference smoothing diagnostic

Fixed checkpoint, varying δ $\mathcal{E}_{\mathrm{ref}}$



≈0.65 at δ = 1, near 0 at small δ

Measured at a frozen training checkpoint.

Paper
NeurIPS paper page
neurips.cc

Code
PyTorch reference implementation
github.com/ZL-IN

Supplement
Proofs & experiments in paper appendices
OpenReview · PDF

NeurIPS 2026 · Poster 152782

Zelin Wei · weizl26@mails.jlu.edu.cn
Bin Gu · gubin@jlu.edu.cn