Understanding the Convergence of Direct Training of SNNs with Surrogate Gradients

Zelin Wei[1] **Tianxing Man**[2] **Wanli Shi**[1] **Bin Gu**[1*]
[1]School of Artificial Intelligence, Jilin University, China
[2]College of Computer Science, Inner Mongolia University, China
weizl26@mails.jlu.edu.cn wanli_shi@jlu.edu.cn
mantx@imu.edu.cn gubin@jlu.edu.cn

Abstract

Spiking Neural Networks (SNNs) offer high energy efficiency, but direct optimization is challenging because spike generation is governed by a non-differentiable hard threshold. Practical direct-training methods, including surrogate gradients (SG) and local zeroth-order (LocalZO) estimators, maintain hard spikes in the forward pass while using surrogate or perturbation-induced derivatives in the backward pass. This forward–backward mismatch raises a central theoretical question: how do local gradient discrepancies propagate through the spatiotemporal computation graph, and can the resulting biased updates still provide guarantees for the original discrete-spike objective? We develop a unified convergence analysis framework for SG and LocalZO direct SNN training. A smoothed spike objective is introduced only as an analytical bridge, allowing us to separate the core error into two components: the gradient-level mismatch between the direct-training mean direction and the smoothed-objective gradient, and the objective-level consistency gap between the smoothed and original discrete objectives. Both components are controlled by a threshold-tail functional that measures how often hard-trajectory membrane potentials lie near the firing threshold. Under suitable regularity assumptions, we obtain a non-asymptotic guarantee for the original discrete-spike objective. Experiments across representative SNN architectures and benchmarks verify convergence consistency, the existence of errors, and demonstrate that the threshold-tail statistic is an effective diagnostic for direct SNN training.

1 Introduction

Spiking Neural Networks (SNNs), which emulate the sparse, event-driven computation of biological brains, offer exceptional energy efficiency for resource-constrained edge AI and low-power embedded scenarios [1, 2]. Unlike conventional artificial neural networks that perform continuous dense matrix operations, SNNs consume energy only when neurons emit spikes, drastically reducing runtime power overhead. Yet their optimization remains fundamentally challenging due to the intrinsic non-differentiability of the spiking neuron's step activation function [3].

To address this challenge, direct SNN training mainly relies on surrogate gradients (SG) [4–7] and local zeroth-order (LocalZO) estimators [7, 8]. SG preserves the hard-threshold spike mechanism in the forward pass while replacing the non-differentiable step-function derivative with a smooth surrogate in backward propagation [9, 10]. LocalZO achieves a similar effect by using local perturbation-based estimation in the backward pass to approximate the gradient. While LocalZO avoids explicit surrogate functions, its gradient estimates are essentially expectations over local two-point estimators, which inherently smooth the discrete forward trajectory [8]. Together, these two paradigms enable

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

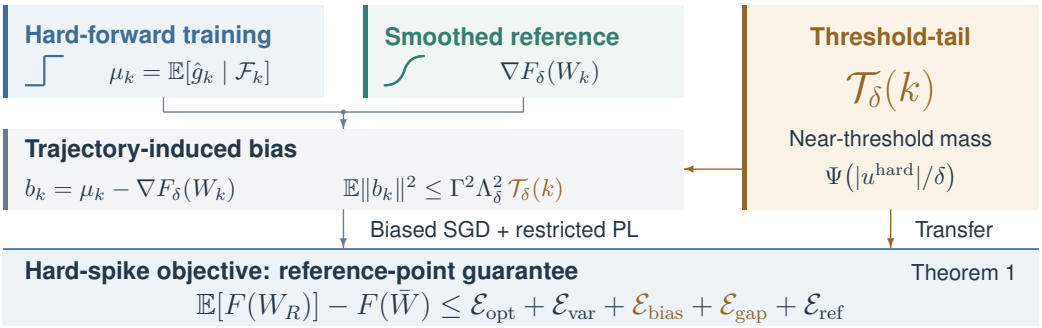

Figure 1: Convergence roadmap for hard-forward direct SNN training. The mean update and matched smoothed reference define the trajectory-induced bias. The threshold-tail functional bounds this bias and the discrete–smooth consistency gap. Biased-SGD descent and the restricted PL conversion then yield the reference-point guarantee for F, including the smoothing residual at $\bar{W}$.

gradient-based optimization for SNNs, forming the mainstream training framework and yielding competitive performance across various architectures and tasks [11, 12].

Despite their empirical success, the exact objective optimized by these methods remains theoretically unclear. In both SG and LocalZO, the forward pass follows the original discrete spike dynamics, while the backward pass propagates a surrogate signal [3, 8]. Taken together, this asymmetry induces a local gradient mismatch at each spiking node, which then propagates across layers and accumulates into a systematic bias in the update direction. Consequently, the resulting mean update direction generally coincides neither with the gradient of the original discrete-spike objective nor exactly with the gradient of a smoothed surrogate objective, leaving the objective effectively optimized by direct training implicit.

Building on these observations, existing theory has clarified several components of direct SNN training, including gradient estimators for discrete neurons, proxy-objective optimization, and neuron-level connections between SG and LocalZO estimators [10, 13, 8, 14]. However, these results do not yet provide a network-level route from local forward-backward mismatch to convergence guarantees for the original discrete-spike objective. The key difficulty is that the algorithm follows hard spike trajectories, whereas the natural analytical proxy is defined by smoothed spike dynamics.

To address this gap, this paper establishes a unified convergence framework for direct SNN training covering both SG and LocalZO methods. The smoothed objective is introduced only as an analytical intermediate, whereas the final guarantee is stated for the original discrete-spike objective used in inference. The analysis then quantifies two mismatch-related errors arising in this transfer: the trajectory-level systematic bias between the direct-training mean direction and the gradient of the smoothed objective, and the discrete–smooth consistency gap between the original discrete objective and its smoothed counterpart. Both errors are controlled by a common threshold-tail functional that characterizes how often membrane potentials lie near the firing threshold. Under the restricted PL condition, the stationarity guarantee for the smoothed objective is transferred to a reference-point guarantee for the original discrete-spike objective, yielding a non-asymptotic convergence bound with explicit tail-controlled residual terms. Figure 1 summarizes the convergence roadmap from direct-training mismatch to the final discrete-objective guarantee.

The main contributions are as follows:

- We develop a unified convergence framework for SG and LocalZO direct SNN training that models their shared hard-forward/backward-surrogate mismatch.
- We establish a reference-point non-asymptotic guarantee for the original discrete-spike objective, with explicit residuals from gradient mismatch and discrete–smooth inconsistency.
- We identify a common threshold-tail functional that controls these residuals and yields an interpretable diagnostic for direct SNN training.

2 Related Work

Theoretical perspectives on direct SNN training. Recent theory has begun to explain why direct SNN training can progress despite the non-differentiability of spike generation. Bengio et al. [10] analyzed gradient estimators for stochastic and discrete neurons, but not for deterministic deep SNNs with hard-threshold forward dynamics. Frei and Gu [13] showed that gradient methods on an objective satisfying proxy-convexity or proxy-PL conditions yield guarantees for a related proxy quantity. Here, the direct-training direction additionally differs from the gradient of the smoothed reference because it is evaluated along hard-spike trajectories. Gygax and Zenke [14] connected surrogate gradients to stochastic automatic differentiation and showed that they are generally not gradients of a single global surrogate loss; Eilers et al. [15] developed an NTK perspective on surrogate-gradient learning. These works clarify important components of surrogate-based optimization but do not provide a network-level convergence guarantee for the original discrete-spike objective under hard-spike trajectories.

Algorithmic perspectives on surrogate-based direct training. Most direct SNN training methods keep discrete spikes in the forward pass and replace the spike derivative in backpropagation with a smooth surrogate. Early works such as Lee et al. [7], Wu et al. [5], and Shrestha and Orchard [6] established the feasibility of direct spatiotemporal training. Subsequent methods improved trainability through differentiable formulations, architectural refinements, and adaptive surrogate designs [16–19]. Representative SG studies demonstrate strong empirical performance [4, 3, 20], but mainly treat the surrogate derivative as a backward rule rather than quantifying its objective-level bias. Zeroth-order perturbation and smoothing methods are well studied in optimization [21–25], mostly for smooth or explicitly smoothed black-box objectives rather than hard-threshold SNNs. Mukhoty et al. [8] connected SG and LocalZO by representing common surrogate derivatives as expectations of local two-point estimators, but only at the neuron level. Recent works also studied gradient-error accumulation and surrogate-width mismatch [26, 27]. Together, these works motivate a unified analysis of SG and LocalZO that helps us quantify forward-backward mismatch at the level of the original discrete-spike objective.

3 Preliminaries

3.1 Spiking Neural Networks

To characterize temporal spiking dynamics, we use the widely used leaky integrate-and-fire (LIF) neuron model as an example [28, 29]. Under a discrete-time hard-reset formulation, the state transitions at layer $l \in \{1, \dots, L\}$ and time step $t \in \{1, \dots, T\}$ are

$$\begin{aligned} X_l[t] &= g_l(S_{l-1}[t]; W_l), \qquad & H_l[t] &= \beta V_l[t-1] + X_l[t], \\ S_l[t] &= h(H_l[t] - V_{\mathrm{th}}), \qquad & V_l[t] &= H_l[t]\big(1 - S_l[t]\big), \end{aligned} \tag{1}$$

where W_l is the trainable parameter of layer l, and $W = (W_1, \dots, W_L)$ collects all trainable parameters. Here $X_l[t]$ is the input current induced by the synaptic map $g_l(\cdot; W_l)$, while $H_l[t]$ and $V_l[t]$ denote the membrane potentials before and after spike emission. The dynamics use decay factor $\beta \in (0, 1)$ and threshold $V_{\mathrm{th}} > 0$. Writing the signed distance from the membrane potential to the firing threshold as $u_l[t] \coloneqq H_l[t] - V_{\mathrm{th}}$, spike emission is given by the Heaviside function $h : \mathbb{R} \to \{0, 1\}$, $h(u) \coloneqq \mathbf{1}\{u > 0\}$; hence $S_l[t] = 1$ exactly when the membrane potential exceeds the threshold.

The discontinuity of h makes the exact gradient of the discrete-spike network unavailable to standard backpropagation, motivating the surrogate constructions reviewed next.

3.2 Surrogate Derivatives and Local Zeroth-Order Estimation

Surrogate derivatives. To circumvent the non-differentiability of the Heaviside step function $h(\cdot)$ in direct SNN training, a surrogate derivative is used in the backward pass. We call $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$ a surrogate derivative if it is even, non-decreasing on $(-\infty, 0)$, and integrable, namely $c_g \coloneqq \int_{-\infty}^{\infty} g(z)\, dz < \infty$. Simple examples of surrogate gradients include the Gaussian surrogate

$g_\delta^{\mathrm{gau}}(u) = \frac{1}{\delta\sqrt{2\pi}} \exp\left(-\frac{u^2}{2\delta^2}\right)$ [8], the Sigmoid surrogate $g_\delta^{\mathrm{sig}}(u) = \frac{1}{\delta} \cdot \frac{\exp(-u/\delta)}{\left(1+\exp(-u/\delta)\right)^2}$ [5], and the Atan surrogate $g_\delta^{\mathrm{ATan}}(u) = \frac{1}{2\delta\left(1+\left(\frac{\pi u}{2\delta}\right)^2\right)}$ [16].

Local Zeroth-Order estimator. Besides deterministic surrogate gradients, one may use a LocalZO estimator [8]. Let $\delta > 0$ be the perturbation width and draw z from an even probability density λ. The standard local two-point estimator is

$$G_{\mathrm{raw}}^2(u; z, \delta) := \frac{h(u+\delta z) - h(u-\delta z)}{2\delta}\, z = \begin{cases} 0, & |u| > |z|\delta, \\ \dfrac{|z|}{2\delta}, & |u| \le |z|\delta, \end{cases} \qquad z \sim \lambda, \tag{2}$$

If $\int_0^\infty t\lambda(t)\, dt < \infty$, then $\mathbb{E}_{z\sim\lambda}[G_{\mathrm{raw}}^2(u; z, \delta)] = \frac{1}{\delta}\int_{|u|/\delta}^\infty t\lambda(t)\, dt$.

Mukhoty et al. [8] establish a nodewise correspondence between LocalZO estimators and surrogate derivatives. The original correspondence is recalled in Appendix A.1. We use its normalized form, $G^2 = c^{-1}G_{\mathrm{raw}}^2$ with $c = \mathbb{E}[z^2]$, so that $\mathbb{E}_z[G^2] = g_\delta = h'_\delta$; Lemma 1 states the precise identity. The smoothed spike h_δ and its surrogate derivative g_δ have distinct roles in the analysis.

4 Problem Setup

4.1 Notations and Direct-Training Update Rule

We consider a finite training set $D = \{\xi_i = (x_i, y_i)\}_{i=1}^n$ and write $\xi \sim \mathrm{Unif}(D)$. Accordingly, $\mathbb{E}_\xi[\cdot]$ denotes the empirical expectation. Let $W \in \mathbb{R}^d$ collect all trainable parameters, and let M be the number of spiking nodes in the finite spatiotemporal computation graph.

Forward propagation. For a parameter vector W and sample ξ, let $s^{\mathrm{discrete}}(W; \xi) \in \{0, 1\}^M$ and $u^{\mathrm{hard}}(W; \xi) \in \mathbb{R}^M$ denote the hard-spike trajectory and the signed distances from the membrane potentials to the firing threshold. They satisfy, componentwise, $s_i^{\mathrm{discrete}}(W; \xi) = h(u_i^{\mathrm{hard}}(W; \xi))$ with $h(u) = \mathbf{1}\{u > 0\}$. The analytical smoothed trajectory $s^{\mathrm{smooth}}(W; \xi) \in [0, 1]^M$ is obtained by replacing every hard threshold with $h_\delta(u) = \mathbb{E}_{\tilde{z}\sim\tilde{\lambda}}[h(u + \delta\tilde{z})]$; we denote $g_\delta(u) = h'_\delta(u)$ whenever the derivative exists.

Backward propagation. For a fixed trajectory argument $s \in [0, 1]^M$, let $G_\delta(s, W; \xi) \in \mathbb{R}^d$ be the parameter-space backward signal using s as the realized forward trajectory and applying the chain rule along this trajectory with the local spike derivative g_δ. Thus, s determines the active spatiotemporal path, whereas g_δ determines the local backward signal.

Symbol	Role in direct training
$h_\delta,\ g_\delta$	Smoothed spike and its derivative, $g_\delta = h'_\delta$
$G_{\mathrm{raw}}^2,\ G^2$	Raw and normalized nodewise LocalZO estimators
$G_\delta(s, W; \xi)$	Parameter-space backward map on a fixed trajectory s
$\hat{g}(W; \xi, z)$	Stochastic parameter-space update for one sample
$g(W; \xi),\ \mu_k$	Perturbation mean and mini-batch conditional mean

Direct-training update rule. Let $\mathcal{F}_k$ be the σ-field generated by all randomness prior to sampling at iteration k. Conditioned on $\mathcal{F}_k$, the algorithm draws a mini-batch $\{\xi_{k,j}\}_{j=1}^B$. For each sample, SG computes the direction from the hard forward trajectory and the surrogate backward derivative; LocalZO additionally draws m independent collections of local perturbations $z_{k,j,1:m}$ and averages the resulting replicated estimators. With $m = 1$ and a dummy perturbation that recovers deterministic SG, the update is

$$\hat{g}_k = \frac{1}{B}\sum_{j=1}^B \hat{g}_{k,j}, \qquad \hat{g}_{k,j} = \frac{1}{m}\sum_{r=1}^m \hat{g}(W_k; \xi_{k,j}, z_{k,j,r}), \qquad W_{k+1} = W_k - \eta\hat{g}_k. \tag{3}$$

To isolate the systematic update direction from stochastic fluctuations, define $g(W_k; \xi) := \mathbb{E}_z[\hat{g}(W_k; \xi, z) \mid \mathcal{F}_k, \xi]$ and $\mu_k := \mathbb{E}[\hat{g}_k \mid \mathcal{F}_k] = \mathbb{E}_\xi[g(W_k; \xi) \mid \mathcal{F}_k]$. Here $g(W_k; \xi)$ averages

only LocalZO perturbations for a fixed sample, whereas μ_k averages further over data sampling. Throughout, $\|\cdot\|$ denotes the Euclidean norm.

4.2 Discrete Objective and Smooth Objective

The primary objective is the empirical risk induced by the discrete-spike objective. For a sample $\xi = (x, y)$, let $\ell(s, W; \xi)$ denote the task loss evaluated on a spike trajectory s, with the same trainable parameters W used in the readout and recurrent dynamics when applicable. The discrete-spike objective is

$$F(W) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \ell\big(s^{\mathrm{discrete}}(W;\xi_i), W; \xi_i\big) = \mathbb{E}_{\xi}\big[\ell\big(s^{\mathrm{discrete}}(W;\xi), W; \xi\big)\big]. \tag{4}$$

The objective F can be discontinuous across parameter surfaces where the spike pattern changes, and can be locally flat along directions that preserve the pattern when the loss has no explicit dependence on those parameters. Smoothness is assumed only for F_δ below; standard smooth nonconvex analysis cannot be applied directly to F.

To construct a differentiable analytical reference, we replace each hard threshold h in the forward dynamics by the smoothed spike h_δ, while keeping the data, parameters, and loss unchanged. This defines the smoothed trajectory $s^{\mathrm{smooth}}(W;\xi)$ and the smoothed objective

$$F_\delta(W) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \ell\big(s^{\mathrm{smooth}}(W;\xi_i), W; \xi_i\big) = \mathbb{E}_{\xi}\big[\ell\big(s^{\mathrm{smooth}}(W;\xi), W; \xi\big)\big]. \tag{5}$$

The role of F_δ is purely analytical: the update in (3) follows hard-spike forward trajectories, while F_δ provides the reference used to quantify the forward–backward mismatch. Matched smooth forward and backward passes would instead give mean direction $\nabla F_\delta(W)$; the hard-forward methods studied here retain a trajectory-induced bias. The final guarantee in Section 5.1 is transferred back to the original objective F.

4.3 Assumptions

Assumptions 1–2 support biased-SGD stationarity; Assumption 3 propagates trajectory discrepancies, Assumption 4 transfers them to objective values, and Assumption 5 is used for the final function-value conversion.

Assumption 1 (Smoothness of the smoothed objective [23, 30–32])**.** *The smoothed objective F_δ is differentiable, bounded below by $F_\delta^\star \coloneqq \inf_W F_\delta(W) > -\infty$, and has an L_δ-Lipschitz continuous gradient, i.e., $\|\nabla F_\delta(W) - \nabla F_\delta(W')\| \le L_\delta \|W - W'\|$ for all $W, W' \in \mathbb{R}^d$. Moreover, the sample gradient is integrable and differentiation commutes with empirical expectation: $\nabla F_\delta(W) = \mathbb{E}_\xi[\nabla_W \ell(s^{\mathrm{smooth}}(W;\xi), W; \xi)]$.*

Assumption 2 (Sampling and variance bounds [23, 31, 32])**.** *For every $k \ge 0$, conditioned on $\mathcal{F}_k$, the mini-batch samples $\{\xi_{k,j}\}_{j=1}^{B}$ are i.i.d.; conditioned on $(\mathcal{F}_k, \xi_{k,j})$, the perturbations $\{z_{k,j,r}\}_{r=1}^{m}$ are independent across both samples and replicates. In addition, there exist constants $\sigma^2_{\mathrm{mb},1}, \sigma^2_{\mathrm{sg},1} \ge 0$ such that $\mathbb{E}_\xi[\|g(W_k;\xi) - \mu_k\|^2 \mid \mathcal{F}_k] \le \sigma^2_{\mathrm{mb},1}$ and $\mathbb{E}_z[\|\hat{g}(W_k;\xi,z) - g(W_k;\xi)\|^2 \mid \mathcal{F}_k, \xi] \le \sigma^2_{\mathrm{sg},1}$ for all k and all samples ξ.*

Remark: We use the effective variances $\sigma^2_{\mathrm{mb}} \coloneqq \sigma^2_{\mathrm{mb},1}/B$ and $\sigma^2_{\mathrm{sg}} \coloneqq \sigma^2_{\mathrm{sg},1}/B$. For deterministic SG, $\sigma^2_{\mathrm{sg}} = 0$.

Assumption 3 (Local regularity of the SNN computation graph)**.** *Let $u(s, W; \xi) \in \mathbb{R}^M$ be the stacked vector of signed distances from membrane potentials to the firing threshold induced by trajectory argument s. Uniformly over the region visited by the iterates and over all samples ξ, the following hold: (i) the backward map is Lipschitz in the trajectory, $\|G_\delta(s, W; \xi) - G_\delta(s', W; \xi)\| \le \Gamma \|s - s'\|$ for some $\Gamma > 0$; (ii) the signed-threshold-distance map is Lipschitz in the trajectory, $\|u(s, W; \xi) - u(s', W; \xi)\| \le L_u \|s - s'\|$ for some $L_u > 0$; (iii) the smoothed spike function is Lipschitz, $|h_\delta(a) - h_\delta(b)| \le L_h|a - b|$ for some $L_h > 0$ and all $a, b \in \mathbb{R}$; (iv) in the finite-horizon setting, writing s_t and u_t for the temporal blocks at step t, there exists $L_{\mathrm{c}} > 0$ such that $\big\|u_t^{\mathrm{hard}} - u_t^{\mathrm{smooth}}\big\| \le L_{\mathrm{c}} \big\|s_{t-1}^{\mathrm{discrete}} - s_{t-1}^{\mathrm{smooth}}\big\|$ for all t, when the discrete and smoothed systems share the same initialization.*

For a finite unrolled graph on a bounded visited region, Lipschitz synaptic/state maps and bounded, Lipschitz derivatives of the maps entering the backward computation give finite L_u and Γ by composition. One may take $L_h = \sup_u |g_\delta(u)|$ for the smoothings above. Here L_u describes the stacked trajectory map, while L_c is the one-step sensitivity in (iv). The finite-horizon branch additionally assumes that causal estimate and does not require $L_h L_u < 1$; its amplification factor can nevertheless grow with the horizon.

Assumption 4 (Local loss Lipschitzness [33])**.** *There exists $L_s > 0$ such that, for every iterate W_k, every sample ξ, and all trajectories $s, s' \in [0,1]^M$, $|\ell(s, W_k; \xi) - \ell(s', W_k; \xi)| \le L_s \|s - s'\|$.*

Assumption 5 (Restricted Polyak–Łojasiewicz condition [34])**.** *There exist a deterministic set $\mathcal{U} \subseteq \mathbb{R}^d$, a constant $\mu_{\mathcal{U}} > 0$, and a reference minimizer $\bar{W} \in \arg\min_{W \in \mathcal{U}} F_\delta(W)$ such that $W_k \in \mathcal{U}$ almost surely for all $0 \le k \le K$, and $F_\delta(W) - F_\delta(\bar{W}) \le (2\mu_{\mathcal{U}})^{-1} \|\nabla F_\delta(W)\|^2$ for all $W \in \mathcal{U}$.*

The restricted PL inequality describes the geometry of F_δ on $\mathcal{U}$; containment of the biased iterates in $\mathcal{U}$ is a separate condition. The stationarity and mismatch results below do not require this PL assumption.

5 Main Theorem

5.1 Non-asymptotic Convergence for the Discrete Objective

We state a reference-point non-asymptotic guarantee for the discrete-spike objective F. Write $\Psi(a) \coloneqq \mathbb{P}(|\tilde{z}| \ge a)$ for the smoothing tail and define

$$\mathcal{T}_\delta(W) \coloneqq \mathbb{E}_\xi \left[\sum_{i=1}^{M} \Psi\left(\frac{|u_i^{\mathrm{hard}}(W;\xi)|}{\delta} \right) \right], \qquad \mathcal{T}_\delta(k) \coloneqq \mathbb{E}[\mathcal{T}_\delta(W_k)]. \tag{6}$$

This aggregates near-threshold mass along hard trajectories. The trajectory amplification is $\Lambda_\delta = (1 - L_h L_u)^{-1}$ in the contraction branch and $\Lambda_\delta = \sum_{t=0}^{T-1} (L_h L_c)^t$ in the finite-horizon branch. Set $\mathcal{B}_\delta^2(k) \coloneqq \Gamma^2 \Lambda_\delta^2 \mathcal{T}_\delta(k)$, $\bar{\mathcal{T}}_K \coloneqq K^{-1} \sum_{k=0}^{K-1} \mathcal{T}_\delta(k)$, $\bar{\mathcal{B}}_{\delta,K}^2 \coloneqq K^{-1} \sum_{k=0}^{K-1} \mathcal{B}_\delta^2(k)$, and $\Delta_\delta \coloneqq F_\delta(W_0) - F_\delta(\bar{W})$. The reference $\bar{W} \in \arg\min_{W \in \mathcal{U}} F_\delta(W)$ is a minimizer of the smoothed objective on $\mathcal{U}$.

Theorem 1 (Non-asymptotic convergence for the discrete objective)**.** *Draw $R \sim \mathrm{Unif}\{0, \dots, K-1\}$ independently of the optimization randomness and return W_R. Assume $g(W_k; \xi) = G_\delta(s^{\mathrm{discrete}}(W_k;\xi), W_k; \xi)$, as established for SG and normalized nodewise LocalZO in Proposition 1. Suppose Assumptions 1, 2, 4, and 5 hold, and suppose Assumption 3(i)–(iii) holds together with either $L_h L_u < 1$ or Assumption 3(iv). If $\eta \le 1/(4L_\delta)$, then*

$$\begin{aligned} \mathbb{E}[F(W_R)] - F(\bar{W}) \le\ & \underbrace{\frac{2\Delta_\delta}{\mu_{\mathcal{U}} \eta K}}_{\mathcal{E}_{\mathrm{opt}}} + \underbrace{\frac{L_\delta \eta}{\mu_{\mathcal{U}}} \left(\sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m} \right)}_{\mathcal{E}_{\mathrm{var}}} + \underbrace{\frac{3}{2\mu_{\mathcal{U}}} \bar{\mathcal{B}}_{\delta,K}^2}_{\mathcal{E}_{\mathrm{bias}}} \\ & + \underbrace{L_s \Lambda_\delta \sqrt{\bar{\mathcal{T}}_K}}_{\mathcal{E}_{\mathrm{gap}}} + \underbrace{\left| F(\bar{W}) - F_\delta(\bar{W}) \right|}_{\mathcal{E}_{\mathrm{ref}}}. \end{aligned} \tag{7}$$

The expectation includes the random index R, mini-batch sampling, and LocalZO perturbations when present. The labeled terms separate the main effects in the bound. $\mathcal{E}_{\mathrm{opt}}$ is the finite-time descent error on the smoothed objective after the restricted PL conversion, and $\mathcal{E}_{\mathrm{var}}$ combines mini-batch variance with LocalZO perturbation variance, whose contribution decreases as m increases. $\mathcal{E}_{\mathrm{bias}}$ captures the accumulated forward-backward mismatch from evaluating the backward map along hard rather than smoothed trajectories; it is controlled by $\mathcal{T}_\delta(k)$ through $\mathcal{B}_\delta^2(k) = \Gamma^2 \Lambda_\delta^2 \mathcal{T}_\delta(k)$. $\mathcal{E}_{\mathrm{gap}}$ is the objective-level consistency gap between F and F_δ, and $\mathcal{E}_{\mathrm{ref}}$ is the reference-point smoothing residual. Thus, the same threshold-tail functional controls both the gradient-level bias and the objective-level transfer.

With $\eta = \Theta(K^{-1/2})$, $\mathcal{E}_{\mathrm{opt}} + \mathcal{E}_{\mathrm{var}} = O(K^{-1/2})$. The remaining terms quantify the residual cost of transferring the smoothed-objective guarantee to the discrete-spike objective and are small when

hard-trajectory membrane potentials stay away from the firing threshold, and the reference smoothing residual is small. The smoothing width δ trades consistency against regularity: at fixed nonzero threshold margins, smaller δ improves spike approximation but may increase L_δ and L_h, reducing the admissible stepsize and enlarging Λ_δ. Appendix E.1 gives an exact one-spike example and discusses the reference residual.

5.2 Proof Sketch

The proof follows the decomposition in Theorem 1; the next section formalizes the corresponding steps. First, F_δ is introduced only as an analytical bridge. Proposition 1 shows that SG and normalized LocalZO share the same conditional mean direction through the trajectory-dependent backward map, yielding the bias decomposition between the hard-trajectory update and the smoothed-objective gradient.

Second, the update is analyzed as biased stochastic gradient descent on F_δ. Smoothness and the variance decomposition give an average stationarity bound with finite-time descent, stochastic variance, and accumulated bias. Proposition 2 then controls this bias by the threshold-tail functional.

Finally, the guarantee is transferred from F_δ to the discrete-spike objective F. Proposition 3 bounds the discrete–smooth consistency gap using the same trajectory discrepancy, while the restricted PL condition converts the stationarity of F_δ into a reference-point function-value bound. Adding the residual $|F(\bar{W}) - F_\delta(\bar{W})|$ proves Theorem 1.

6 Framework for Analyzing Convergence of the Discrete Objective

6.1 Gradient Mismatch in Direct Training

In the convergence analysis, we adopt the normalized version of the nodewise LocalZO estimator introduced in Section 3.2, whose expectation exactly matches the surrogate derivative used by SG.

Lemma 1. *Let $z \sim \lambda$ be drawn from an even density with $0 < c := \mathbb{E}[z^2] < \infty$, and define $G^2(u; z, \delta) := c^{-1} G^2_{\mathrm{raw}}(u; z, \delta)$. With $\tilde{\lambda}(s) := c^{-1} \int_{|s|}^{\infty} t\lambda(t)\, dt$, $h_\delta(u) := \mathbb{E}_{\tilde{z}\sim\tilde{\lambda}}[h(u + \delta\tilde{z})]$, and $g_\delta(u) := h'_\delta(u)$, we have, for almost every $u \in \mathbb{R}$, $\mathbb{E}_{z\sim\lambda}[G^2(u; z, \delta)] = h'_\delta(u) = g_\delta(u)$.*

The proof is given in Appendix A.2. Thus, SG inserts g_δ deterministically, while normalized LocalZO inserts an unbiased nodewise estimator of the same derivative. Their conditional means agree under independent local perturbations; their realized updates differ, with LocalZO contributing the additional perturbation variance in Assumption 2.

We next express the conditional mean direction using the trajectory-dependent backward map G_δ. For SG, the identity holds by construction. For LocalZO, it follows from Lemma 1 and the standard nodewise construction with independent local perturbations on the fixed hard-trajectory backward graph; see Appendix A.3.

Proposition 1. *For every iteration k and almost every sample ξ, under deterministic SG or the normalized nodewise LocalZO construction above,*

$$g(W_k; \xi) = G_\delta\big(s^{\mathrm{discrete}}(W_k; \xi), W_k; \xi\big). \tag{8}$$

Proposition 1 implies $\mu_k = \mathbb{E}_\xi[G_\delta(s^{\mathrm{discrete}}(W_k; \xi), W_k; \xi) \mid \mathcal{F}_k]$, whereas $\nabla F_\delta(W_k) = \mathbb{E}_\xi[G_\delta(s^{\mathrm{smooth}}(W_k; \xi), W_k; \xi) \mid \mathcal{F}_k]$. Define the systematic bias by $b_k := \mu_k - \nabla F_\delta(W_k)$. The mismatch is therefore trajectory-level: direct training backpropagates along the hard-spike path, while ∇F_δ is defined along the smoothed path.

6.2 Biased Stationarity for the Smoothed Objective

With the mismatch isolated in b_k, the update in (3) can be analyzed as biased stochastic gradient descent on F_δ.

Theorem 2. *Under Assumptions 1 and 2, if $\eta \le 1/(4L_\delta)$, then for every $K \ge 1$,*

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\Big[\|\nabla F_\delta(W_k)\|^2\Big] \le \frac{4\big(F_\delta(W_0) - F_\delta^\star\big)}{\eta K} + 2L_\delta\eta\left(\sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m}\right) + \frac{3}{K}\sum_{k=0}^{K-1} \mathbb{E}\Big[\|b_k\|^2\Big]. \tag{9}$$

The first two terms in (9) are the standard finite-time optimization and stochastic-variance terms for optimizing F_δ. The last term is specific to direct SNN training and measures the accumulated forward-backward mismatch caused by evaluating the backward map along hard spike trajectories.

6.3 Tail-Probability Control of the Systematic Bias

Theorem 2 leaves one SNN-specific term to control: the accumulated bias $\mathbb{E}\|b_k\|^2$. By the definition of b_k, this bias comes from evaluating the backward map on the hard trajectory, whereas ∇F_δ is defined on the smoothed trajectory. We control this discrepancy by the probability that local smoothing crosses a firing threshold.

Recall the tail functional $\mathcal{T}_\delta(k)$ and amplification factor Λ_δ defined before Theorem 1. The former measures aggregate near-threshold mass along hard trajectories; the latter accounts for propagation through the spatiotemporal graph.

Proposition 2. *Suppose Proposition 1, Assumption 1, and Assumption 3(i)–(iii) hold. Suppose further that either $L_h L_u < 1$ or Assumption 3(iv) holds. Then, for every iteration k,*

$$\mathbb{E}\Big[\|b_k\|^2\Big] \le \Gamma^2 \Lambda_\delta^2 \, \mathcal{T}_\delta(k) = \mathcal{B}_\delta^2(k). \tag{10}$$

Substituting Proposition 2 into Theorem 2 replaces the bias term by $3K^{-1}\sum_{k=0}^{K-1} \mathcal{B}_\delta^2(k)$. Thus, beyond the standard optimization and stochastic-variance terms, the smoothed-objective stationarity error has a single explicit mismatch term controlled by near-threshold mass and amplified by Γ and Λ_δ.

6.4 Consistency Gap Between the Smoothed and Discrete Objectives

We now transfer the smoothed-objective guarantee to the original discrete objective F. The required transfer is an objective-level consistency bound along the returned iterate.

Proposition 3. *Let W_R be selected by drawing $R \sim \mathrm{Unif}\{0, \dots, K-1\}$ independently of the optimization randomness. Under Assumption 4, Assumption 3(ii)–(iii), and either $L_h L_u < 1$ or Assumption 3(iv),*

$$\mathbb{E}\big|F(W_R) - F_\delta(W_R)\big| \le L_s \Lambda_\delta \sqrt{\frac{1}{K}\sum_{k=0}^{K-1} \mathcal{T}_\delta(k)}. \tag{11}$$

The same threshold-tail functional $\mathcal{T}_\delta(k)$ controls both the gradient-level bias in (10) and the objective-level consistency gap in (11). To assemble the final guarantee, split the difference $\mathbb{E}[F(W_R)] - F(\bar{W})$ into the consistency gap at W_R, the smoothed-objective difference, and the reference smoothing residual. By Assumption 5, $\mathbb{E}[F_\delta(W_R) - F_\delta(\bar{W})] \le (2\mu_{\mathcal{U}})^{-1}\mathbb{E}[\|\nabla F_\delta(W_R)\|^2]$. Combining this inequality with Theorem 2, Proposition 2, the consistency bound (11), and this decomposition proves Theorem 1. Full derivations are provided in Appendices B–E.

7 Experiments

7.1 Verification of Theoretical Results

Convergence Behavior of Direct Training We first test whether SG and LocalZO exhibit consistent direct-training behavior under the unified mean-update framework. For LocalZO, we vary $m \in \{5, 10, 20, 40\}$ and compare its update direction and training loss with SG.

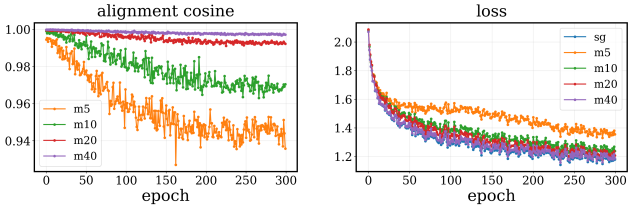

Figure 2: SG-LocalZO alignment and training loss on CIFAR-10 with an SCNN.

Figure 2 shows that LocalZO remains aligned with SG throughout training, with alignment improving as m increases. All LocalZO runs descend stably; $m = 20$ and $m = 40$

closely track SG, while $m = 5$ has larger fluctuations and a higher final loss. These results support the unified treatment: SG and LocalZO share the same mean direct-training mechanism, while LocalZO adds a perturbation-variance term that decreases with m.

Objective Consistency Gap Analysis We next examine the objective-level consistency term in Proposition 3. At matched checkpoints, we evaluate the empirical hard and smoothed losses $\widehat{F}(W_k)$ and $\widehat{F}_\delta(W_k)$ on the same examples and compute their gap $G_k := |\widehat{F}(W_k) - \widehat{F}_\delta(W_k)|$, and the empirical threshold-tail statistic $\widehat{\mathcal{T}}_\delta(k)$ from hard-trajectory signed threshold distances. We compare G_k with $\widehat{C}B_k$, where B_k is the square root of the prefix-averaged empirical tail statistic. The constant $\widehat{C}$ is chosen once from the plotted run as a post-hoc visualization scale; Appendix F.3 gives the definition.

Figure 3 shows that the discrete and smoothed empirical objectives follow similar decreasing trends, supporting F_δ as an analytical bridge. Their nonzero gap confirms that the two objectives are distinct, while the calibrated threshold-tail proxy tracks the gap scale under a fixed calibration. This comparison illustrates the temporal relation between threshold proximity and discrete–smooth mismatch.

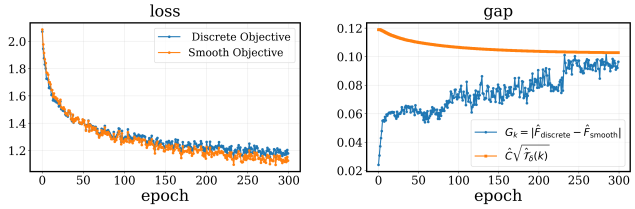

Figure 3: Discrete–smooth objective consistency on CIFAR-10 with an SCNN.

Gradient-level bias diagnostic. The remaining SNN-specific term in Theorem 1 is the gradient-level bias $\mathcal{B}^2_{\delta,K}$, defined by the discrepancy between the direct-training mean update and ∇F_δ. We examine it in Appendix F.5. Empirically, the hard–smooth backward mismatch is consistently nonzero and, under controlled interventions, follows the ordering induced by the threshold-tail statistic $D(k)$. This supports the threshold-tail functional as a diagnostic proxy for the SNN-specific error terms in the convergence bound.

Reference Point Error Analysis The final residual term in Theorem 1, $\left|F(\bar{W}) - F_\delta(\bar{W})\right|$, is a reference-point smoothing discrepancy. To isolate it from optimization dynamics, we freeze a converged checkpoint W_{fix} and evaluate $\widehat{R}_{\text{fix}}(\delta) := |\widehat{F}(W_{\text{fix}}) - \widehat{F}_\delta(W_{\text{fix}})|$ while varying only δ.

Figure 4 shows that $\widehat{R}_{\text{fix}}(\delta)$ rapidly decreases as δ becomes smaller and is nearly negligible in the small-δ regime. Since the parameters are fixed, this trend isolates smoothing approximation at W_{fix}. This probes smoothing approximation at a trained checkpoint, while the theorem evaluates the reference residual at the smoothed minimizer $\bar{W}$.

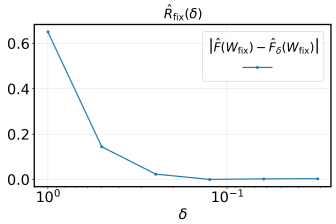

Figure 4: Smoothing residual at a frozen CIFAR-10 SCNN checkpoint.

7.2 Cross-Model Diagnostic Application

We further evaluate $D(k) = \widehat{\mathcal{T}}_\delta(k)/M$, the empirical per-node threshold-tail diagnostic, across datasets, architectures, and seeds. Appendix F.1 gives its mini-batch estimator. For each trained checkpoint, we report test accuracy and $D(k)$, where lower values indicate fewer near-threshold membrane states and a smaller tail contribution to the mismatch bounds.

Table 1 shows that, within each completed multi-seed setting, higher-accuracy checkpoints generally coincide with smaller $D(k)$. The diagnostic is computed directly from hard-forward threshold distances and complements validation loss by monitoring mismatch within a fixed setting. Its scale depends on architecture, time horizon, neuron dynamics, and smoothing width; the seed-wise association is descriptive, with matched comparisons providing the relevant interpretation.

Table 1: Cross-model diagnostic evaluation. For each dataset–model pair, we report seed-wise test accuracy and diagnostic value $D(k)$.

Dataset	Model	Epoch	T	Seed 1		Seed 2		Seed 3	
				Acc.	$D(k)$	Acc.	$D(k)$	Acc.	$D(k)$
DVSGesture	7B-Net	190	16	95.14	0.2650	94.79	0.2685	94.44	0.2701
CIFAR10-DVS	SpikingVGG	100	10	86.10	0.2297	85.80	0.2305	86.90	0.2294
Sequential CIFAR-10	SCNN	300	32	82.36	0.1811	82.63	0.1763	82.34	0.1815
Sequential CIFAR-100	SCNN	300	32	57.23	0.1728	57.06	0.1737	56.97	0.1777
SHD	SHDSNN	100	100	83.22	0.1233	84.18	0.1146	83.39	0.1179
TinyImageNet	SEW ResNet-34	120	4	61.80	0.2137	61.78	0.2195	61.19	0.2216

8 Conclusion

In this work, we develop a unified convergence framework for direct SNN training with surrogate gradients and LocalZO estimators. The analysis decomposes the hard-forward/surrogate-backward mismatch into a gradient bias and a discrete–smooth objective gap, both controlled by a threshold-tail functional measuring near-threshold membrane activity. This yields a reference-point non-asymptotic guarantee for the original discrete-spike objective and motivates threshold proximity as a diagnostic for direct-training mismatch. The results rely on smoothness, bounded variance, graph regularity, and restricted PL-type conditions, with potentially conservative constants for long-horizon, deep recurrent, or heterogeneous SNNs. Future work may relax these assumptions, design tail-aware surrogates or regularizers, and extend the framework to online and event-driven settings.

Acknowledgments and Disclosure of Funding

Funding. The authors received no third-party funding or support for this work.

Competing Interests. The authors declare no competing financial interests relevant to this work.

References

[1] Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997.

[2] Mike Davies, Narayan Srinivasa, Tsung-Han Lin, Gautham Chinya, Yongqiang Cao, Sri Harsha Choday, Georgios Dimou, Prasad Joshi, Nabil Imam, Shweta Jain, et al. Loihi: A neuromorphic manycore processor with on-chip learning. *Ieee micro*, 38(1):82–99, 2018.

[3] Friedemann Zenke and Surya Ganguli. Superspike: Supervised learning in multilayer spiking neural networks. *Neural computation*, 30(6):1514–1541, 2018.

[4] Emre O Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

[5] Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. Spatio-temporal backpropagation for training high-performance spiking neural networks. *Frontiers in neuroscience*, 12:331, 2018.

[6] Sumit B Shrestha and Garrick Orchard. Slayer: Spike layer error reassignment in time. *Advances in neural information processing systems*, 31, 2018.

[7] Jun Haeng Lee, Tobi Delbruck, and Michael Pfeiffer. Training deep spiking neural networks using backpropagation. *Frontiers in neuroscience*, 10:508, 2016.

[8] Bhaskar Mukhoty, Velibor Bojkovic, William De Vazelhes, Xiaohan Zhao, Giulia De Masi, Huan Xiong, and Bin Gu. Direct training of snn using local zeroth order method. *Advances in Neural Information Processing Systems*, 36:18994–19014, 2023.

[9] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[10] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[11] Guillaume Bellec, Franz Scherr, Anand Subramoney, Elias Hajek, Darjan Salaj, Robert Legenstein, and Wolfgang Maass. A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature communications*, 11(1):3625, 2020.

[12] Bojian Yin, Federico Corradi, and Sander M Bohté. Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks. *Nature Machine Intelligence*, 3(10):905–913, 2021.

[13] Spencer Frei and Quanquan Gu. Proxy convexity: A unified framework for the analysis of neural networks trained by gradient descent. *Advances in Neural Information Processing Systems*, 34:7937–7949, 2021.

[14] Julia Gygax and Friedemann Zenke. Elucidating the theoretical underpinnings of surrogate gradient learning in spiking neural networks. *Neural Computation*, 37(5):886–925, 2025.

[15] Luke Eilers, Raoul-Martin Memmesheimer, and Sven Goedeke. A generalized neural tangent kernel for surrogate gradient learning. *Advances in Neural Information Processing Systems*, 37:9026–9085, 2024.

[16] Dongsung Huh and Terrence J Sejnowski. Gradient descent for spiking neural networks. *Advances in neural information processing systems*, 31, 2018.

[17] Chankyu Lee, Syed Shakib Sarwar, Priyadarshini Panda, Gopalakrishnan Srinivasan, and Kaushik Roy. Enabling spike-based backpropagation for training deep neural network architectures. *Frontiers in neuroscience*, 14:497482, 2020.

[18] Wei Fang, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, and Yonghong Tian. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2661–2671, 2021.

[19] Yuhang Li, Yufei Guo, Shanghang Zhang, Shikuang Deng, Yongqing Hai, and Shi Gu. Differentiable spike: Rethinking gradient-descent for training spiking neural networks. *Advances in neural information processing systems*, 34:23426–23439, 2021.

[20] Nicolas Perez-Nieves, Vincent CH Leung, Pier Luigi Dragotti, and Dan FM Goodman. Neural heterogeneity promotes robust learning. *Nature communications*, 12(1):5791, 2021.

[21] James C Spall. A one-measurement form of simultaneous perturbation stochastic approximation. *Automatica*, 33(1):109–112, 1997.

[22] Abraham D Flaxman, Adam Tauman Kalai, and H Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007*, 2004.

[23] Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

[24] Sijia Liu, Pin-Yu Chen, Bhavya Kailkhura, Gaoyuan Zhang, Alfred O Hero III, and Pramod K Varshney. A primer on zeroth-order optimization in signal processing and machine learning: Principals, recent advances, and applications. *IEEE Signal Processing Magazine*, 37(5):43–54, 2020.

[25] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high dimensionality, and saddle points. *Foundations of Computational Mathematics*, 22(1):35–76, 2022.

[26] Shikuang Deng, Hao Lin, Yuhang Li, and Shi Gu. Surrogate module learning: Reduce the gradient error accumulation in training spiking neural networks. In *International Conference on Machine Learning*, pages 7645–7657. PMLR, 2023.

[27] Shuang Lian, Jiangrong Shen, Qianhui Liu, Ziming Wang, Rui Yan, and Huajin Tang. Learnable surrogate gradient for direct training spiking neural networks. In *Ijcai*, pages 3002–3010, 2023.

[28] Yang Li, Yinqian Sun, Xiang He, Yiting Dong, Dongcheng Zhao, and Yi Zeng. Parallel spiking unit for efficient training of spiking neural networks. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2024.

[29] Wulfram Gerstner and Werner M Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

[30] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

[31] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[32] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[33] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

[34] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.

[35] Alex Krizhevsky, Geoff Hinton, et al. Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40(7):1–9, 2010.

[36] Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11:244131, 2017.

[37] Arnon Amir, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo Di Nolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, et al. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7243–7252, 2017.

[38] Benjamin Cramer, Yannik Stradmann, Johannes Schemmel, and Friedemann Zenke. The heidelberg spiking data sets for the systematic evaluation of spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2744–2757, 2020.

[39] Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[40] Wei Fang, Zhaofei Yu, Zhaokun Zhou, Ding Chen, Yanqi Chen, Zhengyu Ma, Timothée Masquelier, and Yonghong Tian. Parallel spiking neurons with high efficiency and ability to learn long-term dependencies. *Advances in Neural Information Processing Systems*, 36:53674–53687, 2023.

[41] Wei Fang, Zhaofei Yu, Yanqi Chen, Tiejun Huang, Timothée Masquelier, and Yonghong Tian. Deep residual learning in spiking neural networks. *Advances in neural information processing systems*, 34: 21056–21069, 2021.

[42] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[43] Yifan Huang, Wei Fang, Zecheng Hao, Zhengyu Ma, and Yonghong Tian. Towards lossless memory-efficient training of spiking neural networks via gradient checkpointing and spike compression. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=nrBJ0Uvj7c`.

A Technical Preliminaries for Direct Training

This appendix collects the technical preliminaries supporting the direct-training estimator. We first record the normalized LocalZO-SG smoothing identity, then verify the mean-alignment property used in the main text, and finally derive a distribution-dependent variance bound for the LocalZO estimator.

A.1 Nodewise LocalZO–Surrogate Correspondence

The following two lemmas from Mukhoty et al. [8] characterize the correspondence between LocalZO estimators and surrogate derivatives: LocalZO induces a surrogate derivative in expectation, and conversely, any admissible surrogate admits a suitable LocalZO representation.

Lemma 2 ([8]). *Assume further that $\int_0^\infty t^2\lambda(t)\,dt < \infty$. Then $\mathbb{E}_{z\sim\lambda}[G^2_{\mathrm{raw}}(u;z,\delta)]$ is a surrogate derivative. Moreover, let $\tilde{\lambda}$ be the distribution with corresponding density $\tilde{\lambda}(s) := \frac{1}{c}\int_{|s|}^\infty t\lambda(t)\,dt$, where c is the scaling constant such that $\int_{-\infty}^\infty \tilde{\lambda}(s)\,ds = 1$. Then the expected LocalZO estimator is the derivative of a smoothed Heaviside function induced by $\tilde{\lambda}$, namely, $\mathbb{E}_{z\sim\lambda}[G^2_{\mathrm{raw}}(u;z,\delta)] = \frac{d}{du}\mathbb{E}_{\tilde{z}\sim\tilde{\lambda}}[c\,h(u+\delta\tilde{z})]$.*

Lemma 3 ([8]). *Let g be a surrogate derivative. Suppose further that $c := -2\delta^2\int_0^\infty \frac{1}{z}g'(z\delta)\,dz < \infty$ and put $\lambda(z) = -\frac{\delta^2}{cz}g'(z\delta)$ (so that λ is a probability density function). Then g admits a LocalZO representation in expectation, namely, $c\,\mathbb{E}_{z\sim\lambda}[G^2_{\mathrm{raw}}(u;z,\delta)] = \mathbb{E}_{z\sim\lambda}[c\,G^2_{\mathrm{raw}}(u;z,\delta)] = g(u)$.*

A.2 Proof of the LocalZO–SG Smoothing Identity

We prove Lemma 1. Recall that $z \sim \lambda$ is drawn from an even density with a finite second moment, $c = \mathbb{E}[z^2]$, and

$$G^2(u;z,\delta) = \frac{h(u+\delta z) - h(u-\delta z)}{2\delta}\frac{z}{c}.$$

The induced density is

$$\tilde{\lambda}(s) = \frac{1}{c}\int_{|s|}^\infty t\lambda(t)\,dt,$$

and $h_\delta(u) = \mathbb{E}_{\tilde{z}\sim\tilde{\lambda}}[h(u+\delta\tilde{z})]$ with $g_\delta(u) = h'_\delta(u)$ whenever the derivative exists.

Proof. We first verify that $\tilde{\lambda}$ is normalized. By Fubini's theorem,

$$\int_{-\infty}^\infty \tilde{\lambda}(s)\,ds = \frac{2}{c}\int_0^\infty\left(\int_s^\infty t\lambda(t)\,dt\right)ds = \frac{2}{c}\int_0^\infty t^2\lambda(t)\,dt = \frac{\mathbb{E}[z^2]}{c} = 1.$$

Thus $\tilde{\lambda}$ is a probability density.

For almost every z,

$$h(u+\delta z) - h(u-\delta z) = \mathrm{sign}(z)\mathbf{1}_{\{|u|<\delta|z|\}},$$

and hence

$$G^2(u;z,\delta) = \frac{|z|}{2\delta c}\mathbf{1}_{\{|u|<\delta|z|\}}.$$

Using the evenness of λ, we obtain

$$\mathbb{E}_{z\sim\lambda}[G^2(u;z,\delta)] = \int_{|u|<|z|\delta}\frac{|z|}{2\delta c}\lambda(z)\,dz = \frac{1}{\delta c}\int_{|u|/\delta}^\infty t\lambda(t)\,dt = \frac{1}{\delta}\tilde{\lambda}\Big(\frac{u}{\delta}\Big).$$

On the other hand,

$$h_\delta(u) = \mathbb{P}(\tilde{z} > -u/\delta) = \int_{-u/\delta}^\infty \tilde{\lambda}(s)\,ds.$$

Therefore, for almost every u,

$$h'_\delta(u) = \frac{1}{\delta}\tilde{\lambda}\Big(-\frac{u}{\delta}\Big) = \frac{1}{\delta}\tilde{\lambda}\Big(\frac{u}{\delta}\Big),$$

where the last equality follows from the evenness of $\tilde{\lambda}$. Comparing the two expressions gives

$$\mathbb{E}_{z\sim\lambda}[G^2(u;z,\delta)] = h'_\delta(u) = g_\delta(u),$$

which proves Lemma 1. □

A.3 Mean Alignment Verification

We next verify the direct-training mean identity stated in Proposition 1. The statement is immediate for deterministic SG. For LocalZO, it follows from the normalized nodewise identity and the fact that the hard-trajectory BPTT gradient is a finite multilinear function of the local spike-derivative factors.

Proof of Proposition 1. For deterministic SG, the forward pass uses the hard-spike trajectory $s^{\mathrm{discrete}}(W_k;\xi)$, while the backward pass replaces the derivative of each hard threshold by g_δ. By the definition of the trajectory-conditioned backward map G_δ, the resulting single-sample direction is exactly

$$g(W_k;\xi) = G_\delta\big(s^{\mathrm{discrete}}(W_k;\xi), W_k;\xi\big).$$

Since no additional local perturbation randomness is present, the identity holds directly.

For LocalZO, conditional on $(\mathcal{F}_k,\xi)$ and the hard-spike trajectory, the unrolled spatiotemporal computation graph is fixed. The resulting reverse-mode derivative can be expanded over a finite collection $\mathcal{P}(W_k;\xi)$ of directed spatiotemporal paths:

$$\hat{g}(W_k;\xi,z) = \sum_{p\in\mathcal{P}(W_k;\xi)} a_p(W_k;\xi)\prod_{\ell\in p} G^2\big(u_\ell^{\mathrm{hard}}(W_k;\xi); z_\ell, \delta\big),$$

where $a_p(W_k;\xi)$ collects all factors on path p that are measurable with respect to $(\mathcal{F}_k,\xi)$ and independent of the LocalZO perturbations. Terms not involving a spike derivative are included through the empty product convention. The local perturbations $\{z_\ell\}$ are conditionally independent given $(\mathcal{F}_k,\xi)$.

Taking conditional expectation and using Lemma 1, we obtain

$$\begin{aligned}\mathbb{E}_z[\hat{g}(W_k;\xi,z)\mid\mathcal{F}_k,\xi] &= \sum_{p\in\mathcal{P}(W_k;\xi)} a_p(W_k;\xi)\prod_{\ell\in p}\mathbb{E}_z\big[G^2\big(u_\ell^{\mathrm{hard}}(W_k;\xi); z_\ell,\delta\big)\,\big|\,\mathcal{F}_k,\xi\big]\\ &= \sum_{p\in\mathcal{P}(W_k;\xi)} a_p(W_k;\xi)\prod_{\ell\in p} g_\delta\big(u_\ell^{\mathrm{hard}}(W_k;\xi)\big).\end{aligned}$$

The last expression is exactly the trajectory-conditioned backward signal obtained by applying the local derivative g_δ along the hard-spike forward trajectory. Hence,

$$\mathbb{E}_z[\hat{g}(W_k;\xi,z)\mid\mathcal{F}_k,\xi] = G_\delta\big(s^{\mathrm{discrete}}(W_k;\xi), W_k;\xi\big),$$

which proves (8). □

A.4 LocalZO Variance Bound

We now derive a distribution-dependent conditional variance bound for the LocalZO estimator. This result justifies Assumption 2 under the standard nodewise linear representation.

Proposition 4 (Distribution-dependent variance bound for LocalZO)**.** *Let the LocalZO local estimator be*

$$G^2(u;z,\delta) := \frac{h(u+\delta z)-h(u-\delta z)}{2\delta}\,\frac{z}{c}, \qquad c=\mathbb{E}[z^2],$$

and let $r := |u|/\delta$. Then

$$\mathbb{E}_z\big[G^2(u;z,\delta)^2\big] = \frac{1}{2\delta^2c^2}\int_r^\infty t^2\lambda(t)\,dt, \tag{12}$$

and

$$\mathrm{Var}_z\big(G^2(u;z,\delta)\big) = \frac{1}{2\delta^2c^2}\int_r^\infty t^2\lambda(t)\,dt - \left(\frac{1}{\delta c}\int_r^\infty t\lambda(t)\,dt\right)^2. \tag{13}$$

If the single-sample LocalZO backward direction admits the representation

$$\hat{g}(W_k;\xi,z) = \sum_{\ell=1}^{N(W_k;\xi)} a_\ell(W_k;\xi)\, G^2\big(u_\ell^{\mathrm{hard}}(W_k;\xi); z_\ell, \delta\big),$$

where the coefficients $a_\ell(W_k;\xi)$ are measurable with respect to $(\mathcal{F}_k,\xi)$, the local perturbations z_ℓ are conditionally independent, and

$$\sum_{\ell=1}^{N(W_k;\xi)} \|a_\ell(W_k;\xi)\|^2 \le C_{\mathrm{bp}}^2,$$

then

$$\mathbb{E}_z\Big[\|\hat{g}(W_k;\xi,z) - g(W_k;\xi)\|^2 \,\Big|\, \mathcal{F}_k,\xi\Big] \le C_{\mathrm{bp}}^2\, \bar{v}_\lambda(\delta), \qquad \bar{v}_\lambda(\delta) := \sup_{r\ge 0} \mathrm{Var}_z\big(G^2(u;z,\delta)\big). \tag{14}$$

In particular, Assumption 2 holds with $\sigma_{\mathrm{sg},1}^2 = C_{\mathrm{bp}}^2\, \bar{v}_\lambda(\delta)$.

Proof. Since

$$h(u+\delta z) - h(u-\delta z) = \mathrm{sign}(z)\cdot \mathbf{1}_{\{|u|<\delta|z|\}}$$

holds for almost every z, we have

$$G^2(u;z,\delta) = \frac{|z|}{2\delta c}\cdot \mathbf{1}_{\{|u|<\delta|z|\}}.$$

Hence

$$G^2(u;z,\delta)^2 = \frac{z^2}{4\delta^2 c^2}\cdot \mathbf{1}_{\{|u|<\delta|z|\}}.$$

Using the even symmetry of λ gives

$$\mathbb{E}_z\big[G^2(u;z,\delta)^2\big] = \frac{1}{2\delta^2 c^2}\int_r^\infty t^2\lambda(t)\,dt, \qquad r = \frac{|u|}{\delta},$$

which proves (12). Moreover, by Lemma 1,

$$\mathbb{E}_z[G^2(u;z,\delta)] = \frac{1}{\delta c}\int_r^\infty t\lambda(t)\,dt,$$

which yields (13).

Now define

$$X_\ell := G^2\big(u_\ell^{\mathrm{hard}}(W_k;\xi); z_\ell,\delta\big) - \mathbb{E}_z\big[G^2\big(u_\ell^{\mathrm{hard}}(W_k;\xi); z_\ell,\delta\big)\,\big|\,\mathcal{F}_k,\xi\big].$$

Then

$$\hat{g}(W_k;\xi,z) - g(W_k;\xi) = \sum_{\ell=1}^{N(W_k;\xi)} a_\ell(W_k;\xi)\, X_\ell.$$

Conditioned on $\mathcal{F}_k$ and ξ, the variables X_ℓ are independent across ℓ and have a zero conditional mean. Therefore

$$\mathbb{E}_z\Big[\|\hat{g}(W_k;\xi,z) - g(W_k;\xi)\|^2 \,\Big|\, \mathcal{F}_k,\xi\Big] = \sum_{\ell=1}^{N(W_k;\xi)} \|a_\ell(W_k;\xi)\|^2 \times \mathrm{Var}_z\big(G^2\big(u_\ell^{\mathrm{hard}}(W_k;\xi); z_\ell,\delta\big)\big).$$

By the definition of $\bar{v}_\lambda(\delta)$ and the bound on the squared chain-rule coefficients,

$$\mathbb{E}_z\Big[\|\hat{g}(W_k;\xi,z) - g(W_k;\xi)\|^2 \,\Big|\, \mathcal{F}_k,\xi\Big] \le C_{\mathrm{bp}}^2\, \bar{v}_\lambda(\delta).$$

This proves (14). □

The constant C_{bp} absorbs the size and gain of the unrolled backward graph, so the LocalZO variance bound should be interpreted as network-dependent rather than dimension-free.

Remark 1. For common unit-variance even distributions, (13) has explicit closed forms. Let $r = |u|/\delta$.

If $z \sim \mathcal{N}(0,1)$, then

$$\mathrm{Var}_z\big(G^2(u;z,\delta)\big) = \frac{r\phi(r) + 1 - \Phi(r)}{2\delta^2} - \frac{\phi(r)^2}{\delta^2},$$

where ϕ and Φ denote the standard normal density and cdf.

If $z \sim \mathrm{Unif}([-\sqrt{3}, \sqrt{3}])$, then

$$\mathrm{Var}_z\big(G^2(u;z,\delta)\big) = \frac{1}{4\delta^2}\left(1 - \frac{r^3}{3\sqrt{3}}\right)_+ - \frac{(3-r^2)_+^2}{48\delta^2}.$$

If $z \sim \mathrm{Laplace}(0, 1/\sqrt{2})$, then

$$\mathrm{Var}_z\big(G^2(u;z,\delta)\big) = \frac{e^{-\sqrt{2}r}(r^2 + \sqrt{2}r + 1)}{4\delta^2} - \frac{e^{-2\sqrt{2}r}(\sqrt{2}r+1)^2}{8\delta^2}.$$

Thus, for these common distributions, the LocalZO variance can be computed explicitly from λ and δ.

B Proof of the Smoothed-Objective Stationarity Result

This appendix proves the biased near-stationarity result for the smoothed objective. We first establish the variance decomposition of the aggregated estimator, and then prove Theorem 2.

B.1 Conditional Variance of the Aggregated Update

Lemma 4 (Conditional variance of the aggregated update)**.** *Under Assumption 2,*

$$\mathbb{E}\Big[\|\hat{g}_k - \mathbb{E}[\hat{g}_k \mid \mathcal{F}_k]\|^2 \,\Big|\, \mathcal{F}_k\Big] \le \sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m}. \tag{15}$$

Proof. Let

$$\hat{g}_{k,j,r} := \hat{g}(W_k; \xi_{k,j}, z_{k,j,r}), \qquad g_{k,j} := \mathbb{E}[\hat{g}_{k,j,r} \mid \mathcal{F}_k, \xi_{k,j}] = g(W_k; \xi_{k,j}),$$

and define

$$\hat{g}_{k,j} := \frac{1}{m}\sum_{r=1}^{m} \hat{g}_{k,j,r}, \qquad \hat{g}_k = \frac{1}{B}\sum_{j=1}^{B} \hat{g}_{k,j}, \qquad \bar{g}_k := \mathbb{E}[\hat{g}_k \mid \mathcal{F}_k, \xi_{k,1:B}] = \frac{1}{B}\sum_{j=1}^{B} g_{k,j}.$$

Then

$$\hat{g}_k - \mathbb{E}[\hat{g}_k \mid \mathcal{F}_k] = (\hat{g}_k - \bar{g}_k) + (\bar{g}_k - \mathbb{E}[\hat{g}_k \mid \mathcal{F}_k]).$$

Taking conditional squared norms and expanding, the cross term has zero conditional expectation, hence

$$\mathbb{E}\Big[\|\hat{g}_k - \mathbb{E}[\hat{g}_k \mid \mathcal{F}_k]\|^2 \,\Big|\, \mathcal{F}_k\Big] = \mathbb{E}\Big[\|\hat{g}_k - \bar{g}_k\|^2 \,\Big|\, \mathcal{F}_k\Big] + \mathbb{E}\Big[\|\bar{g}_k - \mathbb{E}[\hat{g}_k \mid \mathcal{F}_k]\|^2 \,\Big|\, \mathcal{F}_k\Big].$$

For the first term, conditioned on $\mathcal{F}_k$ and $\xi_{k,1:B}$, the random vectors $\hat{g}_{k,j} - g_{k,j}$ are independent across j and have zero conditional mean. Therefore

$$\mathbb{E}\Big[\|\hat{g}_k - \bar{g}_k\|^2 \,\Big|\, \mathcal{F}_k, \xi_{k,1:B}\Big] = \frac{1}{B^2}\sum_{j=1}^{B} \mathbb{E}\Big[\|\hat{g}_{k,j} - g_{k,j}\|^2 \,\Big|\, \mathcal{F}_k, \xi_{k,j}\Big].$$

Conditioned on $\mathcal{F}_k$ and $\xi_{k,j}$, the random vectors $\hat{g}_{k,j,r} - g_{k,j}$ are independent across r and have zero conditional mean. Hence

$$\mathbb{E}\Big[\|\hat{g}_{k,j} - g_{k,j}\|^2 \,\Big|\, \mathcal{F}_k, \xi_{k,j}\Big] = \frac{1}{m^2}\sum_{r=1}^{m} \mathbb{E}\Big[\|\hat{g}_{k,j,r} - g_{k,j}\|^2 \,\Big|\, \mathcal{F}_k, \xi_{k,j}\Big] \le \frac{\sigma_{\mathrm{sg},1}^2}{m}.$$

Substituting back and removing the conditioning yields

$$\mathbb{E}\Big[\|\hat{g}_k - \bar{g}_k\|^2 \,\Big|\, \mathcal{F}_k\Big] \le \frac{\sigma_{\mathrm{sg},1}^2}{Bm}.$$

For the second term, since

$$\mathbb{E}[\hat{g}_k \mid \mathcal{F}_k] = \mathbb{E}_\xi[g(W_k;\xi) \mid \mathcal{F}_k] \qquad \text{and} \qquad \bar{g}_k = \frac{1}{B}\sum_{j=1}^{B} g_{k,j},$$

we have

$$\mathbb{E}\Big[\|\bar{g}_k - \mathbb{E}[\hat{g}_k \mid \mathcal{F}_k]\|^2 \,\Big|\, \mathcal{F}_k\Big] = \mathbb{E}\left[\left\|\frac{1}{B}\sum_{j=1}^{B}\Big(g_{k,j} - \mathbb{E}_\xi[g(W_k;\xi) \mid \mathcal{F}_k]\Big)\right\|^2 \,\middle|\, \mathcal{F}_k\right].$$

Conditioned on $\mathcal{F}_k$, the summands are independent across j and have zero conditional mean, so

$$\mathbb{E}\Big[\|\bar{g}_k - \mathbb{E}[\hat{g}_k \mid \mathcal{F}_k]\|^2 \,\Big|\, \mathcal{F}_k\Big] \le \frac{\sigma_{\mathrm{mb},1}^2}{B}.$$

Combining the two bounds gives

$$\mathbb{E}\Big[\|\hat{g}_k - \mathbb{E}[\hat{g}_k \mid \mathcal{F}_k]\|^2 \,\Big|\, \mathcal{F}_k\Big] \le \frac{\sigma_{\mathrm{mb},1}^2}{B} + \frac{\sigma_{\mathrm{sg},1}^2}{Bm} = \sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m}.$$

This proves (15). □

B.2 Proof of Theorem 2

Proof. By L_δ-smoothness, for any k,

$$F_\delta(W_{k+1}) \le F_\delta(W_k) + \langle \nabla F_\delta(W_k), W_{k+1} - W_k\rangle + \frac{L_\delta}{2}\|W_{k+1} - W_k\|^2.$$

Substituting $W_{k+1} - W_k = -\eta\hat{g}_k$ yields

$$F_\delta(W_{k+1}) \le F_\delta(W_k) - \eta\,\langle \nabla F_\delta(W_k), \hat{g}_k\rangle + \frac{L_\delta\eta^2}{2}\|\hat{g}_k\|^2.$$

Taking conditional expectation with respect to $\mathcal{F}_k$, let

$$\mu_k = \mathbb{E}[\hat{g}_k \mid \mathcal{F}_k] = \nabla F_\delta(W_k) + b_k \qquad \text{and} \qquad \xi_k = \hat{g}_k - \mu_k,$$

so that $\mathbb{E}[\xi_k \mid \mathcal{F}_k] = 0$. Since

$$\|\hat{g}_k\|^2 = \|\mu_k + \xi_k\|^2 = \|\mu_k\|^2 + \|\xi_k\|^2 + 2\langle \mu_k, \xi_k\rangle,$$

and $\mathbb{E}[\langle \mu_k, \xi_k\rangle \mid \mathcal{F}_k] = 0$, we have

$$\mathbb{E}[F_\delta(W_{k+1}) \mid \mathcal{F}_k] \le F_\delta(W_k) - \eta\,\langle \nabla F_\delta(W_k), \mu_k\rangle + \frac{L_\delta\eta^2}{2}\|\mu_k\|^2 + \frac{L_\delta\eta^2}{2}\mathbb{E}[\|\xi_k\|^2 \mid \mathcal{F}_k].$$

Decompose

$$-\langle \nabla F_\delta(W_k), \mu_k\rangle = -\|\nabla F_\delta(W_k)\|^2 - \langle \nabla F_\delta(W_k), b_k\rangle.$$

Young's inequality gives

$$-\langle \nabla F_\delta(W_k), b_k\rangle \le \frac{1}{2}\|\nabla F_\delta(W_k)\|^2 + \frac{1}{2}\|b_k\|^2.$$

Moreover,

$$\|\mu_k\|^2 = \|\nabla F_\delta(W_k) + b_k\|^2 \le 2\|\nabla F_\delta(W_k)\|^2 + 2\|b_k\|^2.$$

Substituting these bounds and using Lemma 4 yields

$$\mathbb{E}[F_\delta(W_{k+1}) \mid \mathcal{F}_k] \le F_\delta(W_k) - \eta\Big(\frac{1}{2} - L_\delta\eta\Big)\,\|\nabla F_\delta(W_k)\|^2 + \eta\Big(\frac{1}{2} + L_\delta\eta\Big)\,\|b_k\|^2 + \frac{L_\delta\eta^2}{2}\left(\sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m}\right). \tag{16}$$

The condition $\eta \le 1/(4L_\delta)$ implies

$$\frac{1}{2} - L_\delta\eta \ge \frac{1}{4} \qquad \text{and} \qquad \frac{1}{2} + L_\delta\eta \le \frac{3}{4}.$$

Taking full expectation and summing from $k = 0$ to $K-1$, and using $\mathbb{E}[F_\delta(W_K)] \ge F_\delta^\star$, yields

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\Big[\|\nabla F_\delta(W_k)\|^2\Big] \le \frac{4\big(F_\delta(W_0) - F_\delta^\star\big)}{\eta K} + 2L_\delta\eta\left(\sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m}\right) + \frac{3}{K}\sum_{k=0}^{K-1}\mathbb{E}\Big[\|b_k\|^2\Big].$$

This proves (9). □

C Proof of the Systematic Bias Bound

This appendix proves the systematic bias bound in Proposition 2. We first control the pointwise discrepancy between the hard spike and the smoothed spike, then lift this estimate to the trajectory level, and finally combine it with Proposition 1 and the Lipschitz continuity of the trajectory-conditioned backward map.

C.1 Pointwise Tail Bound

Lemma 5 (Pointwise discrete–smooth spike discrepancy)**.** *For every $u \in \mathbb{R}$,*

$$|h(u) - h_\delta(u)| \le \Psi\left(\frac{|u|}{\delta}\right). \tag{17}$$

Proof. Recall that

$$h(u) = \mathbf{1}\{u > 0\}, \qquad h_\delta(u) = \mathbb{P}(\tilde{z} > -u/\delta),$$

where $\tilde{z} \sim \tilde{\lambda}$ and

$$\Psi(a) = \mathbb{P}(|\tilde{z}| \ge a), \qquad a \ge 0.$$

If $u > 0$, then $h(u) = 1$, and therefore

$$|h(u) - h_\delta(u)| = 1 - \mathbb{P}(\tilde{z} > -u/\delta) = \mathbb{P}(\tilde{z} \le -u/\delta) \le \mathbb{P}(|\tilde{z}| \ge u/\delta) = \Psi\left(\frac{|u|}{\delta}\right).$$

If $u \le 0$, then $h(u) = 0$, and hence

$$|h(u) - h_\delta(u)| = \mathbb{P}(\tilde{z} > -u/\delta) \le \mathbb{P}(|\tilde{z}| \ge -u/\delta) = \Psi\left(\frac{|u|}{\delta}\right).$$

This proves (17). □

C.2 Trajectory Recursion

Lemma 6 (Trajectory discrepancy bound)**.** *Under Assumption 3(ii)–(iii), suppose either:*

- *(a) the contraction condition $L_h L_u < 1$ holds, in which case $\Lambda_\delta = 1/(1 - L_h L_u)$; or*
- *(b) the finite-horizon causal condition in Assumption 3(iv) holds, in which case $\Lambda_\delta = \sum_{t=0}^{T-1}(L_h L_{\mathrm{c}})^t$.*

Then, for every parameter W and sample ξ,

$$\left\|s^{\mathrm{discrete}}(W;\xi) - s^{\mathrm{smooth}}(W;\xi)\right\|_2^2 \le \Lambda_\delta^2 \sum_{i=1}^{M} \Psi\left(\frac{|u_i^{\mathrm{hard}}(W;\xi)|}{\delta}\right). \tag{18}$$

Proof. We write

$$s^d = s^{\mathrm{discrete}}(W;\xi), \qquad s^s = s^{\mathrm{smooth}}(W;\xi), \qquad u^d = u(s^d, W;\xi), \qquad u^s = u(s^s, W;\xi).$$

Case (a): contraction branch. Since $s^d = h(u^d)$ and $s^s = h_\delta(u^s)$, Assumption 3(ii)–(iii) gives

$$\begin{aligned}\left\|s^d - s^s\right\|_2 &\le \left\|h(u^d) - h_\delta(u^d)\right\|_2 + \left\|h_\delta(u^d) - h_\delta(u^s)\right\|_2 \\ &\le \left\|h(u^d) - h_\delta(u^d)\right\|_2 + L_h L_u \left\|s^d - s^s\right\|_2.\end{aligned}$$

Therefore,

$$\left\|s^d - s^s\right\|_2 \le \frac{1}{1 - L_h L_u} \left\|h(u^d) - h_\delta(u^d)\right\|_2 = \Lambda_\delta \left\|h(u^d) - h_\delta(u^d)\right\|_2.$$

Squaring both sides and applying Lemma 5 together with $\Psi(\cdot) \in [0,1]$ yields

$$\left\|s^d - s^s\right\|_2^2 \le \Lambda_\delta^2 \sum_{i=1}^{M} |h(u_i^d) - h_\delta(u_i^d)|^2 \le \Lambda_\delta^2 \sum_{i=1}^{M} \Psi\left(\frac{|u_i^{\mathrm{hard}}(W;\xi)|}{\delta}\right).$$

Case (b): finite-horizon causal branch. Let s_t and u_t denote the temporal blocks at time step t, and define

$$x_t := \left\|s_t^{\mathrm{discrete}} - s_t^{\mathrm{smooth}}\right\|_2, \qquad d_t := \left\|h(u_t^{\mathrm{hard}}) - h_\delta(u_t^{\mathrm{hard}})\right\|_2.$$

Then, for each $t = 1, \dots, T$,

$$x_t \le d_t + L_h L_{\mathrm{c}}\, x_{t-1}, \qquad x_0 = 0.$$

Unrolling the recursion yields

$$x_t \le \sum_{r=0}^{t-1} (L_h L_{\mathrm{c}})^r d_{t-r}.$$

Define the kernel sequence $k_r = (L_h L_{\mathrm{c}})^r$. By the discrete Young inequality,

$$\sum_{t=1}^{T} x_t^2 \le \left(\sum_{r=0}^{T-1} (L_h L_{\mathrm{c}})^r\right)^2 \sum_{t=1}^{T} d_t^2 = \Lambda_\delta^2 \sum_{t=1}^{T} d_t^2.$$

Since

$$\left\|s^{\mathrm{discrete}}(W;\xi) - s^{\mathrm{smooth}}(W;\xi)\right\|_2^2 = \sum_{t=1}^{T} x_t^2,$$

using Lemma 5 and $\Psi(\cdot) \in [0,1]$ gives

$$\left\|s^{\mathrm{discrete}}(W;\xi) - s^{\mathrm{smooth}}(W;\xi)\right\|_2^2 \le \Lambda_\delta^2 \sum_{i=1}^{M} \Psi\left(\frac{|u_i^{\mathrm{hard}}(W;\xi)|}{\delta}\right).$$

This proves (18). □

C.3 Proof of Proposition 2

Proof. Fix $\mathcal{F}_k$. Since mini-batch sampling is unbiased,

$$\mu_k = \mathbb{E}_\xi[g(W_k;\xi) \mid \mathcal{F}_k].$$

By Proposition 1,

$$g(W_k;\xi) = G_\delta\big(s^{\mathrm{discrete}}(W_k;\xi), W_k;\xi\big).$$

On the other hand, by the definition of the smoothed sample gradient and Assumption 1,

$$\nabla F_\delta(W_k) = \mathbb{E}_\xi\big[G_\delta\big(s^{\mathrm{smooth}}(W_k;\xi), W_k;\xi\big) \,\big|\, \mathcal{F}_k\big].$$

Substituting into $b_k = \mu_k - \nabla F_\delta(W_k)$, we obtain

$$b_k = \mathbb{E}_\xi[\Delta_k(\xi) \mid \mathcal{F}_k],$$

where

$$\Delta_k(\xi) := G_\delta\big(s^{\mathrm{discrete}}(W_k;\xi), W_k;\xi\big) - G_\delta\big(s^{\mathrm{smooth}}(W_k;\xi), W_k;\xi\big).$$

By Jensen's inequality and Assumption 3(i),

$$\mathbb{E}\,\|b_k\|^2 \le \mathbb{E}\Big[\mathbb{E}_\xi\Big[\|\Delta_k(\xi)\|^2 \,\Big|\, \mathcal{F}_k\Big]\Big] \le \Gamma^2\mathbb{E}\Big[\mathbb{E}_\xi\Big[\big\|s^{\mathrm{discrete}}(W_k;\xi) - s^{\mathrm{smooth}}(W_k;\xi)\big\|_2^2 \,\Big|\, \mathcal{F}_k\Big]\Big].$$

Applying Lemma 6 and the definition of $\mathcal{T}_\delta(k)$ gives

$$\mathbb{E}\,\|b_k\|^2 \le \Gamma^2\Lambda_\delta^2\mathbb{E}\left[\mathbb{E}_\xi\left[\sum_{i=1}^{M}\Psi\Big(\frac{|u_i^{\mathrm{hard}}(W_k;\xi)|}{\delta}\Big) \,\middle|\, \mathcal{F}_k\right]\right] = \Gamma^2\Lambda_\delta^2\,\mathcal{T}_\delta(k).$$

Equivalently,

$$\mathbb{E}\,\|b_k\|^2 \le \mathcal{B}_\delta^2(k),$$

which proves (10). □

D Proof of Proposition 3

This appendix proves the objective-level transfer bound between the smoothed objective and the original discrete objective.

Proof. For any fixed parameter W and sample ξ,

$$L(\mathrm{SNN}(x;W), y) = \ell(s^{\mathrm{discrete}}(W;\xi), W;\xi), \qquad L(\mathrm{SNN}_\delta(x;W), y) = \ell(s^{\mathrm{smooth}}(W;\xi), W;\xi).$$

By Assumption 4,

$$\big|\ell(s^{\mathrm{discrete}}(W;\xi), W;\xi) - \ell(s^{\mathrm{smooth}}(W;\xi), W;\xi)\big| \le L_s\,\big\|s^{\mathrm{discrete}}(W;\xi) - s^{\mathrm{smooth}}(W;\xi)\big\|_2.$$

Taking expectation over ξ and applying Cauchy–Schwarz gives

$$\big|F(W) - F_\delta(W)\big| \le L_s\sqrt{\mathbb{E}_\xi\Big[\|s^{\mathrm{discrete}}(W;\xi) - s^{\mathrm{smooth}}(W;\xi)\|_2^2\Big]}.$$

Applying Lemma 6, we obtain for any k,

$$\big|F(W_k) - F_\delta(W_k)\big| \le L_s\Lambda_\delta\sqrt{\mathcal{T}_\delta(W_k)}.$$

Now let $R \sim \mathrm{Unif}\{0,\dots,K-1\}$ be independent of the optimization randomness. Then

$$\mathbb{E}\big|F(W_R) - F_\delta(W_R)\big| = \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\big|F(W_k) - F_\delta(W_k)\big| \le \frac{L_s\Lambda_\delta}{K}\sum_{k=0}^{K-1}\mathbb{E}\Big[\sqrt{\mathcal{T}_\delta(W_k)}\Big].$$

Using Jensen's inequality and then the concavity of the square root yields

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\Big[\sqrt{\mathcal{T}_\delta(W_k)}\Big] \le \frac{1}{K}\sum_{k=0}^{K-1}\sqrt{\mathbb{E}[\mathcal{T}_\delta(W_k)]} \le \sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\mathcal{T}_\delta(k)}.$$

Therefore,

$$\mathbb{E}\big|F(W_R) - F_\delta(W_R)\big| \le L_s\Lambda_\delta\sqrt{\frac{1}{K}\sum_{k=0}^{K-1}\mathcal{T}_\delta(k)},$$

which proves (11). □

E Proof of the Main Theorem

This appendix assembles the smoothed-objective stationarity bound, the systematic bias bound, and the discrete–smooth consistency bound to prove the final convergence theorem for the discrete objective.

Proof. For any parameter vector W, we decompose

$$F(W) - F(\bar{W}) = \big(F_\delta(W) - F_\delta(\bar{W})\big) + \big(F(W) - F_\delta(W)\big) + \big(F_\delta(\bar{W}) - F(\bar{W})\big).$$

Applying this identity at $W = W_R$ and taking expectation yields

$$\mathbb{E}[F(W_R)] - F(\bar{W}) \le \mathbb{E}\big[F_\delta(W_R) - F_\delta(\bar{W})\big] + \mathbb{E}|F(W_R) - F_\delta(W_R)| + \big|F_\delta(\bar{W}) - F(\bar{W})\big|. \quad (19)$$

We first bound the smoothed-objective term. Since $R \sim \mathrm{Unif}\{0, \dots, K-1\}$ and $W_k \in \mathcal{U}$ almost surely for all $k = 0, 1, \dots, K$, Assumption 5 gives

$$\mathbb{E}\big[F_\delta(W_R) - F_\delta(\bar{W})\big] \le \frac{1}{2\mu_{\mathcal{U}}} \mathbb{E}\Big[\|\nabla F_\delta(W_R)\|^2\Big].$$

Because R is uniform on $\{0, \dots, K-1\}$,

$$\mathbb{E}\Big[\|\nabla F_\delta(W_R)\|^2\Big] = \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\Big[\|\nabla F_\delta(W_k)\|^2\Big].$$

We next derive an average gradient bound relative to the basin reference value $F_\delta(\bar{W})$. From the descent inequality (16) in the proof of Theorem 2, the step-size condition $\eta \le 1/(4L_\delta)$, and the bias bound $\mathbb{E}\,\|b_k\|^2 \le \mathcal{B}_\delta^2(k)$ from Proposition 2, we obtain

$$\frac{\eta}{4} \mathbb{E}\Big[\|\nabla F_\delta(W_k)\|^2\Big] \le \mathbb{E}[F_\delta(W_k)] - \mathbb{E}[F_\delta(W_{k+1})] + \frac{3\eta}{4} \mathcal{B}_\delta^2(k) + \frac{L_\delta \eta^2}{2} \left(\sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m}\right).$$

Summing over $k = 0, 1, \dots, K-1$ gives

$$\frac{\eta}{4} \sum_{k=0}^{K-1} \mathbb{E}\Big[\|\nabla F_\delta(W_k)\|^2\Big] \le F_\delta(W_0) - \mathbb{E}[F_\delta(W_K)] + \frac{3\eta}{4} \sum_{k=0}^{K-1} \mathcal{B}_\delta^2(k) + \frac{K L_\delta \eta^2}{2} \left(\sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m}\right).$$

Since $W_K \in \mathcal{U}$ almost surely and $\bar{W} \in \arg\min_{W \in \mathcal{U}} F_\delta(W)$, we have

$$F_\delta(W_K) \ge F_\delta(\bar{W}) \qquad \text{almost surely.}$$

Therefore,

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\Big[\|\nabla F_\delta(W_k)\|^2\Big] \le \frac{4\big(F_\delta(W_0) - F_\delta(\bar{W})\big)}{\eta K} + 2L_\delta \eta \left(\sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m}\right) + \frac{3}{K} \sum_{k=0}^{K-1} \mathcal{B}_\delta^2(k).$$

Consequently,

$$\mathbb{E}\big[F_\delta(W_R) - F_\delta(\bar{W})\big] \le \frac{1}{2\mu_{\mathcal{U}}} \left[\frac{4\big(F_\delta(W_0) - F_\delta(\bar{W})\big)}{\eta K} + 2L_\delta \eta \left(\sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m}\right) + \frac{3}{K} \sum_{k=0}^{K-1} \mathcal{B}_\delta^2(k)\right].$$

For the second term in (19), Proposition 3 yields

$$\mathbb{E}|F(W_R) - F_\delta(W_R)| \le L_s \Lambda_\delta \sqrt{\frac{1}{K} \sum_{k=0}^{K-1} \mathcal{T}_\delta(k)}.$$

The third term in (19) is precisely the residual discrete–smooth mismatch at the reference point:

$$\big|F_\delta(\bar{W}) - F(\bar{W})\big| = \big|F(\bar{W}) - F_\delta(\bar{W})\big|.$$

Substituting the three bounds above into (19) proves (7). □

E.1 Threshold-Tail Scaling and the Reference Residual

Exact one-spike discrepancy. For the even smoothing density $\tilde{\lambda}$, symmetry gives the exact identity

$$|h(u) - h_\delta(u)| = \frac{1}{2}\Psi\left(\frac{|u|}{\delta}\right), \qquad u \in \mathbb{R}. \tag{20}$$

Indeed, for $u > 0$ the discrepancy is $\mathbb{P}(\tilde{z} \le -u/\delta)$, and for $u < 0$ it is $\mathbb{P}(\tilde{z} > |u|/\delta)$; each is half the two-sided tail. At $u = 0$, both sides equal $1/2$. Thus, for one spike with loss $\ell(s) = s$, the objective gap is exactly $\frac{1}{2}\Psi(|u(W)|/\delta)$. This example exhibits exact tail dependence at the local and objective levels. The network bounds additionally account for trajectory and loss sensitivities. A positive lower bound on network mismatch would require corresponding non-degeneracy conditions: for example, a backward map independent of s gives $b_k = 0$ even when $\mathcal{T}_\delta(k) > 0$.

ATan smoothing at a fixed margin. For the ATan family in Appendix F.1, fix a margin $|u| = \gamma > 0$ and scale $\alpha > 0$. Its tail satisfies

$$\Psi_\alpha(\gamma/\delta) = 1 - \frac{2}{\pi}\arctan\Big(\frac{\pi\alpha\gamma}{2\delta}\Big) = \frac{4\delta}{\pi^2\alpha\gamma} + o(\delta), \qquad \delta \to 0. \tag{21}$$

The scalar spike discrepancy is therefore $O(\delta)$. For one isolated node, $\Lambda_\delta = 1$, and the general squared-trajectory bound followed by Cauchy–Schwarz yields the more conservative $O(\sqrt{\delta})$ objective-gap bound for fixed loss regularity. At the same time, $L_h = \sup_u |g_{\delta,\alpha}(u)| = \alpha/(2\delta)$. Consequently, the finite-horizon factor $\sum_{t=0}^{T-1}(L_h L_c)^t$ may increase as δ decreases. The tail, regularity constants, and admissible stepsize together determine the bound; their combination need not improve monotonically with sharper smoothing. Uniform worst-case sensitivities can also overestimate realized amplification in deep or long-horizon networks.

Reference residual and optimal values. At a one-spike reference point with $|u(\bar{W})| = \gamma > 0$ and an L_s-Lipschitz loss, the reference residual is at most $(L_s/2)\Psi_\alpha(\gamma/\delta) = O(\delta)$; equality holds for $\ell(s) = s$ and $L_s = 1$. More generally, a uniform approximation $\sup_{W\in\mathcal{U}}|F(W) - F_\delta(W)| \le \varepsilon_\delta$ would imply $|\inf_{W\in\mathcal{U}} F(W) - \inf_{W\in\mathcal{U}} F_\delta(W)| \le \varepsilon_\delta$ by taking infima in $F_\delta - \varepsilon_\delta \le F \le F_\delta + \varepsilon_\delta$. This uniform condition is stronger than tail control along the iterates and at a single reference point. In particular, the scalar spike discrepancy remains $1/2$ at a threshold crossing for every $\delta > 0$. Theorem 1 therefore retains the reference residual explicitly while quantifying the trajectory-dependent terms separately.

F Additional Experimental Results

F.1 Experimental Details

This appendix records the experimental protocol used in Section 7 and provides the additional diagnostic results omitted from the main text.

Datasets and Models We evaluate on representative frame-based and event-based spiking recognition benchmarks: Sequential CIFAR-10/100 (SCIFAR-10/100) [35], CIFAR10-DVS [36], DVSGesture [37], SHD [38], and TinyImageNet [39]. The corresponding backbones follow standard task-specific SNN architectures used in prior work, including `SCNN` [40], `7B-Net` [41], `SpikingVGG` [42], `SHDSNN` [18, 43], `SEWResNet34` [41, 43]. For architectures with configurable neuron types, we use the default `MELIF` setting described in prior work [43]. Detailed task-specific hyperparameters are summarized in Table 3.

Baselines All methods are implemented within a unified Lightning-based memory-efficient direct-training framework [43]. This framework supports direct BPTT training with discrete spiking forward dynamics, memory optimization, and task-specific training recipes. The reference training paradigm is surrogate-gradient (SG) learning, and LocalZO is implemented in the same framework for matched comparisons under identical runtime infrastructure [43, 8]. Unless otherwise stated, spike compression is enabled for all experiments [43].

Compute resources. All experiments were run on a machine equipped with a cluster of NVIDIA A100-SXM4-80GB GPUs. Unless otherwise stated, each training or diagnostic run used a single GPU; multiple seeds were run independently and, when available, in parallel on different GPUs of the same server. The implementation follows a Lightning-based memory-efficient SNN training framework, and spike compression is enabled unless otherwise stated. Wall-clock time was measured from the start of training to the final reported checkpoint, excluding data downloading and one-time environment setup. For diagnostic-only probes, wall-clock time was measured from the start of checkpoint loading to the end of the corresponding probe evaluation. GPU-hours are computed as

$$\text{GPU-hours} = \sum_i (\#\text{GPUs used in run } i) \times (\text{wall-clock hours of run } i).$$

Common protocol. Training uses hard spikes in the forward pass. SG replaces the hard-threshold derivative with a surrogate derivative in the backward pass; LocalZO replaces it with the local perturbation estimator in Eq. (24). Unless otherwise stated, diagnostic quantities are evaluated at frozen checkpoints in evaluation mode on validation mini-batches. Training-loss curves are the standard training logs. For gradient-based probes, all trainable-parameter gradients are flattened before computing norms, squared differences, or cosine similarities.

Spike surrogate and smoothing Let u denote the signed distance from the pre-spike membrane potential to the firing threshold. All reported diagnostic experiments use the ATan surrogate with scale $\alpha = 2.0$ and default smoothing width $\delta = 1.0$. The smooth spike and its derivative are

$$h_{\delta,\alpha}(u) = \frac{1}{2} + \frac{1}{\pi}\arctan\Big(\frac{\pi\alpha}{2}\frac{u}{\delta}\Big), \qquad \frac{\partial h_{\delta,\alpha}(u)}{\partial u} = \frac{\alpha}{2\delta\left[1 + \left(\frac{\pi\alpha}{2}\frac{u}{\delta}\right)^2\right]}. \tag{22}$$

When evaluating F_δ, each hard spiking nonlinearity in the recurrent forward dynamics is replaced by $h_{\delta,\alpha}$, while the parameters, data, and loss are kept fixed.

LocalZO estimator LocalZO uses the perturbation distribution matched to the ATan surrogate. Specifically,

$$z \sim \lambda_{\mathrm{ATan},\alpha}, \qquad z = \frac{2}{\pi\alpha\sqrt{3}}t, \qquad t \sim t_3, \qquad c_\alpha = \mathbb{E}[z^2] = \frac{4}{\pi^2\alpha^2}, \tag{23}$$

where t_3 is the Student-t distribution with three degrees of freedom. At a spiking node, the m-sample LocalZO derivative estimator is

$$\widehat{h}'_{\mathrm{LZO}}(u) = \frac{1}{m}\sum_{j=1}^{m}\frac{H(u+\delta z_j) - H(u-\delta z_j)}{2\delta}\frac{z_j}{c_\alpha}, \qquad z_j \overset{\text{i.i.d.}}{\sim} \lambda_{\mathrm{ATan},\alpha}, \tag{24}$$

with $H(u) = \mathbf{1}\{u > 0\}$. This normalization makes the nodewise LocalZO mean equal to the ATan surrogate derivative, so increasing m reduces perturbation variance without changing the target mean direction.

Threshold-tail diagnostic For ATan smoothing, the normalized near-threshold statistic reported in the diagnostic tables is

$$D(k) = \frac{1}{|\mathcal{B}_k|M}\sum_{\xi\in\mathcal{B}_k}\sum_{i=1}^{M}\Psi_\alpha\left(\frac{|u_i^{\mathrm{hard}}(W_k;\xi)|}{\delta}\right), \qquad \Psi_\alpha(r) = 1 - \frac{2}{\pi}\arctan\Big(\frac{\pi\alpha}{2}r\Big), \tag{25}$$

where M is the number of traced spatiotemporal membrane states. This is the empirical normalized counterpart of the threshold-tail functional $\mathcal{T}_\delta(k)$. In the objective-gap plot, graph-size and Lipschitz constants are absorbed into a single calibration constant $\widehat{C}$; the calibration is used only to compare the scale and temporal trend of the empirical gap with the theorem-motivated threshold-tail proxy.

F.2 SG–LocalZO Alignment and Convergence

This experiment corresponds to Fig. 2. We train Sequential CIFAR-10 with the SCNN backbone under the common settings in Table 2. SG uses a hard-spike forward pass and the ATan surrogate

Table 2: Shared settings for the diagnostic experiments.

Item	Setting
Hard spike	$H(u) = \mathbf{1}\{u > 0\}$
Default surrogate	ATan surrogate in Eq. (22)
Surrogate scale and width	$\alpha = 2.0$, $\delta = 1.0$ unless δ is explicitly swept
LocalZO perturbation distribution	$\lambda_{\mathrm{ATan},\alpha}$ in Eq. (23)
LocalZO sample sizes in Fig. 2	$m \in \{5, 10, 20, 40\}$
Alignment probes	4 validation mini-batches; cosine after flattening all trainable gradients
Objective-gap probes	matched hard and smoothed forward passes at the same checkpoint
Gradient-bias probes	matched hard-trajectory and smooth-trajectory backward maps on the same mini-batches
Cross-model seeds	$1, 2, 3$ when all seeds are available

Table 3: Task-level training configurations used by the diagnostic runs. Code-level class names are shown for reproducibility; in the main text, `GCSequentialCIFARNet` is referred to as SCNN. Here β denotes the LIF decay factor and is distinct from the LocalZO perturbation distribution $\lambda_{\mathrm{ATan},\alpha}$.

Dataset	Backbone	Neuron	T	β	Loss	Optimizer	Scheduler	Batch / Epochs
Sequential CIFAR-10/100	`GCSequentialCIFARNet`	MELIF	32	0.5	CE	SGD, lr 0.1, mom. 0.9	cosine	128 / 300
DVSGesture	`GCSEWResNet`	MELIF	16	0.5	CE	SGD, lr 10^{-3}, mom. 0.9	step$(64, 0.1)$	16 / 192
CIFAR10-DVS	`GCCIFAR10DVSVGG`	MELIF	10	0.25	TET	SGD, lr 0.1, mom. 0.9, wd 5×10^{-4}	cosine	32 / 100
SHD	`GCDHLIFSFNN`	DHLIF	100	learned	CE	Adam, lr 10^{-2}	step$(20, 0.5)$	128 / 100
TinyImageNet	`GCSEWResNet34`	MELIF	4	0.5	TET	SGD, lr 0.025, mom. 0.9	cosine	32 / 120

backward pass. LocalZO uses the same hard-spike forward pass and replaces the surrogate derivative by Eq. (24), with $m \in \{5, 10, 20, 40\}$.

At each probe epoch, the current checkpoint is frozen and both update directions are evaluated on the same validation mini-batches. The LocalZO direction is averaged over repeated LocalZO backward evaluations, using $\max(8, m)$ repetitions unless explicitly overridden. The reported alignment is

$$\mathrm{Align}_k(m) = \frac{\left\langle \widehat{g}_{\mathrm{SG}}(W_k), \widehat{g}^{(m)}_{\mathrm{LZO}}(W_k) \right\rangle}{\|\widehat{g}_{\mathrm{SG}}(W_k)\|_2 \left\|\widehat{g}^{(m)}_{\mathrm{LZO}}(W_k)\right\|_2}. \tag{26}$$

To complement the cosine metric, we also measure the componentwise mean-squared discrepancy between the flattened SG and LocalZO update directions:

$$\mathrm{MSE}_k(m) = \frac{1}{d}\left\|\widehat{g}_{\mathrm{SG}}(W_k) - \widehat{g}^{(m)}_{\mathrm{LZO}}(W_k)\right\|_2^2, \tag{27}$$

where d is the number of trainable parameters. For the query-budget view, one SG update counts as one hard-threshold evaluation, while one LocalZO update with m perturbation samples counts as $1 + 2m$ hard-threshold evaluations. This view is an estimator-cost diagnostic rather than a wall-clock efficiency comparison.

Figure 5 reinforces the main-text observation from Fig. 2. As m increases, LocalZO shows smaller absolute update discrepancy and stronger directional agreement with SG, while requiring more hard-threshold evaluations per update. These trends are consistent with the perturbation-variance term in the bound and support the view that LocalZO follows the same mean direct-training mechanism as SG with an m-dependent variance component.

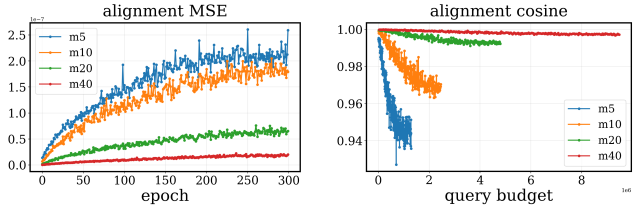

Figure 5: Additional SG–LocalZO diagnostics: update-direction MSE over epochs and cosine alignment under cumulative query budget.

F.3 Discrete–Smooth Objective Consistency Gap

This experiment corresponds to Fig. 3 and evaluates the objective-level transfer term in Proposition 3. For a frozen checkpoint W_k, we evaluate the hard and smooth empirical objectives on the same examples,

$$\begin{aligned}\widehat{F}(W_k) &= \frac{1}{|\mathcal{D}_k|}\sum_{(x,y)\in\mathcal{D}_k}\ell(f_{\mathrm{discrete}}(x;W_k),y)\,,\\ \widehat{F}_\delta(W_k) &= \frac{1}{|\mathcal{D}_k|}\sum_{(x,y)\in\mathcal{D}_k}\ell(f_{\mathrm{smooth},\delta}(x;W_k),y)\,,\end{aligned} \tag{28}$$

where $f_{\mathrm{smooth},\delta}$ is obtained by replacing hard spikes with $h_{\delta,\alpha}$ in the forward dynamics. The empirical consistency gap is

$$G_k = \left|\widehat{F}(W_k) - \widehat{F}_\delta(W_k)\right|. \tag{29}$$

The threshold-tail proxy plotted in the main text is

$$B_k = \left(\frac{1}{k+1}\sum_{j=0}^{k}\widehat{\mathcal{T}}_\delta(j)\right)^{1/2}, \qquad \widehat{C} = \max_{k:B_k>0}\frac{G_k}{B_k}. \tag{30}$$

The constant $\widehat{C}$ is a fixed visualization scale for the plotted run, absorbing the unknown constants in the theoretical bound. It is selected from the plotted run, so the comparison describes within-run scale and temporal trends of the gap and the threshold-tail proxy.

F.4 Fixed-Checkpoint Smoothing Residual

This experiment corresponds to Fig. 4 and isolates the final residual $|F(\bar{W}) - F_\delta(\bar{W})|$ in Theorem 1. We freeze a trained checkpoint W_{fix} and vary only the smoothing width δ. For each value of δ, we evaluate

$$\widehat{R}_{\mathrm{fix}}(\delta) = \left|\widehat{F}(W_{\mathrm{fix}}) - \widehat{F}_\delta(W_{\mathrm{fix}})\right|. \tag{31}$$

The sweep uses $\delta \in \{1, 1/2, 1/4, 1/8, 1/16, 1/32\}$ unless an explicit list is supplied. The checkpoint, data split, model parameters, and loss function are fixed throughout the sweep. Thus, changes in $\widehat{R}_{\mathrm{fix}}(\delta)$ measure the smoothing approximation error rather than an optimization effect.

F.5 Gradient-Level Bias Diagnostic

Theorem 1 contains the gradient-level residual $\bar{\mathcal{B}}^2_{\delta,K}$, which comes from applying the backward map along the hard trajectory while ∇F_δ is defined by the smoothed trajectory. In the notation of Proposition 2,

$$b_k = \mu_k - \nabla F_\delta(W_k) = \mathbb{E}_\xi\left[G_\delta\big(s^{\mathrm{discrete}}(W_k;\xi),W_k;\xi\big) - G_\delta\big(s^{\mathrm{smooth}}(W_k;\xi),W_k;\xi\big)\right], \tag{32}$$

and the theory gives

$$\mathbb{E}\left[\|b_k\|^2\right] \le \Gamma^2\Lambda_\delta^2\mathcal{T}_\delta(k). \tag{33}$$

For a probe mini-batch $\mathcal{B}_k$, we measure the squared hard–smooth backward mismatch

$$\widehat{\mathcal{M}}_\delta(k) = \left\|\frac{1}{|\mathcal{B}_k|}\sum_{\xi\in\mathcal{B}_k}\left[G_\delta\big(s^{\mathrm{discrete}}(W_k;\xi),W_k;\xi\big) - G_\delta\big(s^{\mathrm{smooth}}(W_k;\xi),W_k;\xi\big)\right]\right\|_2^2. \tag{34}$$

Both terms are evaluated at the same checkpoint and on the same mini-batches.

We use Sequential CIFAR-10 with the SCNN backbone. REF is the unmodified hard-spike trajectory. A/B/C are counterfactual threshold-distance interventions evaluated on the same checkpoint and sampled validation mini-batches. For each traced signed threshold distance u, the intervention modifies only the window $|u| \le r$, with $r = 2\delta$:

$$u' = \mathrm{sign}(u)\, r\left(\frac{|u|}{r}\right)^{\gamma_g}, \qquad g \in \{A, B, C\}, \tag{35}$$

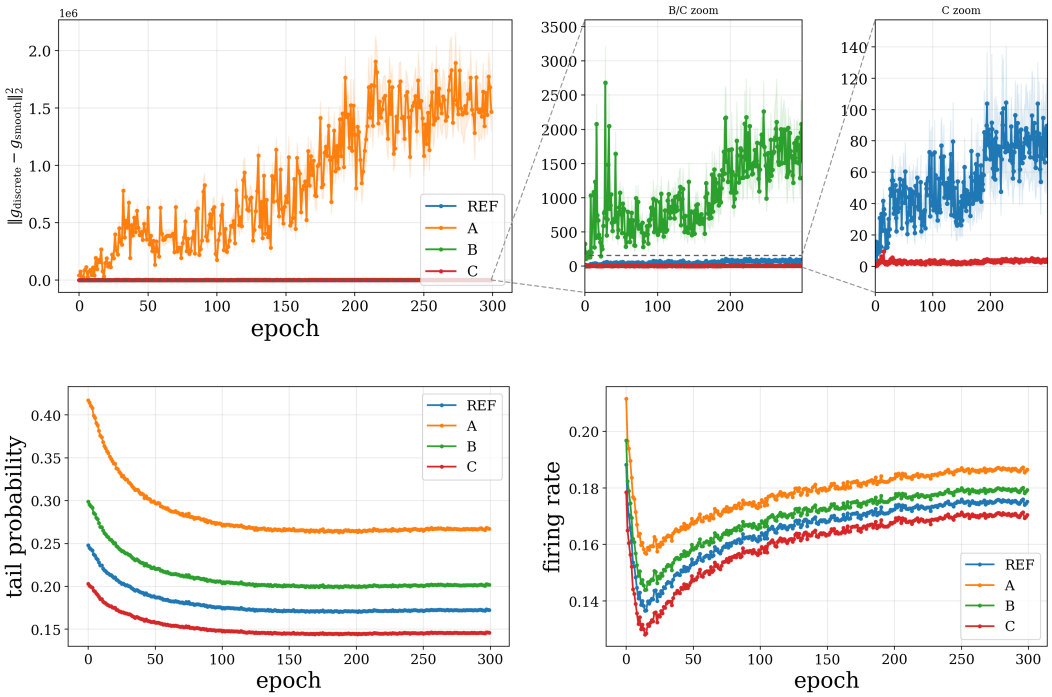

Figure 6: Gradient-level bias diagnostic on Sequential CIFAR-10 with an SCNN. Top: empirical squared hard–smooth backward mismatch $\widehat{\mathcal{M}}_\delta(k)$. Bottom left: normalized threshold-tail diagnostic $D(k)$. Bottom right: mean firing rate $\rho(k)$. REF is the original trajectory; A/B/C are threshold-distance interventions with decreasing threshold proximity from A to C. Solid curves report the mean over three random seeds, and shaded regions denote one standard deviation across seeds.

and leaves u unchanged outside the window. The exponents are

$$\gamma_A = \frac{\log(1/4)}{\log(1/2)}, \qquad \gamma_B = \frac{\log(5/12)}{\log(1/2)}, \qquad \gamma_C = \frac{\log(7/12)}{\log(1/2)}. \tag{36}$$

At the representative point $|u| = r/2$, A, B, and C map the absolute threshold distance to $0.25r$, $(5/12)r$, and $(7/12)r$, respectively. Thus A creates the strongest near-threshold profile, C the weakest, and REF keeps the original signed threshold distances. This construction changes threshold proximity while keeping the data, architecture, and checkpoint fixed.

We also report the mean firing rate

$$\rho(k) = \frac{1}{|\mathcal{B}_k| M} \sum_{\xi \in \mathcal{B}_k} \sum_{i=1}^{M} s_i^{\text{discrete}}(W_k; \xi), \tag{37}$$

as a control for silent or collapsed spiking dynamics.

Figure 6 shows that $\widehat{\mathcal{M}}_\delta(k)$ is nonzero on the REF trajectory. Under the controlled A/B/C interventions, the ordering of $\widehat{\mathcal{M}}_\delta(k)$ follows the ordering of $D(k)$: stronger near-threshold concentration produces a larger hard–smooth backward mismatch. The firing-rate curves remain nondegenerate, so the effect is not explained by silent dynamics. This supports the role of the threshold-tail statistic as an empirical proxy for the gradient-level bias term.

F.6 Cross-Model Threshold-Tail Diagnostic

This experiment corresponds to Table 1. For each dataset–model–seed run, training follows the corresponding recipe in Table 3. At the reported checkpoint, we compute $D(k)$ from hard-trajectory signed threshold distances using Eq. (25) and report it together with test accuracy.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Introduction state the scope as a unified convergence framework for SG and LocalZO direct SNN training, and the main result is stated as a reference-point non-asymptotic guarantee for the hard-spike objective. The assumptions, residual terms, and diagnostic interpretation are specified in Sections 4.3, 5.1, and 7.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the final paragraph of the Conclusion, including the reliance on smoothness, bounded variance, local computation-graph regularity, restricted PL-type conditions, and potentially conservative constants for long-horizon, deep recurrent, or heterogeneous SNNs.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Theoretical results are stated with explicit assumptions in Section 4.3 and in the statements of the main theorem, theorem, propositions, and lemmas. Complete proofs and derivations are provided in Appendices A.2–E, with a proof sketch in Section 5.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experiments are described in Section 7 and Appendix F. Appendix F.1 specifies the datasets, models, common protocol, surrogate and smoothing settings, LocalZO estimator, diagnostic computation, seeds, and task-level configurations used to reproduce the reported diagnostic results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Data and code are provided in https://github.com/ZL-IN/The-Convergence-of-Direct-Training-ofSNNs-with-Surrogate-Gradients

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix F.1, Table 2, and Table 3 specify the datasets, backbones, neuron settings, time horizons, losses, optimizers, schedulers, batch sizes, epochs, seeds, surrogate parameters, LocalZO sample sizes, and diagnostic-evaluation protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports seed-level variability for the diagnostic experiments supporting the empirical claims. Seed-wise results are provided in Table 1, and seed-aggregated curves with one-standard-deviation bands across three seeds are reported in Figure 6.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix F.1 comprehensively reports the compute resources utilized in all experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is a theoretical and diagnostic study of SNN optimization using public benchmark datasets and does not involve human subjects, private data, scraped sensitive data, or high-risk model release. The work does not require deviations from the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper does not explicitly discuss societal impacts, as its main contribution is a theoretical insight into SNN optimization. The work may indirectly support reliable and energy-efficient neuromorphic and edge AI systems, but we do not make specific societal impact claims.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release pretrained language models, image generators, scraped datasets, or other data or models with a high risk of misuse. Therefore no special release safeguards are applicable.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: This paper uses open-source datasets and models in F.1.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The submission does not introduce or release a new dataset, model checkpoint, or public code asset. The proposed contribution is a theoretical framework and associated diagnostic experiments rather than a new released asset.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The research does not involve crowdsourcing, participant studies, user studies, or other research with human subjects. No participant instructions, screenshots, or compensation details are applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research does not involve crowdsourcing or human subjects. Consequently, IRB or equivalent human-subjects review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the core method, theory, or experiments. If LLMs were used only for writing, editing, or formatting, the checklist indicates that a declaration is not required.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.