How Reliable Is Online Training Through Time? A Convergence and Diagnostic Framework for Spiking Neural Networks

Anonymous authors
Paper under double-blind review

Abstract

Spiking neural networks (SNNs) offer remarkable energy efficiency, yet their training typically relies on backpropagation through time (BPTT) with surrogate gradients, whose memory consumption grows linearly with the number of time steps. Online Training Through Time (OTTT) avoids full temporal backpropagation by maintaining presynaptic activity traces and performing same-time spatial backpropagation, enabling direct training with memory cost independent of the number of time steps. However, whether its biased updates can still provide reliable optimization guarantees for the original discrete-spike objective remains an open theoretical question. In this work, we establish a non-asymptotic convergence framework for fixed-within-sequence OTTTA in feedforward LIF SNNs. We introduce a randomized-smoothed spiking model as an analytical reference and use its full soft-BPTT gradient as the differentiable benchmark. The systematic OTTTA bias separates into two components: a hard–soft mismatch controlled by a threshold-tail functional measuring near-threshold membrane activity, and a temporal-decoupling bias quantified by the gradient discrepancy between soft OTTT and full soft-BPTT. Under explicit smoothness, stochastic-variance, trajectory-stability, and restricted-PL assumptions, we derive a finite-step bound directly on the hard-spike risk, relative to a restricted minimizer of the smooth reference risk, with explicit optimization, variance, hard–soft, temporal-decoupling, and risk-transfer terms. Across different static and event-based benchmarks, the two approximation discrepancies exhibit distinct task-dependent behavior, while mechanism-specific checkpoint diagnostics show pronounced seed-level associations with accuracy on several tasks. Together, these results connect a finite-step convergence guarantee with practical diagnostics for characterizing the approximation behavior of OTTT.

1 Introduction

Spiking neural networks (SNNs) offer an energy-efficient computing paradigm through sparse, event-driven communication and are well suited to neuromorphic hardware Maass (1997); Davies et al. (2018). Their training, however, remains challenging because spike generation is non-differentiable and membrane dynamics induce temporal dependencies Neftci et al. (2019); Bellec et al. (2020).

A standard solution combines surrogate gradients with backpropagation through time (BPTT) Shrestha & Orchard (2018); Wu et al. (2018); Zenke & Ganguli (2018); Lee et al. (2017). The forward pass retains hard-threshold spikes, while the step-function derivative is replaced by a smooth surrogate during backpropagation. Although effective, BPTT stores intermediate neuronal states and the unfolded temporal graph, so its memory cost grows linearly with the simulation horizon.

Online Training Through Time (OTTT) reduces this cost by maintaining presynaptic activity traces and performing same-time spatial backpropagation instead of full temporal gradient propagation Xiao et al. (2022). This forward-in-time construction yields memory independent of the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

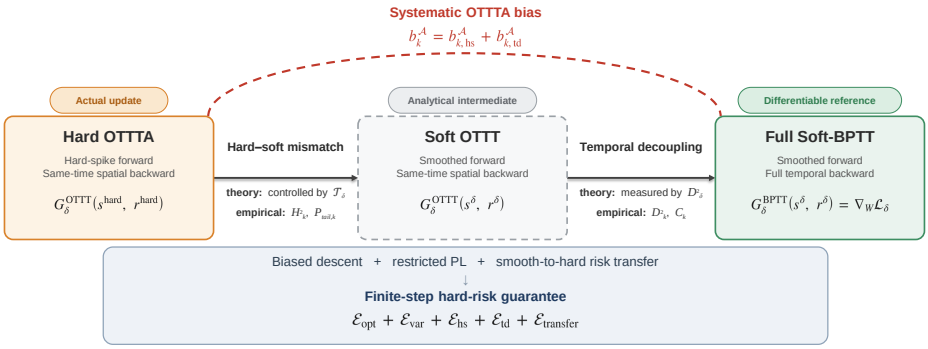

Figure 1: **Overview of the OTTTA bias decomposition.** Using full soft-BPTT as a differentiable reference, we decompose the systematic OTTTA bias into hard–soft and temporal-decoupling components, controlled by $\mathcal{T}_\delta$ and D_δ^2, respectively. These bias controls, combined with biased descent, restricted PL, and smooth-to-hard risk transfer, yield a finite-step guarantee on the hard-spike risk.

simulation horizon. Its accumulation variant, OTTTA, keeps parameters fixed within each sequence, whereas OTTTO updates them at every time step.

This efficiency introduces two distinct sources of bias. First, combining a hard forward trajectory with a smooth surrogate derivative creates a hard–soft mismatch. Second, temporal decoupling makes the OTTT gradient differ from full soft-BPTT even on the same smooth trajectory. Ignoring the reset-mediated temporal derivative is one neuron-level source of this error, but does not generally characterize the full network-level OTTT–BPTT discrepancy. Hence, the expected OTTT update is generally neither the gradient of the discrete-spike objective nor that of the smooth analytical reference.

Existing analyses do not fully resolve these biases. The original OTTT analysis provides a directional justification by relating its accumulated update to descent in spike-representation space under suitable conditions Xiao et al. (2022), but does not establish a non-asymptotic hard-risk guarantee for stochastic OTTTA iterates or explicitly control the hard–soft and temporal-decoupling errors. More broadly, analyses of surrogate-gradient estimators, proxy objectives, and local learning rules Bellec et al. (2020); Bengio et al. (2013); Frei & Gu (2021); Mukhoty et al. (2023) do not provide a network-level guarantee for these biased updates on the discrete objective. The central challenge is therefore to jointly control the hard–soft trajectory discrepancy and the temporal gradient information omitted by OTTT.

We establish a non-asymptotic convergence framework for fixed-parameter OTTTA in feedforward LIF SNNs. Rather than treating the hard-forward OTTT update as the exact gradient of an implicit objective, we introduce a randomized-smoothed spiking model solely as an analytical reference and use its full soft-BPTT gradient as the differentiable benchmark. The systematic OTTTA bias decomposes into a hard–soft mismatch controlled by a threshold-tail functional measuring how often hard-forward membrane margins approach the firing threshold, and a temporal-decoupling bias quantified by the soft OTTT–BPTT gradient discrepancy. Under explicit smoothness, stochastic-variance, trajectory-stability, and restricted-PL assumptions, we derive a finite-step bound directly on the hard-spike risk. The bound is relative to a restricted minimizer of the smooth reference risk and explicitly separates optimization, stochastic variance, hard–soft, temporal-decoupling, and risk-transfer contributions. Figure 1 summarizes the decomposition and its role in the hard-risk analysis.

Our main contributions are summarized as follows:

- We establish a non-asymptotic hard-risk guarantee for fixed-parameter OTTTA in feedforward LIF SNNs, relative to a restricted minimizer of the smooth analytical reference.
- We decompose systematic OTTTA bias into a threshold-tail-controlled hard–soft mismatch and a soft-OTTT–BPTT temporal-decoupling discrepancy.

- We develop mechanism-specific diagnostics and empirically characterize their training dynamics, cross-task magnitudes, seed-level accuracy associations, and smoothing-scale sensitivity.

2 Related Work

Online and memory-efficient training of SNNs. To avoid storing full spatiotemporal graphs, online SNN methods replace temporal backpropagation with local losses, eligibility traces, or forward optimization. DECOLLE uses layerwise errors, e-prop and OSTL factorize temporal credit into eligibility traces and learning signals, and FPTT minimizes a regularized instantaneous risk Kaiser et al. (2020); Bellec et al. (2020); Bohnstingl et al. (2022); Yin et al. (2023). OTTT uses presynaptic traces and same-time spatial backpropagation, while extensions such as NDOT, OPZO, and BrainTrace improve credit assignment or generalize to broader dynamics Xiao et al. (2022); Jiang et al. (2024); Xiao et al. (2026); Wang et al. (2026). These works design scalable online rules; we instead analyze when fixed-parameter OTTTA admits a controlled finite-step hard-risk guarantee and how its approximation error can be diagnosed.

OTTT and temporal-gradient approximation. OTTT approximates BPTT online by backpropagating each instantaneous loss through the same-time spatial graph while carrying temporal information forward via presynaptic traces, yielding memory independent of the number of time steps Xiao et al. (2022). Its fixed-within-sequence variant, OTTTA, accumulates these online directions over a sequence. Related methods adopt different temporal-gradient truncations: SLTT removes temporal backpropagation routes, arguing that their contribution is minor, to achieve time-step-independent memory Meng et al. (2023); Xue et al. analyze online-gradient deviations in completeness and timeliness and propose spatio-temporal online learning to improve fidelity Xue et al. (2024). Rather than introducing another truncation rule, we use full soft-BPTT as a differentiable reference, decompose OTTTA bias into hard–soft and temporal-decoupling components, and propagate both into a finite-step hard-risk guarantee.

Surrogate gradients and biased optimization. OTTT is also related to surrogate-gradient learning and biased stochastic optimization. Biased SGD on smooth nonconvex objectives converges to a bias-dependent neighborhood, with tighter guarantees under Polyak–Łojasiewicz conditions Ajalloeian & Stich (2020). For binary networks, straight-through and surrogate estimators have been analyzed through descent alignment and bias–variance trade-offs Bengio et al. (2013); Yin et al. (2023); Shekhovtsov & Yanush (2021), while recent work connects surrogate gradients to randomized smoothing and kernel dynamics Mukhoty et al. (2023); Gygax & Zenke (2025); Eilers et al. (2024). These results explain why approximate gradients can remain informative, but do not characterize how hard forward dynamics and temporal decoupling jointly bias OTTTA. We specialize biased-gradient theory by using a threshold-tail functional to bound the hard–soft discrepancy and the soft OTTT–BPTT gap to capture temporal decoupling, yielding a finite-step guarantee directly on the hard-spike risk.

3 Preliminaries

We define the hard-forward SNN dynamics and discrete-spike objective, followed by a smoothed analytical reference and its full BPTT gradient. We then review OTTT and its fixed-parameter accumulation variant, OTTTA, the main object of our convergence analysis. Unless stated otherwise, $\|\cdot\|$ denotes the Euclidean norm for vectors and the Frobenius norm for matrices and stacked parameter blocks, while $|\cdot|$ denotes scalar absolute value.

3.1 Spiking Neural Networks and the Discrete Objective

Consider an N-layer feedforward SNN evaluated for T time steps. Let $\xi = (x, y)$ denote a training sample and $W = (W^0, \ldots, W^{N-1})$ collect all trainable parameters, where W^l denotes the synaptic weights from layer l to layer $l+1$. For a leaky integrate-and-fire (LIF) neuron, let $u^l[t]$ and $s^l[t]$ denote the membrane potential and spike output of layer l at time step t, respectively. The hard-

forward dynamics are

$$u^{l+1}[t+1] = \lambda \left(u^{l+1}[t] - V_{\mathrm{th}} s^{l+1}[t]\right) + W^l s^l[t+1] + b^{l+1}, \tag{1}$$

$$s^{l+1}[t+1] = h\left(u^{l+1}[t+1] - V_{\mathrm{th}}\right), \tag{2}$$

where $\lambda \in (0,1)$ is the membrane leak factor, V_{th} is the firing threshold, and $h(a) = \mathbf{1}_{\{a>0\}}$ is the Heaviside step function. We define the signed membrane margin $r^l[t] := u^l[t] - V_{\mathrm{th}}$, so that $s^l[t] = h(r^l[t])$. For later analysis, we stack all spike variables and membrane margins over layers and time into $s^{\mathrm{hard}}(W;\xi) \in \{0,1\}^M$ and $r^{\mathrm{hard}}(W;\xi) \in \mathbb{R}^M$, where M is the total number of spike nodes in the unfolded spatiotemporal computation graph.

Following the instantaneous-loss formulation of OTTT Xiao et al. (2022), we assign each time step the normalized loss $\mathcal{L}[t](s, W;\xi) := \frac{1}{T}\ell\left(s^N[t], y\right), t = 1, \ldots, T$ and define the sequence-level loss as

$$\mathcal{L}_T(s, W;\xi) := \sum_{t=1}^{T} \mathcal{L}[t](s, W;\xi) = \frac{1}{T}\sum_{t=1}^{T} \ell\left(s^N[t], y\right). \tag{3}$$

The objective of interest throughout this paper is the population risk induced by the actual hard-spike dynamics,

$$F(W) := \mathbb{E}_\xi\left[\mathcal{L}_T\big(s^{\mathrm{hard}}(W;\xi), W;\xi\big)\right]. \tag{4}$$

Because h is discontinuous, F is generally non-differentiable and therefore does not directly admit the standard smooth optimization analysis used for gradient-based methods.

3.2 Smoothed Analytical Reference

To analyze the hard-forward update without changing the actual training dynamics, we introduce a smooth spiking model solely as an analytical reference. Throughout the analysis, the smoothed spike is defined by randomized smoothing:

$$h_\delta(a) = \mathbb{E}_{\epsilon\sim p_\epsilon}\left[h(a + \delta\epsilon)\right], \qquad g_\delta(a) := h'_\delta(a), \tag{5}$$

where $\delta > 0$ is the smoothing scale and ϵ has a symmetric continuous density p_ϵ. In this case, h_δ is differentiable with $g_\delta(a) = h'_\delta(a) = \frac{1}{\delta}p_\epsilon\big(-\frac{a}{\delta}\big)$. For the normalized LocalZO construction used in this work, p_ϵ is the induced smoothing density specified in Appendix A.1.

Replacing every occurrence of h in the hard SNN dynamics by h_δ defines the corresponding soft-forward trajectory $s^\delta(W;\xi) \in [0,1]^M$ and $r^\delta(W;\xi) \in \mathbb{R}^M$. Importantly, this soft trajectory is used only as a differentiable analytical reference; the actual OTTTA forward pass remains hard-spiking.

Using the same sequence-level loss, we define the smoothed population risk

$$F_\delta(W) := \mathbb{E}_\xi\left[\mathcal{L}_T\big(s^\delta(W;\xi), W;\xi\big)\right]. \tag{6}$$

Let $G_\delta^{\mathrm{BPTT}}(s, r, W;\xi)$ denote the parameter-gradient mapping obtained by applying the local derivative $g_\delta(r)$ at every spike node and backpropagating through the complete unfolded spatiotemporal graph. On the soft-forward trajectory, this mapping is the exact sample gradient of the smooth objective, i.e., $\nabla_W \mathcal{L}_T(s^\delta(W;\xi), W;\xi) = G_\delta^{\mathrm{BPTT}}(s^\delta(W;\xi), r^\delta(W;\xi), W;\xi)$, and hence

$$\nabla F_\delta(W) = \mathbb{E}_\xi\left[G_\delta^{\mathrm{BPTT}}\big(s^\delta(W;\xi), r^\delta(W;\xi), W;\xi\big)\right]. \tag{7}$$

Thus, full soft-BPTT provides the differentiable reference direction against which we measure the systematic bias of OTTTA.

The smooth spike construction can also be interpreted through the LocalZO theory of Mukhoty et al. Mukhoty et al. (2023), which establishes an exact correspondence between the expected local zeroth-order estimator and the derivative of a randomized smoothing of the Heaviside function; we provide the precise specialization used in our analysis in Appendix A.1.

3.3 OTTT Algorithm and Update Rules

OTTT avoids full temporal backpropagation by maintaining a presynaptic activity trace (Xiao et al., 2022), $\widehat{a}^{\,l}[t] = \lambda \widehat{a}^{\,l}[t-1] + s^l[t]$, and backpropagating each instantaneous loss only through the

same-time spatial graph. Let $g^{l+1}_{u,\delta}[t](s, r, W; \xi)$ denote the resulting error signal at $u^{l+1}[t]$, with g_δ used for spike derivatives. The OTTT direction for W^l is

$$\left[G^{\mathrm{OTTT}}_\delta(s, r, W; \xi)\right]_{W^l} = \sum_{t=1}^{T} g^{l+1}_{u,\delta}[t](s, r, W; \xi) \left(\widehat{a}^{\,l}[t](s)\right)^\top. \tag{8}$$

The temporal approximation is most transparent on the soft reference trajectory. With $M^{l+1}_\delta[t] = \mathrm{diag}(g_\delta(r^{l+1}_\delta[t]))$, full soft BPTT uses $J^{l+1}_\delta[t] = \lambda(I - V_{\mathrm{th}} M^{l+1}_\delta[t])$, whereas OTTT uses $J^{l+1}_{\mathrm{OTTT}}[t] = \lambda I$. Their difference, $-\lambda V_{\mathrm{th}} M^{l+1}_\delta[t]$, is one local source of temporal decoupling; D^2_δ below measures the full network-level discrepancy. See Appendix A.2.

We analyze OTTTA, the fixed-within-sequence gradient-accumulation variant of OTTT. At optimization iteration k, the parameters W_k remain fixed over the T simulation steps, and the instantaneous OTTT directions are accumulated as

$$\widehat{g}^A_k = \sum_{t=1}^{T} G_t(W_k; \zeta_k), \qquad W_{k+1} = W_k - \eta \widehat{g}^A_k, \tag{9}$$

where ζ_k collects mini-batch sampling and any local-estimator randomness. We denote its conditional mean by $\mu^A_k := \mathbb{E}[\widehat{g}^A_k \mid \mathcal{F}_k]$ and its systematic bias relative to the smooth reference gradient by $b^A_k := \mu^A_k - \nabla F_\delta(W_k)$.

4 Assumptions and Bias Measures

4.1 Assumptions

Assumption 1 (Smoothness of the analytical risk)**.** *The smoothed risk F_δ is differentiable and L_δ-smooth, i.e., $\|\nabla F_\delta(W) - \nabla F_\delta(W')\| \le L_\delta \|W - W'\|$ for all W, W'. Moreover, $F^\star_\delta := \inf_W F_\delta(W) > -\infty$, the sample gradients are integrable, and differentiation commutes with expectation so that $\nabla F_\delta(W) = \mathbb{E}_\xi[G^{\mathrm{BPTT}}_\delta(s^\delta(W;\xi), r^\delta(W;\xi), W; \xi)]$.*

Assumption 2 (Variance of the stochastic direction)**.** *At iteration k, the mini-batch is sampled independently from the data distribution conditional on $\mathcal{F}_k$. Any LocalZO perturbations are conditionally independent of the mini-batch and, when used nodewise, conditionally independent across spike nodes. Let $\widehat{g}_k$ denote the stochastic direction used at iteration k, let $\mathcal{F}_k$ denote the optimization history before that iteration, and define $\mu_k := \mathbb{E}[\widehat{g}_k \mid \mathcal{F}_k]$. There exist constants $\sigma^2_{\mathrm{mb}}, \sigma^2_{\mathrm{sg}} \ge 0$ such that $\mathbb{E}[\|\widehat{g}_k - \mu_k\|^2 \mid \mathcal{F}_k] \le \sigma^2_{\mathrm{mb}} + \frac{\sigma^2_{\mathrm{sg}}}{m}$. Here σ^2_{mb} captures mini-batch sampling variance, while σ^2_{sg}/m captures any additional randomness from averaging m local derivative estimators. For deterministic surrogate gradients, we take $m = 1$ and $\sigma^2_{\mathrm{sg}} = 0$.*

Assumption 3 (Trajectory stability of the OTTT mapping)**.** *There exists $L_{\mathrm{OTTT}} > 0$ such that, for any W, sample ξ, and two spatiotemporal trajectories (s, r) and (s', r') generated by the same input, $\left\|G^{\mathrm{OTTT}}_\delta(s, r, W; \xi) - G^{\mathrm{OTTT}}_\delta(s', r', W; \xi)\right\| \le L_{\mathrm{OTTT}}\big(\|s - s'\| + \|r - r'\|\big)$.*

Assumption 4 (Hard–soft trajectory stability)**.** *Define the nodewise smoothing discrepancy evaluated on the hard trajectory by $d_\delta(W;\xi) := h\big(r^{\mathrm{hard}}(W;\xi)\big) - h_\delta\big(r^{\mathrm{hard}}(W;\xi)\big)$, where h and h_δ are applied coordinatewise. There exists a finite amplification factor $\Lambda_\delta > 0$ such that $\|s^{\mathrm{hard}}(W;\xi) - s^\delta(W;\xi)\| + \|r^{\mathrm{hard}}(W;\xi) - r^\delta(W;\xi)\| \le \Lambda_\delta \|d_\delta(W;\xi)\|$ for all relevant W and almost every ξ.*

Assumption 4 is stated directly in the form used by our convergence analysis. In Appendix A.3, we provide two sufficient conditions under which it holds: a global contraction condition and a finite-horizon causal condition.

Assumption 5 (Lipschitz loss with respect to the trajectory)**.** *There exists $L_s > 0$ such that, for all W, samples ξ, and trajectories $s, s' \in [0,1]^M$, $|\mathcal{L}_T(s, W; \xi) - \mathcal{L}_T(s', W; \xi)| \le L_s \|s - s'\|$.*

Assumption 6 (Restricted PL condition)**.** *There exist a deterministic set $\mathcal{U} \subseteq \mathbb{R}^d$, a constant $\mu_{\mathcal{U}} > 0$, and a reference point $W^\star_{\mathcal{U}} \in \arg\min_{W \in \mathcal{U}} F_\delta(W)$ such that $W_k \in \mathcal{U}$ almost surely for all $0 \le k \le K$, and, for every $W \in \mathcal{U}$, $F_\delta(W) - F_\delta(W^\star_{\mathcal{U}}) \le \frac{1}{2\mu_{\mathcal{U}}} \|\nabla F_\delta(W)\|^2$.*

4.2 Two Bias Measures

Threshold-tail functional. Let $\epsilon \sim p_\epsilon$ be the smoothing variable associated with h_δ and define its two-sided tail function $\Psi(a) := \mathbb{P}(|\epsilon| \geq a), a \geq 0$. For a parameter vector W, define

$$\mathcal{T}_\delta(W) := \mathbb{E}_\xi \left[\sum_{i=1}^{M} \Psi\left(\frac{|r_i^{\mathrm{hard}}(W;\xi)|}{\delta} \right) \right]. \tag{10}$$

The quantity $\mathcal{T}_\delta(W)$ is small when few hard-forward membrane margins lie near the firing threshold relative to the smoothing scale δ. It controls the hard–soft trajectory mismatch and, consequently, the corresponding hard–soft component of the OTTTA bias.

temporal-decoupling discrepancy. To isolate the approximation caused only by temporal-gradient decoupling, we evaluate OTTT and full BPTT on the same soft trajectory and define

$$D_\delta^2(W) := \mathbb{E}_\xi \left[\left\| G_\delta^{\mathrm{OTTT}}\big(s^\delta(W;\xi), r^\delta(W;\xi), W;\xi\big) - G_\delta^{\mathrm{BPTT}}\big(s^\delta(W;\xi), r^\delta(W;\xi), W;\xi\big) \right\|^2 \right]. \tag{11}$$

Because both mappings in Eq. equation 11 share the same soft-forward trajectory and local derivative g_δ, $D_\delta^2(W)$ excludes the hard–soft mismatch and measures the full network-level gradient error introduced by replacing full temporal backpropagation with the trace-based OTTT direction.

For the first K optimization iterations, we use the averaged quantities

$$\overline{\mathcal{T}}_{\delta,K} := \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\big[\mathcal{T}_\delta(W_k)\big], \qquad \overline{D}_{\delta,K}^2 := \frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}\big[D_\delta^2(W_k)\big]. \tag{12}$$

These two measures are complementary: $\overline{\mathcal{T}}_{\delta,K}$ quantifies the forward hard–soft mismatch, whereas $\overline{D}_{\delta,K}^2$ quantifies the backward temporal-decoupling error. Both will appear explicitly in the non-asymptotic discrete-objective guarantee in Section 5.

5 Main Convergence Result

Let $W_{\mathcal{U}}^\star$ be the restricted minimizer of F_δ defined in Assumption 6, and recall the averaged approximation measures $\overline{\mathcal{T}}_{\delta,K}$ and $\overline{D}_{\delta,K}^2$ from Eq. equation 12.

Theorem 1 (Non-asymptotic hard-risk guarantee for OTTTA)**.** *Consider the feedforward LIF SNN in Section 3.1 trained by the fixed-parameter OTTTA update in Eq. equation 9. Suppose Assumptions 1–6 hold, and the local backward derivative is either the deterministic g_δ or the normalized nodewise LocalZO estimator in Appendix A.1, with conditionally independent nodewise perturbations. If $\eta \leq 1/(4L_\delta)$ and $R \sim \mathrm{Unif}\{0,\dots,K-1\}$ is independent of the optimization randomness, then*

$$\begin{aligned} \mathbb{E}[F(W_R)] - F(W_{\mathcal{U}}^\star) \leq{}& \underbrace{\frac{2\big(F_\delta(W_0) - F_\delta(W_{\mathcal{U}}^\star)\big)}{\mu_{\mathcal{U}} \eta K}}_{\mathcal{E}_{\mathrm{opt}}} + \underbrace{\frac{L_\delta \eta}{\mu_{\mathcal{U}}} \left(\sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m} \right)}_{\mathcal{E}_{\mathrm{var}}} + \underbrace{\frac{3 L_{\mathrm{OTTT}}^2 \Lambda_\delta^2}{\mu_{\mathcal{U}}} \overline{\mathcal{T}}_{\delta,K}}_{\mathcal{E}_{\mathrm{hs}}} \\ &+ \underbrace{\frac{3}{\mu_{\mathcal{U}}} \overline{D}_{\delta,K}^2}_{\mathcal{E}_{\mathrm{td}}} + \underbrace{L_s \Lambda_\delta \left(\sqrt{\overline{\mathcal{T}}_{\delta,K}} + \sqrt{\mathcal{T}_\delta(W_{\mathcal{U}}^\star)} \right)}_{\mathcal{E}_{\mathrm{transfer}}}. \end{aligned} \tag{13}$$

The first two terms are the standard finite-step optimization and stochastic variance errors. The OTTTA-specific residual separates into a hard–soft term $\mathcal{E}_{\mathrm{hs}}$, controlled by near-threshold activity, and a temporal-decoupling term $\mathcal{E}_{\mathrm{td}}$, measuring the discrepancy between soft OTTT and full soft BPTT on the same trajectory. The final term transfers the guarantee from the smooth reference risk F_δ to the hard-spike risk F. The comparison point $W_{\mathcal{U}}^\star$ minimizes F_δ only over $\mathcal{U}$ and is not assumed to globally minimize the discontinuous objective F.

Section 6 derives the two structural bias controls and shows how they enter the bound. Complete proofs are given in Appendix A.4.

6 Convergence Analysis: Controlling the Two Bias Sources

We isolate the two OTTTA-specific terms in Theorem 1 by comparing the conditional mean hard-forward OTTTA update with ∇F_δ and decomposing their discrepancy into hard–soft and temporal-decoupling components. Bounding these components within a standard biased-descent argument yields the result. Full derivations are in Appendix A.4.

6.1 Bias Decomposition

For deterministic surrogate-gradient OTTTA, or normalized nodewise LocalZO-OTTT with conditionally independent perturbations, conditional averaging removes local estimator randomness and recovers the deterministic OTTT mapping on the hard-forward trajectory:

$$\mu_k^A = \mathbb{E}_\xi\left[G_\delta^{\mathrm{OTTT}}\left(s^{\mathrm{hard}}(W_k;\xi), r^{\mathrm{hard}}(W_k;\xi), W_k;\xi\right) \,\middle|\, \mathcal{F}_k\right]. \tag{14}$$

Thus, the systematic error relative to the smooth reference gradient is $b_k^A := \mu_k^A - \nabla F_\delta(W_k)$.

Inserting the soft-trajectory OTTT direction as an intermediate quantity and using Eq. equation 7, we define

$$\begin{aligned} b_{k,\mathrm{hs}}^A &:= \mathbb{E}_\xi\left[G_\delta^{\mathrm{OTTT}}(s^{\mathrm{hard}}, r^{\mathrm{hard}}, W_k;\xi) - G_\delta^{\mathrm{OTTT}}(s^\delta, r^\delta, W_k;\xi) \,\middle|\, \mathcal{F}_k\right], \\ b_{k,\mathrm{td}}^A &:= \mathbb{E}_\xi\left[G_\delta^{\mathrm{OTTT}}(s^\delta, r^\delta, W_k;\xi) - G_\delta^{\mathrm{BPTT}}(s^\delta, r^\delta, W_k;\xi) \,\middle|\, \mathcal{F}_k\right]. \end{aligned} \tag{15}$$

Here the trajectory arguments depend on $(W_k;\xi)$ and are suppressed for brevity. Hence $b_k^A = b_{k,\mathrm{hs}}^A + b_{k,\mathrm{td}}^A$. The hard–soft term changes only the forward trajectory while keeping the OTTT mapping fixed, whereas the temporal-decoupling term compares OTTT with full BPTT on the same soft trajectory. This isolates the two approximation mechanisms before the optimization analysis.

6.2 Controlling the Two Bias Components

The hard–soft term is controlled through the threshold-tail functional $\mathcal{T}_\delta$, while the temporal term is measured by the soft OTTT–BPTT discrepancy D_δ^2.

Proposition 1 (Control of the two OTTTA bias components)**.** *Under Assumptions 3 and 4,*

$$\mathbb{E}\left[\left\|b_{k,\mathrm{hs}}^A\right\|^2\right] \le L_{\mathrm{OTTT}}^2 \Lambda_\delta^2 \mathbb{E}[\mathcal{T}_\delta(W_k)], \tag{16}$$

$$\mathbb{E}\left[\left\|b_{k,\mathrm{td}}^A\right\|^2\right] \le \mathbb{E}\left[D_\delta^2(W_k)\right]. \tag{17}$$

The first bound combines Lipschitz stability of the OTTT mapping with the hard–soft trajectory bound; the second follows from the definition of D_δ^2 and Jensen's inequality. Using $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$ and averaging over the first K iterations,

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\|b_k^A\|^2\right] \le 2L_{\mathrm{OTTT}}^2 \Lambda_\delta^2 \overline{\mathcal{T}}_{\delta,K} + 2\overline{D}_{\delta,K}^{\,2}. \tag{18}$$

Equation equation 18 is the key structural step: the systematic OTTTA bias is controlled separately by the two approximation mechanisms. The reset-mediated temporal derivative omitted by the OTTT trace is one local source of temporal decoupling, while D_δ^2 captures the full network-level discrepancy; see Appendix A.2.

6.3 From Smooth Descent to the Hard-Risk Guarantee

Under Assumptions 1 and 2, standard L_δ-smooth biased descent with $\eta \le 1/(4L_\delta)$ gives

$$\begin{aligned} \frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\|\nabla F_\delta(W_k)\|^2\right] \le{}& \frac{4\left(F_\delta(W_0) - F_\delta(W_{\mathcal{U}}^\star)\right)}{\eta K} + 2L_\delta \eta \left(\sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m}\right) \\ &+ \frac{3}{K}\sum_{k=0}^{K-1} \mathbb{E}\left[\|b_k^A\|^2\right]. \end{aligned} \tag{19}$$

The first two terms are the usual finite-step optimization and stochastic variance contributions, while the last contains the OTTTA approximation bias. Substituting Eq. equation 18 yields the hard–soft and temporal-decoupling residuals in Theorem 1.

Assumption 6 then converts the averaged gradient bound into a smooth-risk bound at the uniformly sampled iterate W_R. Finally, Assumptions 4 and 5 transfer the smooth reference risk back to the hard-spike objective:

$$|F(W) - F_\delta(W)| \le L_s \Lambda_\delta \sqrt{\mathcal{T}_\delta(W)}. \tag{20}$$

Applying this bound at both W_R and $W_{\mathcal{U}}^\star$ completes the hard-risk guarantee in Eq. equation 13.

7 Experiments

We evaluate three implications of the theoretical decomposition: discrepancy dynamics, relative magnitudes across tasks, and diagnostic–accuracy associations. We further examine the smoothing-scale trade-off induced by δ. We use H_k^2 and D_k^2 for the empirical hard–soft and temporal discrepancies, respectively, and $P_{\mathrm{tail},k}$ as the theory-motivated hard–soft diagnostic. Full measurement details are provided in Appendix B.

7.1 Training Dynamics of the Two Approximation Discrepancies

Figure 2 shows the BPTT-normalized discrepancies from epoch 100 onward, averaged over five seeds with one-standard-deviation shading.

On CIFAR-10 and CIFAR-100, the hard–soft discrepancy increases during late training, whereas the temporal discrepancy remains stable or slightly decreases. On CIFAR10-DVS and DVS Gesture, the temporal discrepancy remains larger and comparatively stable. These contrasting trajectories indicate that training can amplify one source of approximation error without similarly affecting the other, supporting their treatment as distinct mechanisms rather than scaled versions of a single error.

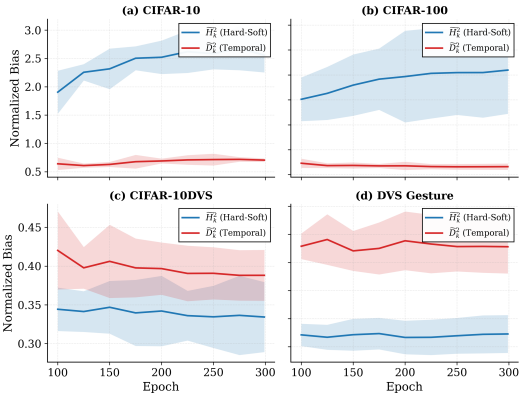

Figure 2: **Training dynamics of the two approximation discrepancies.**

7.2 Cross-Task Discrepancy Magnitudes

At the final checkpoint, we compare the discrepancy ratio D_k^2/H_k^2 with soft-OTTT/soft-BPTT directional agreement C_k, averaged over five seeds.

The ratio D_k^2/H_k^2 is below 0.3 on CIFAR-10 and CIFAR-100, but rises to 1.200 on CIFAR10-DVS and 3.948 on DVS Gesture, showing that the relative importance of the two mechanisms varies substantially across the tested configurations. Directional agreement provides a complementary view: CIFAR10-DVS retains $C_k = 0.9270$ despite a temporal discrepancy comparable to the hard–soft one, whereas DVS Gesture combines a much larger temporal discrepancy with a marked drop to $C_k = 0.5542$. Thus, temporal-error magnitude and its effect on gradient direction need not vary proportionally.

Table 1: **Final-checkpoint discrepancy diagnostics**.

Dataset	H_k^2	D_k^2	D_k^2/H_k^2	C_k
CIFAR-10	75.80	20.77	0.274	0.9610
CIFAR-100	336.77	84.06	0.250	0.9514
CIFAR10-DVS	299.17	358.95	1.200	0.9270
DVS Gesture	2022.82	7985.95	3.948	0.5542

7.3 Diagnostic–Accuracy Association

Following the dominant mechanism identified in Section 7.2, we test seed-level associations between test accuracy and $P_{\mathrm{tail},k}$ on CIFAR-10 and CIFAR-100 or $\log D_k^2$ on CIFAR10-DVS and DVS Gesture, using final-checkpoint measurements from five seeds.

Associations are negative on CIFAR-10 ($r = -0.82$), CIFAR10-DVS ($r = -0.97$), and DVS Gesture ($r = -0.94$), but positive on CIFAR-100 ($r = 0.45$). A joint within-dataset permutation test gives $p_{\text{perm}} \approx 0.002$, providing evidence against a joint zero-association null across the four tasks. The CIFAR-100 result does not contradict the decomposition: $P_{\text{tail},k}$ probes the hard–soft mechanism rather than all factors governing final accuracy, and its association on CIFAR-100 is less stable under the sensitivity analysis. Thus, the diagnostics are informative about mechanism-specific approximation behavior and can covary strongly with performance, but should not be interpreted as universal accuracy predictors. Detailed statistics are provided in Appendix B.2.

Table 2: **Final-checkpoint diagnostic–accuracy association.**

Dataset	Diagnostic	r
CIFAR-10	$P_{\text{tail},k}$	$\mathbf{-0.82}$
CIFAR-100	$P_{\text{tail},k}$	$+0.45$
CIFAR10-DVS	$\log D_k^2$	$\mathbf{-0.97}$
DVS Gesture	$\log D_k^2$	$\mathbf{-0.94}$

7.4 Smoothing-Scale Trade-off

To examine the δ-dependence exposed by the theory, we vary $\delta \in \{1/8, 1/6, 1/4, 1/3, 1/2\}$ on CIFAR-10 while keeping the remaining training configuration fixed.

Figure 3 shows that accuracy varies non-monotonically with δ and peaks at $\delta = 1/3$. In contrast, $P_{\text{tail},k}$ increases with the smoothing width, while D_k^2/H_k^2 reaches its minimum near the default $\delta = 1/4$ and increases sharply for broader smoothing. Thus, δ affects both predictive performance and the relative balance between the two approximation mechanisms, with no single diagnostic varying monotonically with accuracy. Details are provided in Appendix B.3.

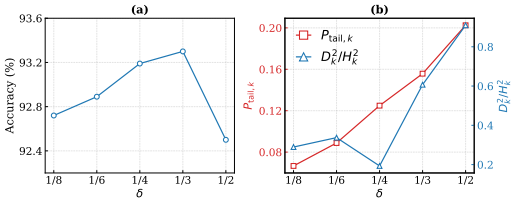

Figure 3: **Smoothing-scale δ trade-off.**

8 Discussion and Limitation

Our results suggest that OTTTA approximation quality is inherently multi-dimensional. Separating hard–soft mismatch from temporal decoupling reveals that the dominant source of error, its magnitude, and its effect on gradient direction need not coincide. This distinction also clarifies the role of the empirical diagnostics: they probe specific approximation mechanisms rather than serve as sufficient predictors of final generalization, which also depends on optimization dynamics, stochastic variation, and risk transfer. This perspective is consistent with the multi-term structure of our theoretical guarantee and suggests that improving OTTTA need not require uniformly reducing every source of error. Instead, corrections may be more effective when targeted at the dominant mechanism—for example, reducing hard–soft mismatch when threshold sensitivity is prominent or restoring temporal credit when decoupling is the main source of gradient distortion. More broadly, the decomposition provides a framework for diagnosing why an online approximation deviates from full BPTT, rather than only measuring the size of that deviation.

Several limitations remain. The theory is restricted to feedforward LIF SNNs and fixed-parameter OTTTA under the stated assumptions. Extending the framework to broader neuronal dynamics, architectures, and genuinely online parameter updates would test whether the same decomposition remains informative beyond the present setting.

9 Conclusion

We establish a non-asymptotic convergence framework for fixed-parameter OTTTA in feedforward LIF SNNs with a guarantee directly on hard-spike risk. Using a smoothed analytical reference, we decompose systematic OTTTA bias into a threshold-tail-controlled hard–soft mismatch and a temporal-decoupling discrepancy between soft OTTT and full soft-BPTT. Across tasks, the two mechanisms exhibit distinct dynamics and relative magnitudes; mechanism-specific diagnostics covary with accuracy on several datasets, and smoothing-scale sensitivity is non-monotone in both performance and approximation behavior. Together, these results characterize when and how OTTTA departs from full temporal-gradient optimization.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript structuring, code editing, experimental design feedback, and interpretation of experimental results. They were also used to provide feedback on mathematical arguments and presentation. All AI-assisted suggestions, code, derivations, and interpretations were independently reviewed and verified by the authors. The authors take full responsibility for the final mathematical claims, proofs, experimental methodology, results, citations, code, and manuscript content.

ETHICS STATEMENT

All experiments use publicly available vision and event-based recognition benchmarks. This work does not involve human-subject research, sensitive data, or deployment in high-stakes decision settings. We do not identify ethical risks specific to this theoretical and diagnostic analysis beyond the general computational and environmental costs of model training.

REPRODUCIBILITY STATEMENT

The model assumptions and convergence statements are given in the main text, with complete proofs in the appendix. Appendix B specifies the experimental diagnostics, checkpoint measurements, seed protocols, smoothing-scale study, and analysis procedures underlying the reported results.

REFERENCES

Ahmad Ajalloeian and Sebastian U Stich. On the convergence of sgd with biased gradients. *arXiv preprint arXiv:2008.00051*, 2020.

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.

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

Thomas Bohnstingl, Stanisław Woźniak, Angeliki Pantazi, and Evangelos Eleftheriou. Online spatio-temporal learning in deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8894–8908, 2022.

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.

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.

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.

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

Haiyan Jiang, Giulia De Masi, Huan Xiong, and Bin Gu. Ndot: neuronal dynamics-based online training for spiking neural networks. In *Forty-first International Conference on Machine Learning*, 2024.

Jacques Kaiser, Hesham Mostafa, and Emre Neftci. Synaptic plasticity dynamics for deep continuous local learning (decolle). *Frontiers in Neuroscience*, 14:424, 2020.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

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

Qingyan Meng, Mingqing Xiao, Shen Yan, Yisen Wang, Zhouchen Lin, and Zhi-Quan Luo. Towards memory-and time-efficient backpropagation for training spiking neural networks. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6143–6153. IEEE, 2023.

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.

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

Alexander Shekhovtsov and Viktor Yanush. Reintroducing straight-through estimators as principled methods for stochastic binary networks. In *DAGM german conference on pattern recognition*, pp. 111–126. Springer, 2021.

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

Chaoming Wang, Xingsi Dong, Zilong Ji, Mingqing Xiao, Jiedong Jiang, Xiao Liu, Yuxiang Huan, and Si Wu. Model-agnostic linear-memory online learning in spiking neural networks. *Nature Communications*, 17(1):1745, 2026.

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.

Mingqing Xiao, Qingyan Meng, Zongpeng Zhang, Di He, and Zhouchen Lin. Online training through time for spiking neural networks. *Advances in neural information processing systems*, 35:20717–20730, 2022.

Mingqing Xiao, Qingyan Meng, Zongpeng Zhang, Di He, Dongsheng Li, and Zhouchen Lin. Online pseudo-zeroth-order training of neuromorphic spiking neural networks. In *International Conference on Learning Representations*, volume 2026, pp. 135364–135391, 2026.

Lang Xue, HanWen Liu, Jing Wang, and Hong Qu. Reasonable gradients for online training algorithms in spiking neural networks. In *ECAI 2024: 27th European Conference on Artificial Intelligence, 19–24 October 2024, Santiago de Compostela, Spain–Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, pp. 2709–2716. SAGE Publications Pvt. Ltd 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2024.

Bojian Yin, Federico Corradi, and Sander M Bohté. Accurate online training of dynamical spiking neural networks through forward propagation through time. *Nature Machine Intelligence*, 5(5): 518–527, 2023.

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

A Additional Theory and Proofs

A.1 Local Zeroth-Order Smoothing as a Theoretical Foundation

We ground our smooth reference F_δ in the Local Zeroth-Order (LocalZO) theory of Mukhoty et al. Mukhoty et al. (2023), which connects a local two-point estimator on the Heaviside function to the derivative of a randomly smoothed spike. This section states the two central theorems from that work in the notation of our paper; the correspondence itself is not our contribution.

LocalZO estimator. Let $h(a) = \mathbf{1}_{\{a>0\}}$ be the Heaviside spike, $\delta > 0$ a perturbation scale, and $z \sim p_z$ with an even density p_z. The generalized local zeroth-order estimator is

$$G_{2,\alpha}(a; z, \delta) = \begin{cases} 0, & |a| > |z|\delta, \\ \dfrac{|z|^\alpha}{2\delta}, & |a| \le |z|\delta, \end{cases} \tag{21}$$

for a nonzero integer α. For $\alpha = 1$, this reduces to the two-point form $G_{2,1}(a; z, \delta) = \frac{h(a+\delta z)-h(a-\delta z)}{2\delta}\, z$ almost everywhere. Using the symmetry of p_z, its expectation is $\mathbb{E}_{z\sim p_z}[G_{2,\alpha}(a; z, \delta)] = \frac{1}{\delta}\int_{|a|/\delta}^{\infty} t^\alpha p_z(t)\, dt$.

Theorem 2 (Mukhoty et al. (2023))**.** *Assume $\int_0^\infty t^\alpha p_z(t)\, dt < \infty$ and $\int_0^\infty t^{\alpha+1} p_z(t)\, dt < \infty$. Define the normalization constant $c_\alpha = 2\int_0^\infty t^{\alpha+1} p_z(t)\, dt$ and the smoothed density $\widetilde{p}_\alpha(r) = \frac{1}{c_\alpha}\int_{|r|}^\infty t^\alpha p_z(t)\, dt$. Then, for almost every a,*

$$\mathbb{E}_{z\sim p_z}[G_{2,\alpha}(a; z, \delta)] = \frac{d}{da}\, \mathbb{E}_{\epsilon\sim\widetilde{p}_\alpha}[c_\alpha\, h(a + \delta\epsilon)]\,. \tag{22}$$

Thus the expected LocalZO estimator is exactly the derivative of a randomized smoothing of the hard spike, up to the constant c_α.

Normalized specialization used in our work. We specialize to $\alpha = 1$. Because p_z is even, $c_1 = 2\int_0^\infty t^2 p_z(t)\, dt = \mathbb{E}_z[z^2] =: c$. We use the normalized estimator $\overline{G}_2(a; z, \delta) = \frac{h(a+\delta z)-h(a-\delta z)}{2\delta}\, \frac{z}{c}$, and define the induced smoothing density $p_\epsilon(r) = \frac{1}{c}\int_{|r|}^\infty t\, p_z(t)\, dt$, which integrates to one. Applying Theorem 2 with $\alpha = 1$ and dividing by c yields the key local correspondence used throughout the paper:

$$\mathbb{E}_{z\sim p_z}\left[\overline{G}_2(a; z, \delta)\right] = g_\delta(a) = h'_\delta(a), \qquad h_\delta(a) = \mathbb{E}_{\epsilon\sim p_\epsilon}[h(a + \delta\epsilon)] \tag{23}$$

almost everywhere. Deterministic surrogate gradients using g_δ and normalized LocalZO therefore share the same expected local derivative; any extra randomness is captured by the estimator variance.

Theorem 3 (Mukhoty et al. (2023))**.** *Let g be a surrogate function such that $c_g = -2\delta^2\int_0^\infty \frac{1}{z^\alpha}\, g'(z\delta)\, dz < \infty$. Define the sampling density $p_z(z) = -\frac{\delta^2}{c_g\, z^\alpha}\, g'(z\delta)$. Then, almost everywhere,*

$$c_g\, \mathbb{E}_{z\sim p_z}[G_{2,\alpha}(a; z, \delta)] = g(a). \tag{24}$$

This converse result shows that a broad class of surrogate derivatives can be recovered by suitable LocalZO distributions, justifying g_δ as a principled smooth derivative rather than an ad hoc replacement.

From local smoothing to the network-level analytical reference. The correspondence above is local: it identifies the expected derivative estimator at an individual hard-threshold node with the derivative $g_\delta = h'_\delta$ of a randomly smoothed spike. In the actual OTTTA algorithm, the forward trajectory remains hard-spike; the smoothing does not replace the network dynamics used for training. Motivated by this local correspondence, we separately construct a differentiable network-level reference by replacing each hard spike h with h_δ, yielding the soft-forward trajectory $s^\delta(W; \xi), r^\delta(W; \xi)$ and the smooth population risk $F_\delta(W) = \mathbb{E}_\xi\left[\mathcal{L}_T(s^\delta(W; \xi), W; \xi)\right]$. Full BPTT on this soft trajectory defines the reference gradient $\nabla F_\delta(W)$. Crucially, the local expectation identity does not imply that the expected hard-forward LocalZO network coincides with this soft-forward reference. Their discrepancy is precisely what our subsequent analysis resolves into the hard–soft mismatch and the temporal-decoupling bias.

A.2 A Local Reset-Mediated Component of Temporal Decoupling

The temporal-decoupling discrepancy D_δ^2 is defined at the network-gradient level as the difference between soft OTTT and full soft-BPTT. We do not attempt to reduce this complete discrepancy to a single temporal path. Instead, this subsection identifies one concrete layer-local source of the

approximation introduced by OTTT: the temporal derivative through the reset term of the LIF dynamics.

The derivation is performed on the soft reference trajectory so that the relevant derivatives are well defined. It is used only to interpret one component of temporal decoupling and does not alter the hard-spike forward dynamics used by OTTTA.

Layer-local state sensitivity. Rewrite the soft LIF dynamics of layer $l+1$ as

$$u_\delta^{l+1}[t+1] = \lambda\left(u_\delta^{l+1}[t] - V_{\mathrm{th}} s_\delta^{l+1}[t]\right) + W^l s_\delta^l[t+1] + b^{l+1}, \qquad s_\delta^{l+1}[t] = h_\delta(r_\delta^{l+1}[t]). \tag{25}$$

For a synapse W_{ij}^l, define the layer-local state sensitivity

$$e_{ij}^l[t] := \frac{\partial u_{\delta,i}^{l+1}[t]}{\partial W_{ij}^l}. \tag{26}$$

Because W^l connects layer l to layer $l+1$, the presynaptic activity $s_{\delta,j}^l$ does not depend on this weight block. Differentiating Eq. equation 25 therefore gives

$$e_{ij}^l[t+1] = \lambda\left(1 - V_{\mathrm{th}} g_\delta(r_{\delta,i}^{l+1}[t])\right) e_{ij}^l[t] + s_{\delta,j}^l[t+1]. \tag{27}$$

By contrast, the presynaptic trace used by OTTT follows

$$\widehat{a}_j^l[t+1] = \lambda \widehat{a}_j^l[t] + s_{\delta,j}^l[t+1]. \tag{28}$$

Comparing Eqs. equation 27 and equation 28 shows that, at this layer-local state sensitivity, the OTTT trace omits the term

$$-\lambda V_{\mathrm{th}} g_\delta(r_{\delta,i}^{l+1}[t]) e_{ij}^l[t]. \tag{29}$$

This term arises from differentiating the reset contribution $-V_{\mathrm{th}} s_\delta^{l+1}[t]$. Thus, OTTT retains the reset in the forward dynamics but does not propagate this particular reset-mediated derivative through its presynaptic trace.

One-step Jacobian interpretation. Let

$$M_\delta^{l+1}[t] = \mathrm{diag}\left(g_\delta(r_\delta^{l+1}[t])\right). \tag{30}$$

The corresponding one-step temporal Jacobian of the soft LIF state is

$$J_\delta^{l+1}[t] = \lambda\left(I - V_{\mathrm{th}} M_\delta^{l+1}[t]\right), \tag{31}$$

whereas the leak recursion implicit in the OTTT trace uses λI. Their one-step difference is therefore

$$J_\delta^{l+1}[t] - \lambda I = -\lambda V_{\mathrm{th}} M_\delta^{l+1}[t]. \tag{32}$$

Equation equation 32 identifies a concrete reset-mediated component of temporal decoupling, but it is deliberately a *local* statement. In a multilayer SNN, full BPTT propagates parameter perturbations through the complete unfolded spatiotemporal computation graph, including their influence on downstream layers and future states. The same-time spatial backward pass and presynaptic traces used by OTTT do not reproduce all such paths. Accordingly, we do not express the full network-gradient discrepancy as a product of the layer-local Jacobians above.

For this reason, the convergence analysis retains

$$D_\delta^2(W) = \mathbb{E}_\xi\left[\left\|G_\delta^{\mathrm{OTTT}}(s^\delta, r^\delta, W; \xi) - G_\delta^{\mathrm{BPTT}}(s^\delta, r^\delta, W; \xi)\right\|^2\right] \tag{33}$$

as the network-level measure of temporal decoupling. The local calculation above explains one mechanism contributing to this quantity, rather than an exact decomposition of D_δ^2 itself. This distinction is consistent with the original OTTT construction, which obtains forward-in-time learning by tracking presynaptic activities and using instantaneous gradients instead of full backward propagation through time Xiao et al. (2022).

A.3 Sufficient Conditions for Hard–Soft Trajectory Stability

Assumption 4 is stated in the main text directly in the network-level form required by our convergence analysis. We now give two concrete sufficient conditions under which it holds. The first imposes a global contraction property, whereas the second only requires causal stability over a finite unfolded horizon.

Throughout this subsection, fix a parameter vector W and a sample ξ, and abbreviate

$$s^{\mathrm{h}} := s^{\mathrm{hard}}(W;\xi), \qquad s^{\delta} := s^{\delta}(W;\xi),$$
$$r^{\mathrm{h}} := r^{\mathrm{hard}}(W;\xi), \qquad r^{\delta} := r^{\delta}(W;\xi).$$

Recall the nodewise smoothing discrepancy

$$d_\delta := h(r^{\mathrm{h}}) - h_\delta(r^{\mathrm{h}}), \tag{34}$$

where h and h_δ are applied coordinatewise. We assume throughout this subsection that h_δ is L_h-Lipschitz:

$$\|h_\delta(a) - h_\delta(b)\| \le L_h\|a-b\| \qquad \text{for all relevant } a, b. \tag{35}$$

Proposition 2 (Sufficient conditions for Assumption 4)**.** *Suppose Eq. equation 35 holds. Then Assumption 4 follows if either of the following conditions is satisfied.*

(a) Global contraction condition. *There exists $L_r > 0$ such that, uniformly over all relevant W and almost every ξ,*

$$\|r^{\mathrm{h}} - r^{\delta}\| \le L_r\|s^{\mathrm{h}} - s^{\delta}\|, \qquad L_h L_r < 1. \tag{36}$$

Then

$$\|s^{\mathrm{h}} - s^{\delta}\| + \|r^{\mathrm{h}} - r^{\delta}\| \le \Lambda_\delta^{\mathrm{glob}}\|d_\delta\|, \tag{37}$$

where

$$\Lambda_\delta^{\mathrm{glob}} := \frac{1+L_r}{1-L_hL_r}. \tag{38}$$

(b) Finite-horizon causal condition. *Partition the unfolded spatiotemporal computation graph into H causal blocks. If multiple feedforward dependencies occur within the same physical time step, they may be refined into micro-blocks and ordered causally. Let (s_t, r_t) denote the spike and signed-margin variables in block t, and assume that the hard and soft systems share the same initialization,*

$$s_0^{\mathrm{h}} = s_0^{\delta}. \tag{39}$$

Suppose there exists $L_c > 0$ such that, for every $t = 1, \dots, H$,

$$\|r_t^{\mathrm{h}} - r_t^{\delta}\| \le L_c\|s_{t-1}^{\mathrm{h}} - s_{t-1}^{\delta}\|. \tag{40}$$

Then

$$\|s^{\mathrm{h}} - s^{\delta}\| + \|r^{\mathrm{h}} - r^{\delta}\| \le \Lambda_\delta^{\mathrm{fh}}\|d_\delta\|, \tag{41}$$

where

$$\Lambda_\delta^{\mathrm{fh}} := (1+L_c)\sum_{q=0}^{H-1}(L_hL_c)^q. \tag{42}$$

Hence Assumption 4 holds with $\Lambda_\delta = \Lambda_\delta^{\mathrm{glob}}$ under condition (a), or with $\Lambda_\delta = \Lambda_\delta^{\mathrm{fh}}$ under condition (b).

Proof. We prove the two cases separately.

Case (a): global contraction. Since $s^{\mathrm{h}} = h(r^{\mathrm{h}})$ and $s^{\delta} = h_\delta(r^{\delta})$, the triangle inequality gives

$$\begin{aligned}\|s^{\mathrm{h}} - s^{\delta}\| &= \|h(r^{\mathrm{h}}) - h_\delta(r^{\delta})\| \\ &\le \|h(r^{\mathrm{h}}) - h_\delta(r^{\mathrm{h}})\| + \|h_\delta(r^{\mathrm{h}}) - h_\delta(r^{\delta})\|.\end{aligned} \tag{43}$$

By the definition of d_δ and the L_h-Lipschitz continuity of h_δ,

$$\|s^{\mathrm{h}} - s^{\delta}\| \le \|d_\delta\| + L_h\|r^{\mathrm{h}} - r^{\delta}\|. \tag{44}$$

Using Eq. equation 36,

$$\|s^{\mathrm{h}} - s^{\delta}\| \le \|d_\delta\| + L_h L_r \|s^{\mathrm{h}} - s^{\delta}\|. \tag{45}$$

Since $L_h L_r < 1$, rearranging yields

$$\|s^{\mathrm{h}} - s^{\delta}\| \le \frac{1}{1 - L_h L_r} \|d_\delta\|. \tag{46}$$

Applying Eq. equation 36 once more,

$$\|r^{\mathrm{h}} - r^{\delta}\| \le \frac{L_r}{1 - L_h L_r} \|d_\delta\|. \tag{47}$$

Therefore,

$$\|s^{\mathrm{h}} - s^{\delta}\| + \|r^{\mathrm{h}} - r^{\delta}\| \le \frac{1 + L_r}{1 - L_h L_r} \|d_\delta\|, \tag{48}$$

which proves Eq. equation 37.

Case (b): finite-horizon causal stability. For each causal block, define

$$x_t := \|s_t^{\mathrm{h}} - s_t^{\delta}\|, \qquad y_t := \|r_t^{\mathrm{h}} - r_t^{\delta}\|, \qquad d_t := \|h(r_t^{\mathrm{h}}) - h_\delta(r_t^{\mathrm{h}})\|. \tag{49}$$

By Eq. equation 40,

$$y_t \le L_c x_{t-1}. \tag{50}$$

Moreover, using the same decomposition as in Eq. equation 43,

$$\begin{aligned} x_t &\le d_t + L_h y_t \\ &\le d_t + L_h L_c x_{t-1}, \qquad x_0 = 0. \end{aligned} \tag{51}$$

Let

$$a := L_h L_c.$$

Unrolling Eq. equation 51 gives

$$x_t \le \sum_{q=0}^{t-1} a^q d_{t-q}. \tag{52}$$

Equation equation 52 is a finite discrete convolution. Young's convolution inequality therefore yields

$$\left(\sum_{t=1}^{H} x_t^2 \right)^{1/2} \le \left(\sum_{q=0}^{H-1} a^q \right) \left(\sum_{t=1}^{H} d_t^2 \right)^{1/2}. \tag{53}$$

Since the full trajectories are obtained by stacking the causal blocks,

$$\|s^{\mathrm{h}} - s^{\delta}\| = \left(\sum_{t=1}^{H} x_t^2 \right)^{1/2}, \qquad \|d_\delta\| = \left(\sum_{t=1}^{H} d_t^2 \right)^{1/2}. \tag{54}$$

Furthermore, Eq. equation 50 and $x_0 = 0$ imply

$$\begin{aligned} \|r^{\mathrm{h}} - r^{\delta}\| &= \left(\sum_{t=1}^{H} y_t^2 \right)^{1/2} \\ &\le L_c \left(\sum_{t=1}^{H} x_{t-1}^2 \right)^{1/2} \\ &\le L_c \left(\sum_{t=1}^{H} x_t^2 \right)^{1/2}. \end{aligned} \tag{55}$$

Combining Eqs. equation 53–equation 55, we obtain

$$\begin{aligned} \|s^{\mathrm{h}} - s^{\delta}\| + \|r^{\mathrm{h}} - r^{\delta}\| &\le (1 + L_c) \left(\sum_{t=1}^{H} x_t^2 \right)^{1/2} \\ &\le (1 + L_c) \left(\sum_{q=0}^{H-1} (L_h L_c)^q \right) \|d_\delta\|, \end{aligned} \tag{56}$$

which proves Eq. equation 41. □

Discussion. The two conditions provide complementary routes to the same network-level stability property. The global contraction condition yields a horizon-independent amplification factor provided $L_h L_r < 1$. In contrast, the finite-horizon causal condition does not require $L_h L_c < 1$: for every finite H, the geometric sum in Eq. equation 42 remains finite, although the amplification factor may grow with the unfolded horizon. This motivates the abstract factor Λ_δ used in Assumption 4.

The proposition only controls how a local smoothing discrepancy propagates through the network. The magnitude of the local discrepancy $\|d_\delta\|$ is controlled separately by the threshold-tail functional $\mathcal{T}_\delta$ in the convergence analysis. This separation allows trajectory stability and near-threshold activity to play distinct roles in the final bias bound.

A.4 Proofs for the Convergence Analysis

We prove the intermediate relations used in Section 6 and then assemble them to obtain Theorem 1.

Threshold-tail control of the local smoothing discrepancy. We first connect randomized smoothing to the threshold-tail functional.

Lemma 1. *For every $a \in \mathbb{R}$,*

$$|h(a) - h_\delta(a)| \le \Psi\left(\frac{|a|}{\delta}\right). \tag{57}$$

Consequently, for every W,

$$\mathbb{E}_\xi\left[\|d_\delta(W;\xi)\|^2\right] \le \mathcal{T}_\delta(W). \tag{58}$$

Proof. Recall that $h_\delta(a) = \mathbb{P}(\epsilon > -a/\delta)$. If $a > 0$, then $h(a) = 1$ and

$$|h(a) - h_\delta(a)| = \mathbb{P}(\epsilon \le -a/\delta) \le \mathbb{P}(|\epsilon| \ge |a|/\delta).$$

If $a < 0$, then $h(a) = 0$ and

$$|h(a) - h_\delta(a)| = \mathbb{P}(\epsilon > |a|/\delta) \le \mathbb{P}(|\epsilon| \ge |a|/\delta).$$

For $a = 0$, $|h(0) - h_\delta(0)| \le 1 = \Psi(0)$. Applying Eq. equation 57 coordinatewise gives

$$\|d_\delta(W;\xi)\|^2 \le \sum_{i=1}^{M} \Psi^2\left(\frac{|r_i^{\mathrm{hard}}(W;\xi)|}{\delta}\right) \le \sum_{i=1}^{M} \Psi\left(\frac{|r_i^{\mathrm{hard}}(W;\xi)|}{\delta}\right),$$

where the second inequality uses $0 \le \Psi \le 1$. Taking expectation over ξ proves Eq. equation 58. □

Conditional mean alignment. For deterministic surrogate-gradient OTTTA, the hard-forward trajectory is $s^{\mathrm{hard}}(W_k;\xi)$ and every spike derivative in the backward mapping is g_δ. By Eq. equation 8, the resulting single-sample direction is

$$G_\delta^{\mathrm{OTTT}}\big(s^{\mathrm{hard}}(W_k;\xi), r^{\mathrm{hard}}(W_k;\xi), W_k;\xi\big).$$

For normalized nodewise LocalZO, condition on $(\mathcal{F}_k, \xi)$. The hard-forward trajectory, membrane margins, presynaptic traces, and same-time spatial backward graph are then fixed. The single-sample direction can be expanded over the finite set of spatial backward paths $\mathcal{P}_{\mathrm{OTTT}}$ as

$$\widehat{g}^A(W_k;\xi,z) = \sum_{p \in \mathcal{P}_{\mathrm{OTTT}}} A_p(W_k;\xi) \prod_{q \in p} \overline{G}_2\big(r_q^{\mathrm{hard}}(W_k;\xi); z_q, \delta\big), \tag{59}$$

where $A_p(W_k;\xi)$ contains the factors independent of the LocalZO perturbations. Conditional independence and Eq. equation 23 imply

$$\mathbb{E}_z\left[\prod_{q \in p} \overline{G}_2\big(r_q^{\mathrm{hard}}; z_q, \delta\big) \,\middle|\, \mathcal{F}_k, \xi\right] = \prod_{q \in p} g_\delta\big(r_q^{\mathrm{hard}}\big). \tag{60}$$

Substitution into Eq. equation 59, followed by averaging over data and mini-batch sampling, gives Eq. equation 14.

Bias decomposition and component control. By Eq. equation 14 and Eq. equation 7,

$$\begin{aligned} b_k^A &= \mathbb{E}_\xi \left[G_\delta^{\mathrm{OTTT}}(s^{\mathrm{hard}}, r^{\mathrm{hard}}, W_k; \xi) - G_\delta^{\mathrm{OTTT}}(s^\delta, r^\delta, W_k; \xi) \,\middle|\, \mathcal{F}_k \right] \\ &\quad + \mathbb{E}_\xi \left[G_\delta^{\mathrm{OTTT}}(s^\delta, r^\delta, W_k; \xi) - G_\delta^{\mathrm{BPTT}}(s^\delta, r^\delta, W_k; \xi) \,\middle|\, \mathcal{F}_k \right] \\ &= b_{k,\mathrm{hs}}^A + b_{k,\mathrm{td}}^A. \end{aligned} \tag{61}$$

For the hard–soft component, Assumption 4 and Lemma 1 give

$$\begin{aligned} &\mathbb{E}_\xi \left[\left(\|s^{\mathrm{hard}}(W;\xi) - s^\delta(W;\xi)\| + \|r^{\mathrm{hard}}(W;\xi) - r^\delta(W;\xi)\| \right)^2 \right] \\ &\qquad \le \Lambda_\delta^2 \mathcal{T}_\delta(W). \end{aligned} \tag{62}$$

Jensen's inequality and Assumption 3 then imply

$$\begin{aligned} \|b_{k,\mathrm{hs}}^A\|^2 &\le \mathbb{E}_\xi \left[\left\| G_\delta^{\mathrm{OTTT}}(s^{\mathrm{hard}}, r^{\mathrm{hard}}, W_k; \xi) - G_\delta^{\mathrm{OTTT}}(s^\delta, r^\delta, W_k; \xi) \right\|^2 \,\middle|\, \mathcal{F}_k \right] \\ &\le L_{\mathrm{OTTT}}^2 \Lambda_\delta^2 \mathcal{T}_\delta(W_k). \end{aligned} \tag{63}$$

Taking expectation over the optimization randomness proves Eq. equation 16.

For the temporal-decoupling component, Jensen's inequality gives

$$\|b_{k,\mathrm{td}}^A\|^2 \le \mathbb{E}_\xi \left[\left\| G_\delta^{\mathrm{OTTT}}(s^\delta, r^\delta, W; \xi) - G_\delta^{\mathrm{BPTT}}(s^\delta, r^\delta, W; \xi) \right\|^2 \,\middle|\, \mathcal{F}_k \right]. \tag{64}$$

By the definition of D_δ^2 in Eq. equation 11, taking expectation over the optimization randomness proves Eq. equation 17. This completes the proof of Proposition 1.

Using Eq. equation 61 and $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$,

$$\mathbb{E}[\|b_k^A\|^2] \le 2L_{\mathrm{OTTT}}^2 \Lambda_\delta^2 \mathbb{E}[\mathcal{T}_\delta(W_k)] + 2\mathbb{E}[D_\delta^2(W_k)].$$

Averaging over $k = 0, \dots, K-1$ yields Eq. equation 18.

Biased descent on the smooth reference risk. Let

$$v_k := \widehat{g}_k^A - \mu_k^A.$$

Then $\mathbb{E}[v_k \mid \mathcal{F}_k] = 0$, and Assumption 2 gives

$$\mathbb{E}[\|v_k\|^2 \mid \mathcal{F}_k] \le \sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m} =: V. \tag{65}$$

By L_δ-smoothness,

$$F_\delta(W_{k+1}) \le F_\delta(W_k) - \eta \left\langle \nabla F_\delta(W_k), \widehat{g}_k^A \right\rangle + \frac{L_\delta \eta^2}{2} \|\widehat{g}_k^A\|^2. \tag{66}$$

Taking conditional expectation and using $\mu_k^A = \nabla F_\delta(W_k) + b_k^A$ gives

$$\begin{aligned} \mathbb{E}[F_\delta(W_{k+1}) \mid \mathcal{F}_k] &\le F_\delta(W_k) - \eta \left\langle \nabla F_\delta(W_k), \mu_k^A \right\rangle \\ &\quad + \frac{L_\delta \eta^2}{2} \|\mu_k^A\|^2 + \frac{L_\delta \eta^2}{2} V. \end{aligned} \tag{67}$$

Using $-\langle a, b\rangle \le \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$ and $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$,

$$\begin{aligned} \mathbb{E}[F_\delta(W_{k+1}) \mid \mathcal{F}_k] &\le F_\delta(W_k) - \eta \left(\frac{1}{2} - L_\delta \eta \right) \|\nabla F_\delta(W_k)\|^2 \\ &\quad + \eta \left(\frac{1}{2} + L_\delta \eta \right) \|b_k^A\|^2 + \frac{L_\delta \eta^2}{2} V. \end{aligned} \tag{68}$$

When $\eta \le 1/(4L_\delta)$,

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

Taking full expectation, summing over $k = 0, \dots, K-1$, and using $W_K \in \mathcal{U}$ together with $F_\delta(W_K) \ge F_\delta(W_{\mathcal{U}}^\star)$ gives

$$\frac{\eta}{4} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla F_\delta(W_k)\|^2] \le F_\delta(W_0) - F_\delta(W_{\mathcal{U}}^\star) + \frac{3\eta}{4} \sum_{k=0}^{K-1} \mathbb{E}[\|b_k^A\|^2] + \frac{L_\delta \eta^2 K}{2} V.$$

Dividing by $\eta K/4$ proves Eq. equation 19.

Hard–soft risk transfer. For any fixed W, Assumption 5 gives

$$|F(W) - F_\delta(W)| \le \mathbb{E}_\xi \left[\left| \mathcal{L}_T(s^{\mathrm{hard}}, W; \xi) - \mathcal{L}_T(s^\delta, W; \xi) \right| \right] \le L_s \mathbb{E}_\xi \left[\|s^{\mathrm{hard}} - s^\delta\| \right]. \tag{69}$$

Assumption 4, Cauchy–Schwarz, and Lemma 1 therefore imply

$$|F(W) - F_\delta(W)| \le L_s \Lambda_\delta \sqrt{\mathcal{T}_\delta(W)}. \tag{70}$$

In particular,

$$|F(W_{\mathcal{U}}^\star) - F_\delta(W_{\mathcal{U}}^\star)| \le L_s \Lambda_\delta \sqrt{\mathcal{T}_\delta(W_{\mathcal{U}}^\star)}. \tag{71}$$

Since R is uniform and independent of the optimization randomness,

$$\mathbb{E}[\mathcal{T}_\delta(W_R)] = \overline{\mathcal{T}}_{\delta,K}.$$

Applying Jensen's inequality to Eq. equation 70 gives

$$\mathbb{E}\left[|F(W_R) - F_\delta(W_R)|\right] \le L_s \Lambda_\delta \sqrt{\overline{\mathcal{T}}_{\delta,K}}. \tag{72}$$

Proof of Theorem 1. Combining Eq. equation 19 with Eq. equation 18 gives

$$\frac{1}{K} \sum_{k=0}^{K-1} \mathbb{E}[\|\nabla F_\delta(W_k)\|^2] \le \frac{4\big(F_\delta(W_0) - F_\delta(W_{\mathcal{U}}^\star)\big)}{\eta K} + 2L_\delta \eta \left(\sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m} \right) + 6L_{\mathrm{OTTT}}^2 \Lambda_\delta^2 \overline{\mathcal{T}}_{\delta,K} + 6\overline{D}_{\delta,K}^2. \tag{73}$$

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

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

Moreover, $W_R \in \mathcal{U}$ almost surely, so Assumption 6 implies

$$\mathbb{E}\left[F_\delta(W_R) - F_\delta(W_{\mathcal{U}}^\star)\right] \le \frac{1}{2\mu_{\mathcal{U}}} \mathbb{E}[\|\nabla F_\delta(W_R)\|^2]. \tag{74}$$

Substituting Eq. equation 73 into Eq. equation 74 gives

$$\mathbb{E}\left[F_\delta(W_R) - F_\delta(W_{\mathcal{U}}^\star)\right] \le \frac{2\big(F_\delta(W_0) - F_\delta(W_{\mathcal{U}}^\star)\big)}{\mu_{\mathcal{U}} \eta K} + \frac{L_\delta \eta}{\mu_{\mathcal{U}}} \left(\sigma_{\mathrm{mb}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{m} \right) + \frac{3L_{\mathrm{OTTT}}^2 \Lambda_\delta^2}{\mu_{\mathcal{U}}} \overline{\mathcal{T}}_{\delta,K} + \frac{3}{\mu_{\mathcal{U}}} \overline{D}_{\delta,K}^2. \tag{75}$$

Finally,

$$\mathbb{E}[F(W_R)] - F(W_{\mathcal{U}}^\star) \le \mathbb{E}\left[F_\delta(W_R) - F_\delta(W_{\mathcal{U}}^\star)\right] + \mathbb{E}\left[|F(W_R) - F_\delta(W_R)|\right] + |F(W_{\mathcal{U}}^\star) - F_\delta(W_{\mathcal{U}}^\star)|. \tag{76}$$

Applying Eqs. equation 72, equation 71, and equation 75 yields Eq. equation 13, completing the proof. □

B Experimental Setup and Diagnostic Protocol

Implementation and reproducibility. All experiments use the OTTTA update implemented in the accompanying codebase (`online_update=False`). The primary experiments use five independent training seeds, $\{2022, 2023, 2024, 2025, 2026\}$, whereas the smoothing-scale study uses seed 2022. We set the Python, NumPy, PyTorch, CUDA, and DataLoader random states accordingly. CuDNN deterministic mode is enabled and benchmark mode is disabled. The implementation uses PyTorch 2.1.0, torchvision 0.16.0, and CUDA 11.8. Unless otherwise stated, all runs use the weight-standardized `online_spiking_vgg11_ws` architecture with OnlineLIF neurons, membrane time constant $\tau = 2$, subtractive reset (`v_reset=None`), and the sigmoid surrogate with $\alpha = 4$.

Training configuration. All models are trained for 300 epochs with cosine annealing (`T_max`=300). SGD runs use momentum 0.9, and no weight decay is applied. The per-step loss is

$$\ell_t = \frac{(1-\beta)\ell_{\mathrm{CE}}(o_t, y) + \beta\ell_{\mathrm{MSE}}\big(o_t, \mathrm{onehot}(y)\big)}{T}, \tag{77}$$

with gradients accumulated over the T simulation steps before each parameter update. Classification uses the accumulated output $\sum_{t=1}^{T} o_t$. Dataset-specific settings are summarized in Table 3.

The theorem analyzes vanilla fixed-step OTTTA, whereas the experiments use practical optimizers; the experiments therefore evaluate the proposed approximation diagnostics rather than directly instantiate the optimization theorem.

Table 3: **Training configurations.**

Dataset	T	Batch	Optimizer	LR	β / Dropout
CIFAR-10	6	128	SGD	0.1	0.05 / 0
CIFAR-100	6	128	SGD	0.1	0.05 / 0
CIFAR10-DVS	10	128	SGD	0.1	0.001 / 0.1
DVS Gesture	20	16	Adam	0.001	0.001 / 0

Following the OTTT notation used in the main analysis, the theoretical per-step loss is written generically as $\ell(s^N[t], y)$. In the implementation, the corresponding current-step classifier output is denoted by o_t (`out_fr`). This output is computed from the current network state and introduces no independent temporal state. We therefore absorb the instantaneous readout into the generic loss notation, so that the theoretical $\ell(s^N[t], y)$ represents the implemented loss in Eq. equation 77 applied to the current-step output. The regularity assumptions in the analysis are imposed on this composed per-step loss over the parameter and trajectory region of interest.

Dataset Description and Processing We evaluate on four standard benchmarks. CIFAR-10 and CIFAR-100 are static RGB image classification datasets with 10 and 100 classes, respectively, each containing 50,000 training and 10,000 test 32×32 images. For both, training images use a 32×32 random crop with four-pixel padding, Cutout, and random horizontal flipping, followed by channel-wise normalization. Test images use tensor conversion and the same normalization. The same static image is presented at all $T = 6$ simulation steps.

CIFAR10-DVS is an event-based version of CIFAR-10 with 10 classes of DVS event streams. Its events are integrated into $T = 10$ frames by equal event count (`split_by=number`) using the default class-wise $90/10$ train/test split of the dataset loader. Training sequences are resized to 48×48, padded by four pixels, and consistently cropped to 48×48 across all frames of an event sequence. Test sequences are resized directly to 48×48. Both use channel mean $(0.2728, 0.1295)$ and standard deviation $(0.2225, 0.1290)$.

DVS Gesture is a standard event-based hand-gesture recognition benchmark with 11 gesture classes. We use its standard train/test split and integrate the events into $T = 20$ frames by equal event count without additional stochastic augmentation.

Diagnostic protocol. For each dataset, gradient diagnostics are evaluated on a fixed subset of $N_{\mathrm{diag}} = 128$ test examples sampled once with seed 2027. The same subset is used across all five training seeds and diagnostic checkpoints. Diagnostics are evaluated every 25 epochs from epoch 100 onward, including the final checkpoint (epoch 299 under zero-based indexing), with the model in evaluation mode. Experiment 1 uses the full checkpoint sequence, whereas Experiments 2–3 and the smoothing-scale study use final-checkpoint diagnostics.

A DataLoader batch size of eight is used only for data loading. Each diagnostic gradient is computed independently for one example, and squared discrepancies are averaged only after per-example differentiation. Thus, the reported quantities are averages of per-example squared gradient differences rather than squared differences between mini-batch-averaged gradients.

Diagnostic gradients. For each diagnostic example ξ, we compute

$$g_A^h(\xi), \qquad g_A^\delta(\xi), \qquad g_B^\delta(\xi).$$

The hard-forward gradient g_A^h is the OTTT gradient used by the training dynamics. The soft-OTTT gradient g_A^δ replaces the hard spike by the primitive sigmoid forward map while retaining OTTT-style temporal detachment. The soft-BPTT gradient g_B^δ uses the same soft forward dynamics while retaining the complete membrane and reset computational graph through time. Hence, g_A^h and g_A^δ isolate the hard–soft forward discrepancy, whereas g_A^δ and g_B^δ isolate temporal decoupling.

All gradients are flattened over the trainable parameters. Parameters receiving no gradient in a particular pass contribute zeros so that the three vectors have identical dimensionality.

Empirical approximation discrepancies. The raw hard–soft and temporal discrepancies are

$$H_k^2 = \frac{1}{N_{\text{diag}}} \sum_{i=1}^{N_{\text{diag}}} \left\| g_A^h(\xi_i) - g_A^\delta(\xi_i) \right\|_2^2, \qquad D_k^2 = \frac{1}{N_{\text{diag}}} \sum_{i=1}^{N_{\text{diag}}} \left\| g_A^\delta(\xi_i) - g_B^\delta(\xi_i) \right\|_2^2. \tag{78}$$

The temporal quantity D_k^2 is the finite-sample analogue of D_δ^2 in the theoretical analysis. The quantity H_k^2 directly measures the empirical hard–soft gradient discrepancy, while the corresponding theoretical hard–soft term is controlled through the threshold-tail quantity $\mathcal{T}_\delta$. Equation equation 78 averages per-example squared norms rather than taking the squared norm of an averaged gradient difference.

Because H_k^2 and D_k^2 have the same units, Experiment 2 uses D_k^2/H_k^2 to compare their relative magnitude within a training configuration. At the final checkpoint, both quantities are computed for each of the five seeds and averaged across seeds; the reported ratio is formed from the corresponding aggregate discrepancies.

Directional agreement. We measure directional agreement between soft-OTTT and soft-BPTT using

$$C_k = \frac{1}{N_{\text{diag}}} \sum_{i=1}^{N_{\text{diag}}} \frac{\left\langle g_A^\delta(\xi_i), g_B^\delta(\xi_i) \right\rangle}{\left\| g_A^\delta(\xi_i) \right\|_2 \left\| g_B^\delta(\xi_i) \right\|_2 + \varepsilon}, \qquad \varepsilon = 10^{-12}. \tag{79}$$

This scale-free statistic complements D_k^2 by measuring changes in gradient direction rather than magnitude alone.

Normalized discrepancies for training dynamics. Experiment 1 uses BPTT-normalized per-example discrepancies,

$$\widetilde{H}_k^2 = \frac{1}{N_{\text{diag}}} \sum_{i=1}^{N_{\text{diag}}} \frac{\left\| g_A^h(\xi_i) - g_A^\delta(\xi_i) \right\|_2^2}{\left\| g_B^\delta(\xi_i) \right\|_2^2 + \varepsilon}, \qquad \widetilde{D}_k^2 = \frac{1}{N_{\text{diag}}} \sum_{i=1}^{N_{\text{diag}}} \frac{\left\| g_A^\delta(\xi_i) - g_B^\delta(\xi_i) \right\|_2^2}{\left\| g_B^\delta(\xi_i) \right\|_2^2 + \varepsilon}. \tag{80}$$

Figure 2 reports their mean over the five seeds at each diagnostic checkpoint, with one-standard-deviation shading. The same positive display factor is applied to both curves within each task, preserving their relative magnitude and ordering. These normalized quantities are used only for the training-dynamics visualization. Experiment 2 uses the raw discrepancies in Eq. equation 78, whereas Experiment 3 uses the threshold-tail diagnostic for the hard–soft mechanism and D_k^2 for temporal decoupling.

Threshold-tail diagnostic. The primary experiments use the sigmoid surrogate with $\alpha = 4$. Its primitive soft forward map is $\sigma(\alpha r)$, with

$$\delta = \frac{1}{\alpha},$$

giving the default $\delta = 0.25$. The smoothing-scale study varies δ and sets $\alpha = 1/\delta$ accordingly. For the hard-forward membrane margin $r = u - V_{\text{th}}$, the implemented two-sided logistic tail is

$$\Psi\left(\frac{|r|}{\delta}\right) = \frac{2}{1 + \exp(|r|/\delta)}.$$

We use the node-normalized threshold-tail diagnostic

$$P_{\text{tail},k} = \frac{1}{N_{\text{diag}}} \sum_{i=1}^{N_{\text{diag}}} \frac{1}{M_i} \sum_{j=1}^{M_i} \Psi\left(\frac{|r_{ij}^{\text{hard}}(W_k)|}{\delta}\right), \tag{81}$$

where M_i is the number of recorded membrane margins for example i. For a fixed dataset, architecture, and temporal horizon, M_i is fixed, so this normalization preserves seed-wise ordering relative to the corresponding threshold-tail sum while mapping the diagnostic to $[0, 1]$. $P_{\mathrm{tail},k}$ is the theory-motivated hard–soft diagnostic used in Experiment 3 and the smoothing-scale study, whereas H_k^2 directly measures the empirical hard–soft gradient discrepancy.

Diagnostic–accuracy analysis. Experiment 3 uses five seed-level observations per dataset at the final checkpoint; different checkpoints from the same training run are not treated as independent samples. The diagnostic is selected according to the larger-magnitude mechanism identified independently of accuracy in Experiment 2. Specifically, we define

$$X_{d,s} = \begin{cases} P_{\mathrm{tail},k}(s), & d \in \{\text{CIFAR-10}, \text{CIFAR-100}\}, \\ \log D_k^2(s), & d \in \{\text{CIFAR10-DVS}, \text{DVS Gesture}\}, \end{cases} \tag{82}$$

where $s \in \{2022, \ldots, 2026\}$ denotes the training seed and k is the final checkpoint. $P_{\mathrm{tail},k}$ is already a normalized probability-type quantity, whereas D_k^2 is a positive squared-gradient quantity with larger scale variation across seeds; we therefore analyze the temporal diagnostic on the log scale.

For each dataset d, the task-wise statistic is

$$r_d = \mathrm{Pearson}\left(X_{d,s},\, A_{d,s}\right), \tag{83}$$

where $A_{d,s}$ is the test accuracy measured at the same final checkpoint. Using same-checkpoint accuracy avoids temporal misalignment between diagnostic and performance measurements.

To assess the four task-wise associations jointly while allowing their directions to differ across datasets, we use

$$T_{\mathrm{joint}} = \sum_{d=1}^{4} r_d^2. \tag{84}$$

Under the joint null of no diagnostic–accuracy association, accuracy values are permuted independently within each dataset while the corresponding diagnostic values are held fixed. We use $B = 200{,}000$ Monte Carlo permutations with random seed 2027 and compute

$$p_{\mathrm{perm}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\left\{T_{\mathrm{joint}}^{(b)} \geq T_{\mathrm{joint}}\right\}}{B + 1}. \tag{85}$$

Because only five seeds are available per dataset, task-wise coefficients and p-values are interpreted cautiously. We additionally report Spearman rank correlations and leave-one-seed-out Pearson correlations as sensitivity analyses.

B.1 Seed-wise Consistency of Discrepancy Dominance

To assess whether the cross-task pattern in Section 7.2 is an artifact of aggregation, we additionally compute the discrepancy ratio D_k^2/H_k^2 separately for each training seed. Table 4 reports the ratio of the across-seed mean discrepancies used in the main text, the mean and standard deviation of the five seed-wise ratios, and the number of seeds for which the temporal discrepancy exceeds the hard–soft discrepancy ($D_k^2/H_k^2 > 1$).

Table 4: **Seed-wise consistency of discrepancy dominance.**

Dataset	Aggregate ratio D_k^2/H_k^2	Seed-wise ratio mean $\pm$ std	$D_k^2 > H_k^2$ seeds
CIFAR-10	0.274	0.279 ± 0.046	$0/5$
CIFAR-100	0.250	0.267 ± 0.048	$0/5$
CIFAR10-DVS	1.200	1.221 ± 0.228	$4/5$
DVS Gesture	3.948	3.668 ± 1.769	$5/5$

The seed-wise results are consistent with the aggregate pattern in Section 7.2. The hard–soft discrepancy exceeds the temporal discrepancy for all five seeds on CIFAR-10 and CIFAR-100. In contrast, the temporal discrepancy is larger for four of five seeds on CIFAR10-DVS and for all five seeds on DVS Gesture. Thus, the cross-task differences in relative discrepancy magnitude are not solely induced by averaging across seeds, although CIFAR10-DVS remains close to the crossover between the two mechanisms.

B.2 Additional Diagnostic–Accuracy Results

Table 5 reports the detailed final-checkpoint analysis. Accuracy is measured at the same checkpoint as the diagnostic and reported as mean $\pm$ standard deviation across five seeds. For CIFAR-10 and CIFAR-100, r is computed using $P_{\mathrm{tail},k}$; for CIFAR10-DVS and DVS Gesture, it is computed using $\log D_k^2$. The task-wise p-values are two-sided and are reported for descriptive context.

Table 5: **Detailed final-checkpoint diagnostic–accuracy results.**

Dataset	Diagnostic	Accuracy (%)	r	p
CIFAR-10	$P_{\mathrm{tail},k}$	93.29 ± 0.14	-0.816	0.092
CIFAR-100	$P_{\mathrm{tail},k}$	69.74 ± 0.77	$+0.450$	0.447
CIFAR10-DVS	$\log D_k^2$	75.70 ± 1.01	-0.967	0.007
DVS Gesture	$\log D_k^2$	95.42 ± 0.93	-0.939	0.018

Three of the four datasets exhibit negative diagnostic–accuracy associations. The strongest occur on CIFAR10-DVS ($r = -0.967$) and DVS Gesture ($r = -0.939$), while CIFAR-10 also shows a large negative coefficient ($r = -0.816$). CIFAR-100 instead shows a moderate positive association ($r = 0.450$). The observed joint statistic is $T_{\mathrm{joint}} = 2.685$, and the within-dataset permutation analysis gives $p_{\mathrm{perm}} \approx 0.002$, providing evidence against the joint null of zero diagnostic–accuracy association across all four tasks. Given the small number of seeds, the task-wise coefficients should nevertheless be interpreted as descriptive evidence rather than as a universal performance-prediction rule.

Sensitivity analyses. We examine whether the task-wise conclusions depend strongly on rank structure or on any single training seed. Table 6 reports Spearman rank correlations and the range of the primary Pearson coefficient obtained by leaving out one of the five seeds at a time.

Table 6: **Sensitivity analyses for diagnostic–accuracy association.**

Dataset	ρ_s	LOO r range
CIFAR-10	-0.600	$[-0.954, -0.354]$
CIFAR-100	$+0.300$	$[-0.025, +0.896]$
CIFAR10-DVS	-1.000	$[-1.000, -0.960]$
DVS Gesture	-0.975	$[-0.997, -0.943]$

The CIFAR10-DVS and DVS Gesture associations remain strongly negative under every leave-one-seed-out analysis, and CIFAR-10 also retains a negative sign throughout. The CIFAR-100 association is less stable and can approach zero when individual seeds are removed. Rank-based correlations show the same overall pattern, with the strongest monotone associations on the two event-based datasets. Using the untransformed D_k^2 instead of $\log D_k^2$ gives Pearson correlations of -0.962 on CIFAR10-DVS and -0.731 on DVS Gesture, preserving the negative direction of both temporal associations.

B.3 Smoothing-Scale Trade-off

The smoothing-scale study isolates the dependence on δ while keeping the remaining CIFAR-10 training configuration fixed. It is intended as a theory-guided sensitivity analysis rather than a search for a universally optimal surrogate width.

Protocol. We evaluate

$$\delta \in \left\{\frac{1}{8}, \frac{1}{6}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}\right\}, \qquad \alpha = \frac{1}{\delta} \in \{8, 6, 4, 3, 2\},$$

using seed 2022. Each setting is trained independently from initialization for 300 epochs. Architecture, optimizer, learning-rate schedule, simulation horizon, augmentation, batch size, and loss are identical to the default CIFAR-10 configuration in Table 3. The default setting used elsewhere in the paper is $\delta = 1/4$ ($\alpha = 4$).

Measurements. For each run, we record the final-checkpoint test accuracy and evaluate $P_{\mathrm{tail},k}$, H_k^2, and D_k^2 at the same final checkpoint on the fixed $N_{\mathrm{diag}} = 128$ diagnostic examples used in the primary experiments. Each diagnostic is computed using the same δ as the corresponding training run, and D_k^2/H_k^2 is formed from these final-checkpoint discrepancies. The individual H_k^2 and D_k^2 values are retained in the result records in addition to their ratio.

Because $P_{\mathrm{tail},k}$ itself depends on δ, comparisons across the sweep quantify threshold sensitivity under each run's actual smoothing scale rather than under a common post-hoc reference width. Likewise, D_k^2/H_k^2 measures relative discrepancy magnitude and should not by itself be interpreted as an absolute change in either component. The sweep uses a single training seed and is therefore treated as a sensitivity study.