APPLIED BUT NOT RECEIVED: DESIGNING EFFECTIVE SURROGATE GRADIENTS FOR SPIKING NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Surrogate gradients are the standard tool for training spiking neural networks (SNNs), yet their true effect on the learning signal remains poorly understood. We show that the surrogate gradient (SG) applied at the spike nonlinearity is not the effective surrogate gradient (ESG) the network receives once it is propagated through the complete adjoint of the leaky integrate-and-fire (LIF) dynamics under backpropagation-through-time. Characterizing the ESG analytically, our analysis shows how common applied surrogate gradient definitions engineered to be smooth and symmetric yield effective surrogate gradients that are neither: discontinuous at threshold and asymmetric around it. These findings are consequential: the discontinuity inflates gradient variance, and the asymmetry biases learning toward non-spiking states, revealing an unintended mismatch between the forward and backward models. We turn this analysis of ESG dynamics into a design principle: rather than reshaping the applied surrogate alone, we design for the effective surrogate directly—shaping the applied surrogate and the LIF adjoint together—so that the effective gradient the network receives is continuous at threshold and restores intended credit assignment between silent and spiking regimes. The resulting design enables more stable and performant training of deep SNNs, demonstrating design of the effective surrogate gradient as a principled route to improved optimization.

1 INTRODUCTION

Spiking neural networks (SNNs) transmit information as sparse binary spikes, emitted when a neuron's membrane potential crosses a threshold, an event-driven asynchronous mechanism that motivates their use for low-power neuromorphic hardware. The same mechanism obstructs gradient-based training: the Heaviside step at the threshold is non-differentiable, leaving the exact gradient undefined. Surrogate gradients circumvent this by replacing the step's derivative with a smooth proxy, unlocking backpropagation through time (BPTT) for gradient-based SNN optimisation. How a surrogate actually shapes the learning signal, however, is less well understood.

Surrogate gradients originate from the straight-through estimator in binary neural networks, where they replace the non-differentiable Dirac-impulse derivative of the binarising Heaviside step function with a unit-gain local gradient (Courbariaux et al., 2015). Zenke & Ganguli (2018) showed that multilayer networks of spiking neurons can adopt this technique, and Neftci et al. (2019) consolidated the approach, making surrogate-based BPTT the standard route to gradient-based SNN training. Current research investigates what the surrogate gradient fundamentally is—connecting it to the derivative of a stochastic firing intensity whose shape follows the escape-noise profile of a probabilistic neuron (Gygax & Zenke, 2025), and to a frequency-domain view that characterises surrogate gradients as a Fourier transform(Chen et al., 2026). Beyond interpretation, the majority of the literature focuses on the design of performant surrogates. This design space has since been mapped chiefly along shape and scale: Zenke & Vogels (2021) find learning to be remarkably robust to the surrogate's shape yet sensitive to its scale. Current trends make the surrogate adaptive rather than fixed—aligned to the network's initialisation and weight statistics (Herranz-Celotti & Rouat, 2022; Chang et al., 2024), adjusted online from global task performance (Van den Berghe et al., 2026) or each neuron's local membrane variance (Shi & Yu, 2026). An alternative strand designs

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.

appropiate surrogate shapes to enhance computational efficiency, such as sparsifying surrogates below a near-threshold boundary to avoid computation Perez-Nieves & Goodman (2021) or through dynamic quantization of the surrogate's shape(Hou et al., 2026). A parallel effort questions the smooth, symmetric, strictly non-negative template itself, admitting asymmetric or locally negative surrogates (Yin et al., 2021), tuning that asymmetry dynamically to stabilise gradient flow (Kang et al., 2026).

Within the current paradigm, surrogate gradients are assessed at the same location where they are applied, while backpropagation through time (BPTT) transforms them through the leaky integrate-and-fire (LIF) adjoint before they reach the rest of the network. In this work, we make this transformation explicit by deriving the received *effective surrogate gradient* (ESG) (Figure 1) and use it to analyse and quantitatively compare common surrogate functions. This perspective complements existing work by proposing the ESG as the more appropriate design space to be explored, and provides the practical tools to determine, analyse and consequently design it in order to shape gradient flow and improve the performance of spiking neural networks.

We make three contributions. (1) **Analytical.** We characterise in closed form the ESG—what a network receives once the applied surrogate is propagated through the complete LIF adjoint under BPTT—and show that applied surrogates built to be smooth and symmetric induce effective ones that are neither: discontinuous at threshold and asymmetric around it, inflating gradient variance and biasing learning toward non-spiking states. (2) **Design.** We turn this into a design rule: because reshaping the applied surrogate alone cannot correct the effective surrogate robustly, we design for the effective gradient directly—shaping the applied surrogate and the LIF adjoint together, realised as our cdx-t* construction—so its effective counterpart is continuous at threshold and restores balanced credit across silent and spiking regimes, removing the variance inflation and bias. (3) **Empirical.** Across three neuromorphic benchmarks and a depth$\times$width grid, holding the applied surrogate fixed, the resulting design trains SNNs more accurately in the majority of configurations, most clearly at depth.

2 METHODOLOGY

We analyse the leaky integrate-and-fire (LIF) spiking neuron with a hard, detached reset(Zenke & Vogels, 2021);. H is the forward Heaviside spike map, G the applied surrogate gradient substituted for H' in the backward pass, and G_{eff} the *effective* surrogate gradient—the shape G impresses on the loss landscape through its interaction with the LIF adjoint, carried upstream to earlier layers and their parameters. Figure 1 shows the computational graph this section follows.

2.1 FORWARD PASS: LIF DYNAMICS

Time is discrete, $t \in \{1, \dots, T\}$; the neuron carries a membrane potential $U_t \in \mathbb{R}$, read before the nonlinearity, and emits a binary spike through the Heaviside step

$$S_t \;=\; H(U_t;\theta) \;=\; \mathbf{1}[\,U_t \geq \theta\,] \;\in\; \{0,1\}, \tag{1}$$

with θ the firing threshold . With leak $\beta \in [0,1)$ and reset potential U_{reset}, the membrane leaks, integrates the input current $I_t \in \mathbb{R}$, and resets at a spike:

$$U_t \;=\; \beta\, V_{t-1} + I_t, \qquad V_{t-1} \;=\; (1 - S_{t-1})\, U_{t-1} + S_{t-1}\, U_{\text{reset}}, \tag{2}$$

where V_{t-1} is the state carried into step t: the retained membrane U_{t-1} on a silent step ($S_{t-1} = 0$), the reset value on a spiking step ($S_{t-1} = 1$). The reset is *hard* (a replacement) and, in the backward pass, *detached*(Zenke & Vogels, 2021): it severs the temporal gradient path at a spike (subsection 2.2).

2.2 BACKWARD PASS: THE SURROGATE GRADIENT AND THE ADJOINT

The Heaviside derivative is zero away from θ and singular on it, so training replaces H' with a smooth applied surrogate G . In the shifted coordinate $z_t = U_t - \theta$, the controlled analysis instantiates G as the Gaussian gate

$$G(u) \;=\; \exp\!\big(-\tfrac{1}{2}\, k\, z^2\big)\,, \qquad z = u - \theta, \tag{3}$$

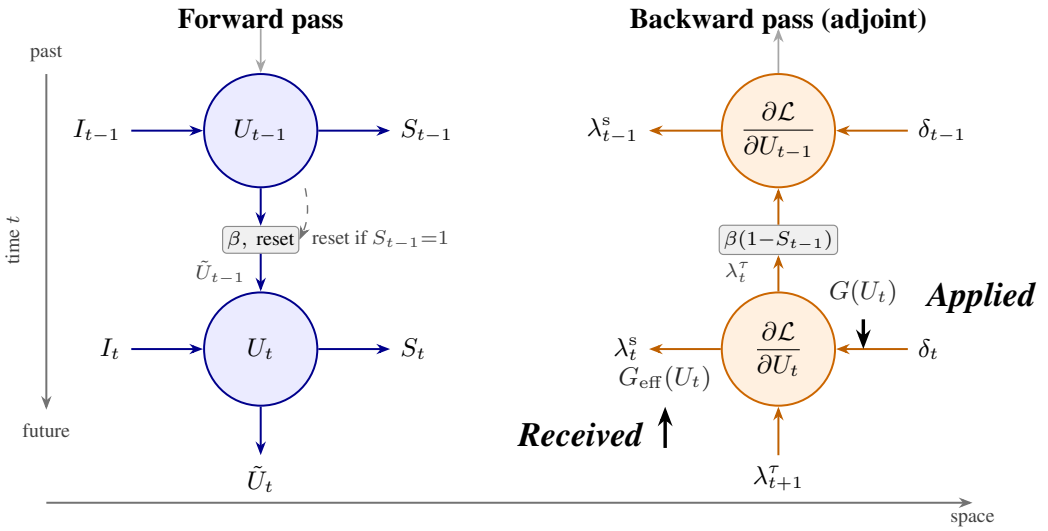


Figure 1: ***Applied but not received***: The surrogate gradient $G(U_t)$ applied at the spike nonlinearity is not equal to one received $G_{eff}(U_t)$ after interaction with the spiking neuron's LIF adjoint. This figure displays the computational graph of a single-neuron LIF neuron across time. *Left:* the forward pass—membrane U_t, Heaviside spike $S_t = H(U_t; \theta)$, and the leaky, hard-reset temporal recurrence. *Right:* the backward pass (adjoint) in cotangents, marking where the applied surrogate G enters and where the effective surrogate G_{eff} is read—on the spatial cotangent edge passed upstream.

a unit-peak bump on θ with slope $k > 0$ (sharper for larger k); section 3 substitutes other surrogates.

Backpropagation-through-time runs a reverse-time adjoint over Equation 2 (Figure 1, right) . It transports a *spatial* cotangent λ^{s}_t (emitted upstream) and a *temporal* cotangent λ^{τ}_t (passed to the previous step), given the downstream cotangent $\delta_t \equiv \partial\mathcal{L}/\partial S_t$ (the loss sensitivity to the emitted spike) and the future λ^{τ}_{t+1}. Inline with proceeding denotations, the *conventional* adjoint then returns

$$\lambda^{\tau}_t \;=\; \underbrace{G(U_t)\,\delta_t}_{\text{spatial drive}} \;+\; \underbrace{\beta\,(1-S_t)\,\lambda^{\tau}_{t+1}}_{\text{temporal transfer}}, \qquad \lambda^{\mathrm{s}}_t \;=\; \lambda^{\tau}_t, \qquad \lambda^{\tau}_{T+1} = 0. \tag{4}$$

The spatial drive $G(U_t)\,\delta_t$ has the surrogate gate the contribution to the arriving spatial gradients; the temporal term leaks the future cotangent back, cut at a spike by $(1 - S_t)$. Since $\beta(1 - S_t)$ vanishes on spikes, the recursion truncates at the next spike $s^*(t) = \min\{s \geq t : S_s = 1\}$, and the spatial cotangent unrolls as

$$\lambda^{\mathrm{s}}_t \;=\; \sum_{s=t}^{s^*(t)} \beta^{\,s-t}\, G(U_s)\,\delta_s, \tag{5}$$

a β-discounted sum of surrogate-graded cotangents along the current inter-spike interval: above threshold it is the single term $G(U_t)\delta_t$, while below threshold a neuron accrues credit from its future near-threshold visits even where its local gate $G(U_t) \approx 0$ (Appendix A).

2.3 The effective surrogate gradient

A spiking neuron influences the loss only when it fires: within an inter-spike interval (ISI) the silent steps emit nothing, and the leak's linear low-pass dynamics only describe how the membrane reached the terminating spike, which the reset then severs from the next ISI. An adjoint should therefore admit credit only there—at the crossing, decayed along the path that led to it. We measure how an adjoint does so through the *effective surrogate gradient*: the credit it delivers upstream, conditioned on the membrane,

$$G_{\mathrm{eff}}(u) \equiv \mathbb{E}\Big[\tfrac{\partial\mathcal{L}}{\partial U_t}\,\Big|\,U_t = u\Big] = \mathbb{E}[\lambda^{\mathrm{s}}_t \,|\, U_t = u]\,, \qquad \bar{\lambda} \equiv \mathbb{E}[\lambda^{\mathrm{s}}_t] = \int G_{\mathrm{eff}}(u)\,\pi(u)\,\mathrm{d}u, \tag{6}$$

with $\pi(u)$ the membrane potentials density; the occupancy-weighted $G_{\text{eff}}(u)\,\pi(u)$ is the density of delivered credit and integrates to the marginal mean cotangent $\bar{\lambda}$.

Over one ISI ending in a spike at $T = s^*(t)$ (phase $\tau = T - t$), the expectation of the unrolled conventional adjoint (Equation 5; δ_t drawn independently of the state) splits the delivered credit in two,

$$\mathbb{E}[\lambda_t^{\text{s}}] = \underbrace{\beta^\tau\, G(U_T)\,\bar{\delta}}_{\text{spike step: one influence, leak-decayed}} + \underbrace{\sum_{s=t}^{T-1} \beta^{\,s-t}\, G(U_s)\,\bar{\delta}}_{\text{silent steps: credit before firing}}, \qquad \bar{\delta} = \mathbb{E}[\delta_t]. \tag{7}$$

The spike term is the credit the neuron's one influence moment deserves, decayed by the leak that carries it; the silent sum credits steps at which the neuron did not act. That accumulation is not purely harmful—sharing future terms across the ISI lowers the variance between successive λ_t^{s}—but it credits the neuron as if it influenced the loss continuously through time, a non-spiking RNN, rather than only at its spike.

Analytically, the conventional adjoint's effective surrogate therefore departs from a profile peaked at θ (visual aid at Figure 5), continuous, and symmetric in three linked ways: a downward threshold jump ΔG (with slope kink κ)—the pedestal is present at θ^- but truncated at θ^+, so membrane states an infinitesimal step apart receive sharply different credit; a sub-threshold-displaced mode $u_-^\star < \theta$ (while the super mode $u_+^\star$ sits at θ), a credit misalignment that spends influence off the decision boundary; and sub-heavy cross-threshold credit mass, $\rho = m_-/m_+ > 1$, so two neurons just below and just above θ are trained with different effective gains. The per-state dispersion—the gradient variance the jump inflates—is a further such departure. Each is a loss of fidelity to the forward pass, which acts only at the spike; the simulation analysis of section 3 confirms them empirically.

These are all departures from one ideal: admitting credit only at the optimal admission step $t^\star$—the instant the membrane crosses θ—decayed to the present,

$$G_{\text{eff}}^{t^\star}(u) = \mathbb{E}\Big[\, \beta^{\,t^\star - t}\, G(U_{t^\star})\, \delta_{t^\star} \;\Big|\; U_t = u \,\Big]. \tag{8}$$

On the discrete grid $t^\star$ is inaccessible—it is the continuous threshold-crossing time evaluated exactly by event-based gradient methods such as Mészáros et al. (2025)—and need not be known: what an adjoint controls is *where* it admits credit, and Equation 8 is the reference a redesigned adjoint should approach. (Appendix A writes the conventional G_{eff} explicitly by sub- and super-threshold domain, gives the equivalent linear applied-to-effective operator and its $\beta = 0$ / recurrence-free limit $G_{\text{eff}} = G$, and illustrates these shape departures.)

2.4 Surrogate Adjoint redesign alligned to the effective surrogate

To remove the conventional adjoint's discrepancies we must move them out of G_{eff}—the object the rest of the network actually uses. One route repositions them onto the *applied* surrogate: reshape G so that, after the adjoint, its effective counterpart is smooth and threshold-centred (subsection B.4). But a static reshaping loads the whole correction onto the gate and is brittle—it fixes the leak and cannot follow the operating point as training moves it. We therefore intervene in the temporal adjoint itself, at the source named in equation 7: the silent sum. Each rule below is a different estimate of the admission step $t^\star$ of Equation 8.

Sparse surrogate A crude intervention inspired by the work by Perez-Nieves & Goodman (2021) deletes the silent accumulation outright, leaving $\lambda_t^{\text{s}} \to \beta^\tau G(U_T)\delta_T$, the decayed spike credit alone. approximate this with a hard relative sub-threshold $\vartheta_{\text{sub}} = \beta\theta$, gating silent steps off below it with $\widehat{G}(u) = G(u)\,\mathbf{1}[u \geq \vartheta_{\text{sub}}]$:

$$\lambda_t^{\text{s}} = \widehat{G}(U_t)\,\delta_t + \beta(1 - S_t)\,\lambda_{t+1}^{\text{s}}, \tag{9}$$

truncating the accumulation to the near-threshold band $[\vartheta_{\text{sub}}, \theta)$—implicitly placing $t^\star$ near the crossing—for computational efficiency, with learning unimpaired provided β stays clear of ϑ_{sub}. This raises fidelity but is brittle—a hard cut, meticulous to tune—and the axis is otherwise little explored; it keeps the sign of the θ-jump and adds a second jump at ϑ_{sub} (Appendix A).

cdx-t* (ours). We estimate $t^\star$ from the realization instead. The crossing lies between the last silent step $T-1$ and the spike T, and quantized input can make U_T overshoot θ, so neither endpoint alone is a reliable gate. We carry only the decayed spike credit, gated at the better of the two *observed* steps $\{T-1, T\}$,

$$\widetilde{G}_T = G(U_{T^\star}), \qquad T^\star = \underset{s\in\{T-1,\,T\}}{\arg\max}\; w_s\, G(U_s), \quad w_T = 1,\;\; w_{T-1} = \beta^{\,1-\eta}, \tag{10}$$

with $\eta \in [0,1]$ a hyperparameter fixed at $\frac{1}{2}$ (only the selection is weighted—the value $\widetilde{G}_T$ is the winner's full surrogate; Appendix A), and keep a bounded instantaneous floor $(1-\beta)$ on silent steps via a carrier seeded at the spike:

$$\Psi^\star_t = S_t\,\delta_t\,\widetilde{G}_T + \beta(1-S_t)\,\Psi^\star_{t+1} = \beta^\tau \widetilde{G}_T\,\delta_T, \qquad \lambda^{\mathrm{s}}_t = (1-\beta)(1-S_t)\,G(U_t)\,\delta_t + \Psi^\star_t. \tag{11}$$

Its expectation carries the same two-part reading as equation 7, the accumulation now replaced by a bounded floor,

$$\mathbb{E}[\lambda^{\mathrm{s}}_t]\big|_{t<T} = \underbrace{(1-\beta)\,G(U_t)\,\bar{\delta}}_{\text{instantaneous floor}} + \underbrace{\beta^\tau\,\widetilde{G}_T\,\bar{\delta}}_{\text{decayed spike credit}}. \tag{12}$$

The rule is bounded, $|\lambda^{\mathrm{s}}_t| \le \max_s |\delta_s|$, and reduces to the instantaneous gate at $\beta = 0$; under stationary δ it places both modes at θ and closes the jump ($\Delta G = 0$; Appendix A). Its effective surrogate $\hat{G}^{t^\star}_{\mathrm{eff}}$ thus approximates the objective $G^{t^\star}_{\mathrm{eff}}$ of Equation 8, forgoing the sparse hard cut and its tuning while removing the accumulation.

adjoint	admits credit at	θ-jump ΔG	bounded
conventional	every step (no $t^\star$)	sub-higher, unbounded	no
sparse	band $[\vartheta_{\mathrm{sub}}, \theta)$	reduced, same sign	no
cdx-t*	$T^\star \in \{T-1, T\}$	0 (stationary δ)	yes

subsection 3.1 quantifies this comparison at a fixed Gaussian surrogate.

3 Results and Discussion

3.1 Simulation analysis

We measure the applied-to-effective map on a single LIF node: a downstream cotangent is drawn and propagated through the adjoint (Figure 1) on an input current that holds the membrane stationary Gaussian (subsection 2.3). At leak β, for sample i and step t,

$$I_{t,i} \sim \mathcal{N}(\mu_I, \sigma_I^2), \qquad U_{t,i} \sim \mathcal{N}(\mu_u, \sigma_u^2), \qquad \delta_{t,i} \sim \mathrm{Lognormal}(0,1), \tag{13}$$

where $t = 1, \dots, T$ indexes time and $i = 1, \dots, N$ the trajectories; $I_{t,i}$ is the input current, whose moments (μ_I, σ_I^2) are calibrated (subsection B.1) so the membrane $U_{t,i}$ from Equation 2 is stationary Gaussian $\mathcal{N}(\mu_u, \sigma_u^2)$, and $\delta_{t,i} = e^\xi$, $\xi \sim \mathcal{N}(0,1)$ is the log-normal downstream cotangent. The adjoint (Equation 4) then gives $\lambda^{\mathrm{s}}_{t,i}$, and G_{eff} is its binned conditional mean over a warm-up/truncation-trimmed window (Appendix B). We use 8 seeds $\times$ 1024 trajectories of $T = 500$ steps and sweep $\beta \in \{0, 0.5, 0.9\}$ at $\mu_u = 1$.

We run this for five common applied surrogate families (Gaussian, fast-sigmoid, sech^2, boxcar, triangular; Appendix B) , their slopes calibrated to equal participation—a matched overlap under a common reference membrane—so shape, not width, is compared.[1] The resulting shape scores of subsection 2.3 are in Table 1. Biasing the applied surrogate toward the super-threshold side—by rotation or translation (subsection B.4)—shifts credit across θ but leaves the mode offsets and the threshold discontinuity intact: no single static asymmetry corrects them consistently across the leak slice. Analyses extended to asymmetric surrogates uphold this conclusion in subsection B.4.

Reshaping the applied surrogate cannot remove the discrepancy consistently, but adjusting the *adjoint* can. Holding the forward LIF and the Gaussian ($k = 4$) applied surrogate fixed, we compare

[1] Each slope k solves $I(\mathrm{sg}, k) = \mathbb{E}_{u\sim\mathcal{N}(\theta,1)}[\mathrm{sg}(u;k)] = I_{\mathrm{ref}}$, with I_{ref} the Gaussian overlap at k=4 ($1/\sqrt{5} \approx 0.447$); see Appendix B.

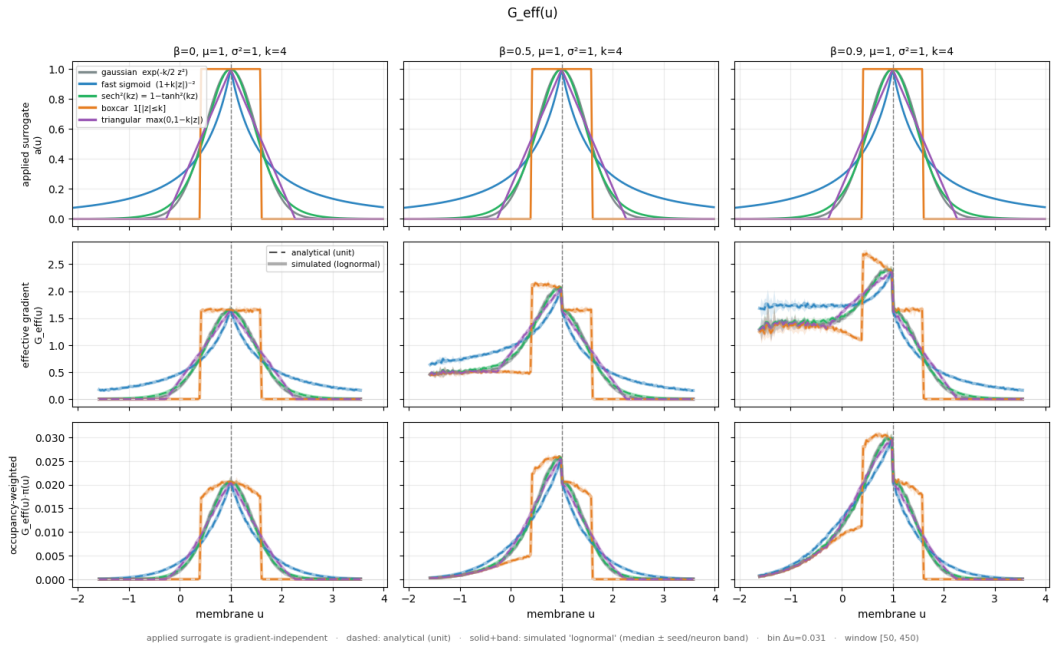


Figure 2: *Symmetric* sub-study (log-normal gradient), five calibrated surrogates overlaid. *Rows:* applied (top), effective G_{eff} (middle), occupancy-weighted $G_{\text{eff}}\pi$ (bottom); *columns:* $\beta \in \{0, 0.5, 0.9\}$. At β=0 applied and effective coincide; leakage grows the sub-threshold flank and threshold jump, differently per shape. Calibration and estimator in Appendix B.

Table 1: *Symmetric* effective-surrogate shape scores (log-normal gradient), five calibrated surrogates across $\beta \in \{0, 0.5, 0.9\}$ at μ=1. Each β block reports the threshold jump ΔG, the sub-/super-threshold modes $u^\star_-/u^\star_+$ (relative to θ), and the occupancy-weighted mass split (sub/super, %); slopes k are calibrated to equal participation (Appendix B). Conventional adjoint throughout; the mass split is shape-invariant (occupancy-set) and grows sub-heavy with the leak. Adjoint variants in Table 2.

	$\beta = 0$			$\beta = 0.5$			$\beta = 0.9$		
Surrogate (k)	$\Delta G\,[-]$	$u^\star_-/u^\star_+\,[-]$	mass [%]	$\Delta G\,[-]$	$u^\star_-/u^\star_+\,[-]$	mass [%]	$\Delta G\,[-]$	$u^\star_-/u^\star_+\,[-]$	mass [%]
Gaussian (4.00)	+0.01	−0.02/+0.02	50/50	+0.41	−0.05/+0.02	62/38	+0.72	−0.14/+0.05	71/29
Fast-sigmoid (0.86)	+0.01	−0.02/+0.02	50/50	+0.42	−0.02/+0.02	62/38	+0.76	−0.02/+0.02	71/29
Sech^2 (1.55)	+0.01	−0.02/+0.02	50/50	+0.42	−0.05/+0.02	62/38	+0.72	−0.05/+0.05	71/29
Boxcar (0.59)	+0.01	−0.39/+0.45	50/50	+0.41	−0.52/+0.30	62/38	+0.71	−0.52/+0.39	71/29
Triangular (0.79)	+0.01	−0.02/+0.02	50/50	+0.42	−0.02/+0.02	62/38	+0.72	−0.02/+0.02	71/29

the conventional spike-detached adjoint against a β-matched *Sparse SG* (its silent-step surrogate gated off below a leak-dependent fraction of θ) and our *cdx-t** operator (subsection 2.4); the effective surrogates are read at the input layer of a four-layer stack (subsection B.5). Across the leak slice, cdx-t* closes the threshold discontinuity and restores the balanced sub/super credit split that the conventional adjoint loses under leakage, while the Sparse SG recovers it only in part (Figure 3, Table 2); this is the basis for the empirical study below.

3.2 Empirical Analysis

We test whether designing the surrogate for its effective counterpart improves training. Because the applied surrogate and the complete LIF adjoint jointly set the effective surrogate, we hold the applied surrogate fixed (Gaussian, k=4) and compare three adjoint constructions that realise different effective surrogates—the conventional spike-detached adjoint, the β-matched Sparse SG, and our cdx-t* (subsection 2.4)—on three neuromorphic classification benchmarks (introduced below) across a grid of depths $L \in \{2, 3, 4\}$ and hidden widths $H \in \{32, 64, 128\}$. Architecture and op-

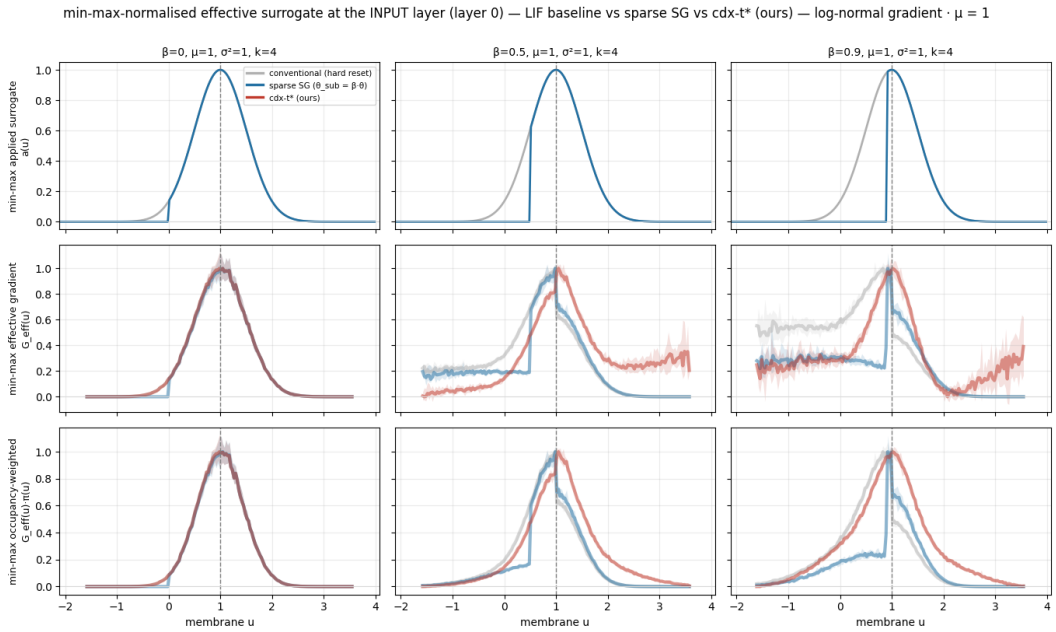


Figure 3: Effective surrogate under three adjoints at a fixed applied surrogate (Gaussian, k=4), read at the input layer of the four-layer stack. Every panel is min–max normalised to $[0, 1]$, so the adjoints are compared by shape, not scale (the scores of Table 2 are correspondingly scale-free). *Rows:* applied (top), effective G_{eff} (middle), occupancy-weighted $G_{\text{eff}}\pi$ (bottom); *columns:* $\beta \in \{0, 0.5, 0.9\}$; overlaid: conventional (grey), Sparse SG (blue), cdx-t* (ours, red), with solid = median, band = seed/neuron spread. cdx-t* has no pointwise applied surrogate, so the top row omits it; under leakage it keeps the effective mode on θ and the credit balanced, where the conventional adjoint grows a sub-threshold flank and jump.

Table 2: Effective-surrogate shape scores under three adjoints at a fixed applied surrogate (Gaussian, k=4), read at the input layer of the four-layer stack (subsection B.5). Columns as in Table 1, with ΔG read on the min–max-normalised curves of Figure 3 (scale-free, comparable across adjoints of different magnitude). Our cdx-t* (subsection 2.4) closes the jump and restores the balanced split at every leak; the conventional adjoint does neither under leakage, and Sparse SG only in part.

	$\beta = 0$			$\beta = 0.5$			$\beta = 0.9$		
Adjoint	$\Delta G\,[-]$	$u^\star_-/u^\star_+\,[-]$	mass [%]	$\Delta G\,[-]$	$u^\star_-/u^\star_+\,[-]$	mass [%]	$\Delta G\,[-]$	$u^\star_-/u^\star_+\,[-]$	mass [%]
Conventional	0.02	−0.02/+0.05	50/50	0.36	−0.02/+0.05	67/33	0.50	−0.17/+0.05	77/23
Sparse SG	0.04	−0.02/+0.05	49/51	0.27	−0.02/+0.05	59/41	0.04	−0.08/+0.05	52/48
cdx-t* (ours)	0.02	−0.02/+0.05	50/50	0.16	−0.02/+0.05	44/56	0.02	−0.02/+0.02	52/48

timiser are common to every cell: each input frame is linearly projected to H units, with an affine batch-normalisation layer (default initialisation), and fed to a stack of L LIF layers of width H; a read-out layer then assigns one LIF neuron to each class and scores it by that neuron's total spike count over the sequence, trained with cross-entropy. Every LIF neuron has a fixed threshold $\theta = 1$, reset $U_{\text{reset}} = 0$, and a neuron-wise trainable leak β, initially 0.9; weights use the default Kaiming-uniform initialisation, and all cells are optimised with Adam at its default hyperparameters, without per-configuration tuning. Figure 4 plots the out-of-sample accuracy over training and Table 3 the final tallies.

We evaluate on three neuromorphic classification benchmarks: *ECG*, heartbeat classification from level-crossing spikes (6 classes, 4 channels, fixed 1300-step sequences, step-wise labels); *SHD*, spoken-digit recognition from model-cochlea spikes (20 classes, 700 channels); and *DVS Gesture*, hand-gesture recognition from an event camera (11 classes, a downsampled $64 \times 64 \times 2$ event grid) (Laguna et al., 1997; Cramer et al., 2020; Amir et al., 2017). The three span a range of input

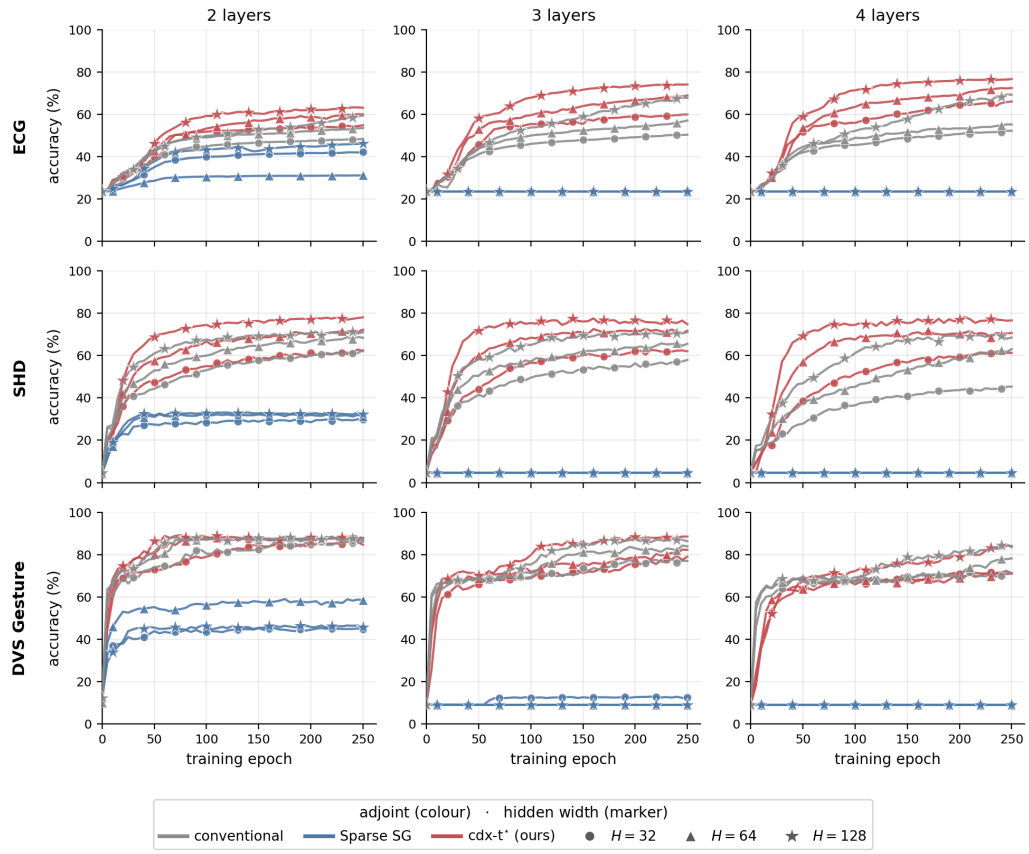


Figure 4: Out-of-sample (test) accuracy over training, three benchmarks. *Rows:* datasets (ECG, SHD, DVS Gesture); *columns:* depth (2–4 LIF layers). Overlaid per panel: conventional (grey), Sparse SG (blue), cdx-t* (ours, red), at widths $H \in \{32, 64, 128\}$ (circle/triangle/star); lines are the seed mean. Applied surrogate fixed (Gaussian k=4). Tallies in Table 3; training curves in Figure 11.

Table 3: Final out-of-sample (test) accuracy (%) at epoch 250, mean$_{\pm\text{std}}$ over five seeds, across datasets, network depth, and hidden width, for the conventional, Sparse SG, and cdx-t* (ours) adjoints at a fixed applied surrogate (Gaussian slope k=4). Winner (dataset, depth, width) is in bold.

		Conventional			Sparse SG			cdx-t* (ours)		
Dataset	**Depth**	32	64	128	32	64	128	32	64	128
ECG	2 layers	$48.3_{\pm 2.3}$	$53.5_{\pm 0.7}$	$59.5_{\pm 2.8}$	$42.2_{\pm 17.3}$	$31.2_{\pm 10.1}$	$46.2_{\pm 20.8}$	$\mathbf{54.8}_{\pm 3.9}$	$\mathbf{60.0}_{\pm 5.0}$	$\mathbf{63.1}_{\pm 4.5}$
	3 layers	$50.4_{\pm 2.7}$	$57.0_{\pm 3.3}$	$67.7_{\pm 3.2}$	$23.6_{\pm 0.0}$	$23.6_{\pm 0.0}$	$23.6_{\pm 0.0}$	$\mathbf{59.9}_{\pm 1.2}$	$\mathbf{68.9}_{\pm 1.9}$	$\mathbf{74.1}_{\pm 1.8}$
	4 layers	$52.2_{\pm 1.2}$	$55.2_{\pm 1.9}$	$69.3_{\pm 2.7}$	$23.6_{\pm 0.0}$	$23.6_{\pm 0.0}$	$23.6_{\pm 0.0}$	$\mathbf{66.1}_{\pm 4.9}$	$\mathbf{72.4}_{\pm 5.2}$	$\mathbf{76.7}_{\pm 1.3}$
SHD	2 layers	$\mathbf{62.5}_{\pm 1.2}$	$68.4_{\pm 0.9}$	$71.1_{\pm 1.0}$	$29.8_{\pm 9.3}$	$31.4_{\pm 4.6}$	$32.5_{\pm 10.6}$	$62.1_{\pm 1.7}$	$\mathbf{72.2}_{\pm 0.7}$	$\mathbf{78.1}_{\pm 0.9}$
	3 layers	$57.9_{\pm 1.8}$	$65.6_{\pm 2.7}$	$71.1_{\pm 1.9}$	$4.5_{\pm 0.0}$	$4.5_{\pm 0.0}$	$4.5_{\pm 0.0}$	$\mathbf{62.0}_{\pm 1.7}$	$\mathbf{71.7}_{\pm 0.8}$	$\mathbf{74.8}_{\pm 1.8}$
	4 layers	$45.4_{\pm 7.9}$	$63.0_{\pm 2.4}$	$68.5_{\pm 1.8}$	$4.5_{\pm 0.0}$	$4.5_{\pm 0.0}$	$4.5_{\pm 0.0}$	$\mathbf{61.3}_{\pm 1.6}$	$\mathbf{70.7}_{\pm 2.8}$	$\mathbf{76.6}_{\pm 1.2}$
DVS Gesture	2 layers	$\mathbf{86.1}_{\pm 1.8}$	$\mathbf{87.9}_{\pm 0.4}$	$86.6_{\pm 1.5}$	$44.6_{\pm 3.2}$	$58.5_{\pm 24.9}$	$45.7_{\pm 8.8}$	$84.6_{\pm 2.1}$	$87.6_{\pm 1.7}$	$\mathbf{87.8}_{\pm 0.8}$
	3 layers	$77.0_{\pm 7.3}$	$\mathbf{83.9}_{\pm 4.6}$	$86.3_{\pm 1.2}$	$12.6_{\pm 7.8}$	$9.1_{\pm 0.0}$	$9.1_{\pm 0.0}$	$\mathbf{79.0}_{\pm 3.8}$	$82.2_{\pm 5.6}$	$\mathbf{88.5}_{\pm 1.7}$
	4 layers	$70.8_{\pm 3.7}$	$\mathbf{78.2}_{\pm 9.0}$	$83.5_{\pm 4.9}$	$9.1_{\pm 0.0}$	$9.1_{\pm 0.0}$	$9.1_{\pm 0.0}$	$\mathbf{71.2}_{\pm 7.9}$	$71.1_{\pm 3.8}$	$\mathbf{84.0}_{\pm 5.9}$

dimensionalities, sequence lengths, and target structures (per-step for ECG, per-sequence for SHD and DVS Gesture), so they probe the surrogate across regimes; encoding and preprocessing are detailed in Appendix C.

Across the grid, the cdx-t* adjoint attains the best final test accuracy in most cells, the conventional adjoint is second, and Sparse SG is last (Figure 4, Table 3). The cdx-t* margin over the conventional adjoint is smallest at two layers—on SHD and DVS Gesture the two lie within each other's seed spread—and clearest at three and four layers, where it exceeds that spread in wider cells; on DVS Gesture the two stay close at every depth. Final accuracy generally rises with hidden width at fixed depth for both adjoints on all three datasets. Sparse SG is weakest throughout, carries the largest seed-to-seed variance at two layers, and hits a near-chance floor at three and four layers on every dataset. Training accuracy preserves this ranking, reported in Appendix C.1 (Figure 11, Table 9), where the train–test gap grows with width and depth. A steeper applied surrogate (Gaussian k=8) reproduces the same ranking and trends (subsection C.2).

3.3 Discussion and Limitations

The effective surrogate gradient is a lens on how credit is transformed as it propagates through a spiking network—set jointly by the applied surrogate and the complete LIF adjoint, not by either alone (subsection 3.1). Through it, the simulation analysis shows that neither lever suffices on its own: common smooth, symmetric applied surrogates cannot correct the effective shape, and intervening on the adjoint is itself delicate, the hard-cut Sparse SG recovering the balanced credit only in part. Taking the effective surrogate as the design point instead—a surrogate backward model shaped around it, not around the applied gate or the adjoint alone—our cdx-t* construction restores that balance; under empirical scrutiny (subsection 3.2) it matches or improves on the conventional adjoint out of sample while training markedly more stably and faster (subsection C.1).

This contrast in training dynamics traces back to the shape of the effective surrogate. Under the conventional adjoint, leakage opens a downward jump ΔG at threshold, pushes the sub-threshold mode below θ, and leaves credit sub-heavy ($\rho = m_-/m_+ > 1$)—crediting a neuron as if it acted continuously in time, an RNN-like profile, rather than only at its spike, and concentrating gradient variance at the discontinuity. We hypothesise that this misaligned, high-variance credit is what unsettles training and biases it toward silent states; by removing the jump and re-centring credit on θ, cdx-t* yields the steadier, faster-converging gradients its training curves show (subsection C.1), even where final out-of-sample accuracy is comparable.

We frame these results as a demonstration of effective-surrogate redesign rather than a complete characterisation. We therefore note the limitations of the study. The evaluation covers neuromorphic benchmark classification; its behaviour in broader practical settings is left out of scope. Moreover, this work has established the effective surrogate analytically and assessed it empirically, yet its sensitivity to the broad range of avaialble shape hyperparameters—in both single neurons and deeper stacks—remains to be fully mapped. The effective-surrogate view still leaves a wide design space of applied-surrogate and adjoint combinations, and of objectives beyond the credit fidelity we target here—computational efficiency, gradient stability, and retained learning capacity—open to explore.

4 Conclusion

We showed that the surrogate gradient applied at the spike nonlinearity is not the effective gradient a spiking network receives: propagated through the complete LIF adjoint under BPTT, a smooth, symmetric applied surrogate becomes discontinuous at threshold and asymmetric around it, inflating gradient variance and biasing credit toward silent states. Because this effective surrogate is set by the applied surrogate and the adjoint together, we designed for it directly—our cdx-t* construction—rather than reshaping the applied surrogate alone, closing the discontinuity and restoring balanced credit. This yields markedly more stable training that matches or improves out-of-sample accuracy, establishing design for the effective surrogate as a principled route to better SNN optimisation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing methods, such as writing and refactoring code and text, and additionally to produce figures and tables and to draft and edit parts of the manuscript for readability. The authors have designed the theory, analysis and empirical scenarios investigated in this study. AI has aided our efforts in analysis and the testing of our hypotheses through practical implementation and visualisation. We have reviewed all AI-assisted work and take responsibility for the final content of this work.

ETHICS STATEMENT

This work is methodological: it studies the surrogate gradients used to train spiking neural networks, in simulation and on three publicly available neuromorphic benchmarks. It involves no human-subject experiments and collects no new data; the benchmark datasets were collected and publicly released by their original authors, cited in Appendix C. We are not aware of ethical concerns specific to this work beyond those common to machine learning research.

REPRODUCIBILITY STATEMENT

The effective surrogate gradient and the adjoints we compare are defined in section 2, with the derivations and their modelling assumptions in Appendix A, and the simulation procedure, applied surrogate functions, and their calibration in Appendix B. The supplementary material contains all of the code required to faithfully reproduce this analysis of effective surrogate gradients, including implementations of every applied surrogate function and LIF adjoint investigated in this work. For the empirical analysis, Appendix C details the benchmark datasets, available from their original authors cited there, and, together with subsection 3.2, the full training configuration.

REFERENCES

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*, pp. 7243–7252, 2017.

Da Chang, Deliang Wang, and Xiao Yang. Ikun: Initialization to keep snn training and generalization great with surrogate-stable variance. *arXiv preprint arXiv:2411.18250*, 2024.

Tianyi Chen, Sihan Chen, Xiaoyi Qu, Dan Zhao, Ruomei Yan, Jongwoo Ko, Luming Liang, and Pashmina Cameron. Stableqat: Stable quantization-aware training at ultra-low bitwidths. *arXiv preprint arXiv:2601.19320*, 2026.

Brian Chmiel, Liad Ben-Uri, Moran Shkolnik, Elad Hoffer, Ron Banner, and Daniel Soudry. Neural gradients are near-lognormal: improved quantized and sparse training, October 2020. URL http://arxiv.org/abs/2006.08173.

Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. *Advances in neural information processing systems*, 28, 2015.

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.

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

Luca Herranz-Celotti and Jean Rouat. Stabilizing spiking neuron training. *arXiv preprint arXiv:2202.00282*, 2022.

Kungjui Hou, Kunlun Wu, and Yongcheng Zhou. Adaptive and lightweight surrogate gradients: enhancing training efficiency of spiking neural networks. *Frontiers in Neuroscience*, 20:1795946, 2026.

Yechan Kang, Yongjin Kweon, Mingyeong Seo, Sohee Park, Yeonguk Jeon, Jongkil Park, Hyun Jae Jang, Jaewook Kim, YeonJoo Jeong, Suyoun Lee, et al. A2sg: Adaptive and asymmetric surrogate gradients for training deep spiking neural networks. *arXiv preprint arXiv:2606.11236*, 2026.

Pablo Laguna, Roger G Mark, A Goldberg, and George B Moody. A database for evaluation of algorithms for measurement of qt and other waveform intervals in the ecg. In *Computers in cardiology 1997*, pp. 673–676. IEEE, 1997.

Gregor Lenz, Kenneth Chaney, Sumit Bam Shrestha, Omar Oubari, Serge Picaud, and Guido Zarrella. Tonic: event-based datasets and transformations., July 2021. URL `https://doi.org/10.5281/zenodo.5079802`. Documentation available under https://tonic.readthedocs.io.

Balázs Mészáros, James C Knight, and Thomas Nowotny. Efficient event-based delay learning in spiking neural networks. *Nature Communications*, 16(1):10422, 2025.

Emre O Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

Nicolas Perez-Nieves and Dan Goodman. Sparse spiking gradient descent. *Advances in Neural Information Processing Systems*, 34:11795–11808, 2021.

Xinyu Shi and Zhaofei Yu. Training deep spiking neural networks without normalization. In *Proceedings of the 43rd International Conference on Machine Learning*. PMLR, 2026. URL `https://icml.cc/virtual/2026/poster/62569`. Accepted at ICML 2026. To appear in PMLR.

Korneel Van den Berghe, Stein Stroobants, Vijay Janapa Reddi, and Guido De Croon. Adaptive surrogate gradients for sequential reinforcement learning in spiking neural networks. *Advances in Neural Information Processing Systems*, 38:147904–147926, 2026.

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.

Friedemann Zenke and Surya Ganguli. SuperSpike: Supervised Learning in Multilayer Spiking Neural Networks. *Neural Computation*, 30(6):1514–1541, June 2018. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco_a_01086. URL `https://direct.mit.edu/neco/article/30/6/1514-1541/8378`.

Friedemann Zenke and Tim P Vogels. The remarkable robustness of surrogate gradient learning for instilling complex function in spiking neural networks. *Neural computation*, 33(4):899–925, 2021.

A Appendix - Methodology

These derivations support section 2 in its own order: the effective-surrogate operator and the shape and dispersion it produces (subsection A.1), then the sparse and cdx-t* redesigns (subsection A.2).

A.1 The effective-surrogate operator

The single-neuron analysis extends verbatim to a layer: the scalar leak β becomes the diagonal recurrent map $\beta\mathbf{I}$, each neuron carrying its own membrane, spike, and reset, and the spatial cotangent λ^{s} is emitted upstream through the inter-layer weights; the base neuron of subsection 2.1 is a strict special case. The *applied* surrogate is the adjoint's space output read as a pure function of the current state (unit downstream cotangent, zeroed future); for the recipe of Equation 4 this is exactly the gate G of Equation 3.

Conditioning the unrolled adjoint (Equation 5) on $U_t = u$ gives the linear, forward-determined applied-to-effective operator

$$G_{\text{eff}}(u) = \bar{\delta}\, \mathbb{E}\left[\sum_{s=t}^{s^*(t)} \beta^{\,s-t}\, G(U_s) \,\middle|\, U_t = u \right] \equiv (\mathcal{G}_\beta\, G)\,(u), \tag{14}$$

fixed by the forward dynamics alone—the β-discounted occupation of the membrane process killed at θ. Two anchors bound it: above threshold the sum is one term, $G_{\text{eff}}^{\text{conv}} = \bar{\delta}\, G$; at $\beta = 0$, $\mathcal{G}_0 = \bar{\delta}\, \mathbb{I}$ recovers the gate everywhere, so any recurrence-free binary unit (a binary-activation MLP still requires a surrogate) has $G_{\text{eff}} \propto G$. For $\beta > 0$ the accumulated terms add a sub-threshold pedestal, with four consequences for the shape of G_{eff}: (1) a downward jump at θ (present at θ^-, truncated at θ^+); (2) a mode shifted below θ; (3) sub-heavy cross-threshold mass; (4) at large leak, a change from a bump to a monotone ramp. subsection 2.3 names the first three; Figure 5 illustrates them on a generic effective surrogate.

Beyond the conditional mean, the same conditioning yields the conditional variance $\mathrm{Var}[\lambda_t^{\mathrm{s}} \mid U_t = u]$ and a per-state signal-to-noise ratio: two adjoints can share G_{eff} yet differ in dispersion, and the threshold discontinuity is where the gradient variance concentrates (a heavy-tailed downstream cotangent inflates this dispersion without moving the mean; Appendix B).

A.2 Adjoint redesign: sparse and cdx-t*

Over one inter-spike interval ending in a spike at $T = s^*(t)$ (phase $\tau = T - t$), the silent steps accumulate the conventional sub-threshold tail $B^{\text{conv}}(u) = \bar{\delta} \sum_{s>t} \beta^{\,s-t} \mathbb{E}[G(U_s) \mid U_t = u] > 0$, which the conventional, sparse, and cdx-t* adjoints admit in full, truncate to a band, and replace by a spike-time carrier, respectively.

Conventional adjoint. Evaluating the operator of Equation 14 in each domain gives

$$G_{\text{eff}}^{\text{conv}}(u) = \begin{cases} G(u)\,\bar{\delta}, & u \geq \theta, \\ G(u)\,\bar{\delta} + B^{\text{conv}}(u), & u < \theta. \end{cases} \tag{15}$$

Above threshold the neuron fires at the current step ($s^*(t) = t$) and the sum collapses to the lone spike term $G(u)\bar{\delta}$; below it the local gate $G(u)\bar{\delta}$ is lifted by the strictly positive pedestal $B^{\text{conv}}(u)$ carried by the silent steps that still precede the spike. Two shape departures follow at once: at θ only the sub side carries the pedestal, a downward jump $\Delta G = B^{\text{conv}}(\theta^-) > 0$; and because $B^{\text{conv}}(u)$ falls toward θ (fewer silent steps remain) while $G(u)$ rises, their sum peaks strictly below threshold—the detached sub-mode $u^\star_- < \theta$. Sparse and cdx-t* are read against this baseline.

Sparse surrogate. With the band-truncated tail $B^{\text{sparse}}(u) = \bar{\delta} \sum_{s>t} \beta^{\,s-t} \mathbb{E}[\widehat{G}(U_s) \mid U_t = u] \leq B^{\text{conv}}(u)$, the effective surrogate is

$$G_{\text{eff}}^{\text{sparse}}(u) = \begin{cases} G(u)\,\bar{\delta}, & u \geq \theta, \\ G(u)\,\bar{\delta} + B^{\text{sparse}}(u), & \vartheta_{\text{sub}} \leq u < \theta, \\ B^{\text{sparse}}(u), & u < \vartheta_{\text{sub}}. \end{cases} \tag{16}$$

Two discontinuities follow: at θ the sub side is higher by $B^{\text{sparse}}(\theta^-) > 0$—same sign as conventional but reduced, since the tail now accumulates only $[\vartheta_{\text{sub}}, \theta)$—and a downward jump of size $G(\vartheta_{\text{sub}})\bar{\delta}$ at ϑ_{sub} where the local gate switches off. Cutting the deep sub-threshold contributions concentrates credit near θ and weakens, but does not remove, the sub-mode.

cdx-t*. Substituting $\widetilde{G}_T = \max(G(U_T), G(U_{T-1}))$ with $G(U_{T-1}) = G(u)$ at the penultimate step ($\tau = 1$),

$$\hat{G}_{\text{eff}}^{t^\star}(u) = \begin{cases} \bar{\delta}_T\, \mathbb{E}\big[\max(G(u), G(U_{T-1}))\big], & u \geq \theta, \\ (1-\beta)\, G(u)\, \bar{\delta}_{T-1} + \beta\, \mathbb{E}\big[\max(G(U_T), G(u))\big]\, \bar{\delta}_T, & u < \theta. \end{cases} \tag{17}$$

Both sub-threshold terms increase in u toward θ, so the mode sits at θ^-; under overshoot ($G(u) \to 0$) the super side is *floored* at $\bar{\delta}_T\, \mathbb{E}[G(U_{T-1})] > 0$, not emptied. Under stationary δ and penultimate-step dominance of the θ^- region (a modelling assumption), using $\max(1, \cdot) = 1$,

$$\hat{G}_{\text{eff}}^{t^\star}(\theta^+) = \bar{\delta}, \qquad \hat{G}_{\text{eff}}^{t^\star}(\theta^-) = (1-\beta)\bar{\delta} + \beta\bar{\delta} = \bar{\delta} \;\Rightarrow\; \Delta G = 0 : \tag{18}$$

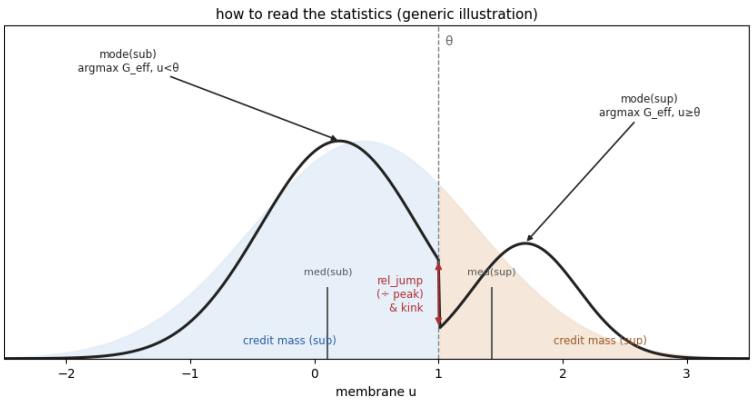


Figure 5: How to read the shape scores on a generic effective surrogate $G_{\text{eff}}(u)$ (placeholder): the sub-/super-threshold modes $u_-^\star, u_+^\star$, the threshold jump ΔG (and slope kink) at θ, and the occupancy-weighted mass m_-, m_+ on each side. A discrepancy-free effective surrogate would be peaked at θ and symmetric about it; the conventional one is neither.

a unit-gain hand-off—as the penultimate membrane approaches θ the gate selects $G(U_{T-1}) \to 1$, passing the spike credit at full strength and lifting the sub side to meet the super side.

Eligibility exponent η. The gate weights the penultimate step by $w_{T-1} = \beta^{1-\eta}$ (Equation 10); the value read is always the winner's full surrogate, so η tunes only the selection. At $\eta = 1$ ($w_{T-1} = 1$) the steps compete at equal strength—the plain $\max(G(U_T), G(U_{T-1}))$—which at $\beta \to 0$ still compares the penultimate step though $\partial U_T / \partial I_{T-1} = \beta = 0$, aliasing credit; at $\eta = 0$ ($w_{T-1} = \beta$) the weight is linear in the leak, diluting the near-threshold selection as the membrane moves off θ. We fix $\eta = \frac{1}{2}$ ($w_{T-1} = \sqrt{\beta}$): it vanishes as $\beta \to 0$ yet preserves the selection at moderate leak. Read through time, the comparison $\beta^{1-\eta} G(U_{T-1}) \gtrless G(U_T)$ is a selection margin—in log-surrogate terms it offsets the squared threshold-distances $r_{T-1}^2 - r_T^2$ ($r_s = |U_s - \theta|$) by $\propto (1-\eta)\ln(1/\beta)$, so $1-\eta$ sets how far back the crossing selection $T^\star$ may reach. At $\eta = 0$ the weight is exactly the raw leak β, the same scaling the sparse window $[\vartheta_{\text{sub}}, \theta)$ applies; the exponent decouples the selection from that leak, reading the gate at full strength while preserving the $\beta \to 0$ collapse—which is why the sparse baseline keeps $\vartheta_{\text{sub}} = \beta\theta$ and does not inherit it (a shared $\beta^{1-\eta}\theta$ would collapse its band onto θ at $\eta = 1$). The two share the leak β as their common anchor, not the exponent.

B Appendix - Simulation analysis

This appendix has two parts. **Part 1** expands the simulation analysis of subsection 3.1—how the effective surrogate is sampled and estimated, the applied surrogate families, and how their slopes are calibrated (symmetric and asymmetric). **Part 2** (subsection B.4) uses the asymmetric variants to ask whether a static asymmetry can remove the stylistic discrepancies.

B.1 Sampling and estimation

Input distribution and stationary membrane potential. The input current is Gaussian, $I_{t,i} \sim \mathcal{N}(\mu_I, \sigma_I^2)$, and its mean and variance (μ_I, σ_I^2) are calibrated—for the given leak β and threshold θ—so that the membrane settles to a stationary Gaussian law, $U_{t,i} \sim \mathcal{N}(\mu_u, \sigma_u^2)$. The target moments μ_u and σ_u^2 carry *no time index*: the same law holds at every step t, which is the sense of stationarity used here. We obtain (μ_I, σ_I^2) through grid search, refined iteratively until convergence. Trajectories are simulated over all steps, while the fit is assessed on the interior of each trajectory—the first and last 50 steps dropped—so that warm-up and truncation do not distort the moment estimates. The fit is verified rather than assumed: at the accepted (μ_I, σ_I^2) the empirical membrane mean and variance match the target (μ_u, σ_u^2), and the spike rate matches a Bernoulli law of mean p and variance $p(1-p)$—checking both the membrane's stationarity, through its first two moments, and its intended symmetry, through the spike statistics. The slice studied sets $\mu_u = 1$,

$\sigma_u^2 = 1$ at $\theta = 1$, giving $p = 0.5$; the sampled input-current and membrane densities are shown in Figure 6 (top, middle).

Ensemble and binning. The forward recurrence (Equation 2) is run over 8 seeds $\times$ 1024 trajectories $\times\, T = 500$ steps, dropping the first 50 (warm-up) and last 50 (truncation), so the retained window is $\mathcal{W} = \{t_0{+}1, \ldots, T{-}t_0\}$ with $t_0 = 50$. The sampled downstream cotangent is injected and the adjoint (Equation 4) is run back; λ^{s} is binned into 256 bins $\{\mathcal{B}_b\}$ over $u \in [-4, 4]$ (centres u_b), and the effective surrogate and occupancy are the per-bin conditional mean and frequency (dropping bins with fewer than 200 samples):

$$\widehat{G}_{\mathrm{eff}}(u_b) = \frac{\sum_{i,\, t\in\mathcal{W}} \lambda^{\mathrm{s}}_{t,i}\, \mathbf{1}[\, U_{t,i} \in \mathcal{B}_b\,]}{\sum_{i,\, t\in\mathcal{W}} \mathbf{1}[\, U_{t,i} \in \mathcal{B}_b\,]}, \qquad \widehat{\pi}(u_b) = \frac{1}{N\,|\mathcal{W}|} \sum_{i,\, t\in\mathcal{W}} \mathbf{1}[\, U_{t,i} \in \mathcal{B}_b\,], \tag{19}$$

with i the sample and t the time index of Equation 13 and N the number of trajectories. The one-sided threshold limits $G_{\mathrm{eff}}(\theta^{\mp})$ use local linear fits extrapolated to θ (skipping the bins straddling θ).

Gradient distribution. The downstream cotangent is drawn i.i.d. per step and trajectory from the log-normal law $\delta = \exp(\xi)$, $\xi \sim \mathcal{N}(0, 1)$: base-domain median 1, mean $e^{1/2} \approx 1.649$, variance $(e-1)e \approx 4.67$, with a heavy right tail (Figure 6, bottom). sSampling synthetic gradients from lognormal distributions for this analysis aligns with the empirical observation that gradients are lognormally distributed(Chmiel et al., 2020). The lognormal distribution selected here, results in relatively large variance which makes it suitable to analyze the rigidity and robustness of our approach even under extreme conditions. Because the spike-detached adjoint is affine in δ, the effective surrogate is the unit-gradient one scaled by the gradient's base-domain mean (≈ 1.649): the distribution's shape leaves the shape of G_{eff} unchanged and rescales only its magnitude (measured ratio $G_{\mathrm{eff}}(\text{log-normal})/G_{\mathrm{eff}}(\text{unit}) = 1.647 \pm 0.02$ across bins). The heavy tail inflates the conditional *variance* of the credit, not the conditional mean reported in Table 1.

B.2 Surrogate definitions

Every surrogate is written in the shifted coordinate $z = u - \theta$ and peaks at 1 on the threshold, with a slope k (a half-width for the boxcar); the asymmetric variants (subsection B.3) apply a slope per side, k_- for $z < 0$ and k_+ for $z \geq 0$:

$$\begin{aligned} \text{Gaussian: } & G(u) = \exp\!\big(-\tfrac{1}{2}kz^2\big), & \text{fast-sigmoid: } & G(u) = \big(1 + k|z|\big)^{-2}, \\ \text{sech}^2\text{: } & G(u) = 1 - \tanh^2(kz), & \text{triangular: } & G(u) = \max\!\big(0,\, 1 - k|z|\big), \\ \text{boxcar: } & G(u) = \mathbf{1}\big[\,|z| \leq k\,\big]. \end{aligned} \tag{20}$$

All share the same adjoint (Equation 4); only G changes, so any difference in the effective surrogate is a pure shape effect.

B.3 Slope calibration

A raw slope is not comparable across shapes, so we fix each by matching the *expected surrogate overlap* (participation) under a reference membrane $U \sim \mathcal{N}(\theta, 1)$. Calibrating a slope *per side* of the threshold, k_- (sub) and k_+ (super), the participation splits as

$$I = I_{\mathrm{sub}}(k_-) + I_{\mathrm{super}}(k_+), \quad I_{\mathrm{super}} = \rho\, I_{\mathrm{ref}}, \quad I_{\mathrm{sub}} = (1-\rho)\, I_{\mathrm{ref}}, \quad I_{\mathrm{ref}} = \tfrac{1}{\sqrt{5}} \approx 0.447, \tag{21}$$

with I_{ref} the Gaussian reference at $k{=}4$ ($I(\text{Gaussian}, k) = 1/\sqrt{k+1}$ in closed form; the other slopes by a root find on a fine quadrature). The super-share ρ selects the case. *Symmetric* ($\rho = \frac{1}{2}$) forces $k_- = k_+$ and a 50:50 split—the main study. The *asymmetric* variants take $\rho = \frac{2}{3}$ (a 1:2 sub:super split): *rotate* solves the two sides independently, holding the total I_{ref} fixed (a narrow sub flank, wide super flank, peak on θ); *translate* instead freezes the symmetric width k and rigidly shifts the shape by an offset (in membrane-potential units) until the super-share is $\frac{2}{3}$, so the total participation drifts a few percent below I_{ref} (the cost of shifting off the mode). The reference membrane is a calibration assumption only, unrelated to the tested cells. The resulting slopes are collected in Table 4.

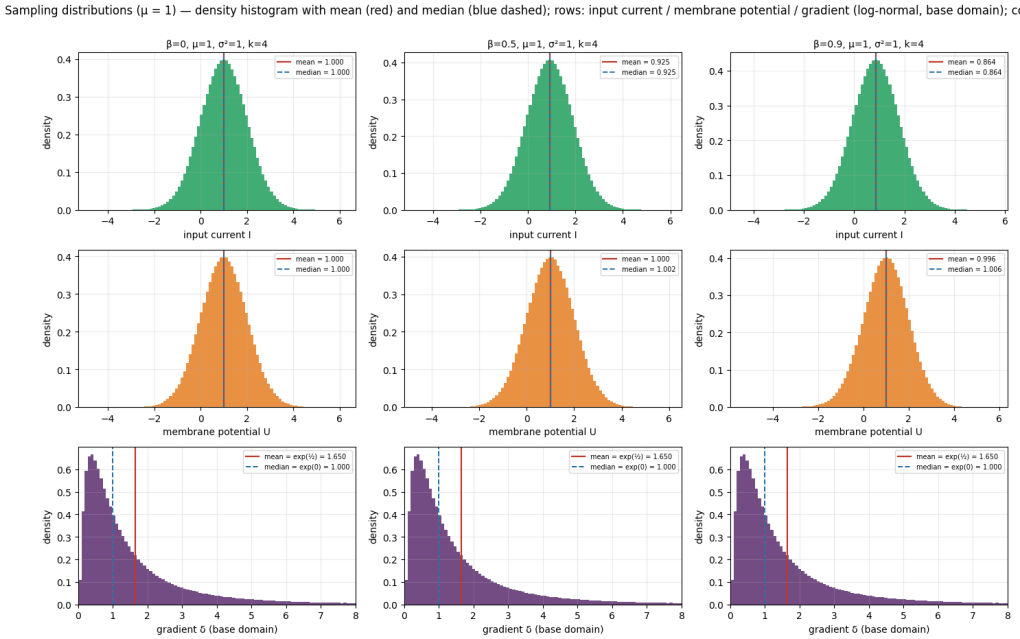
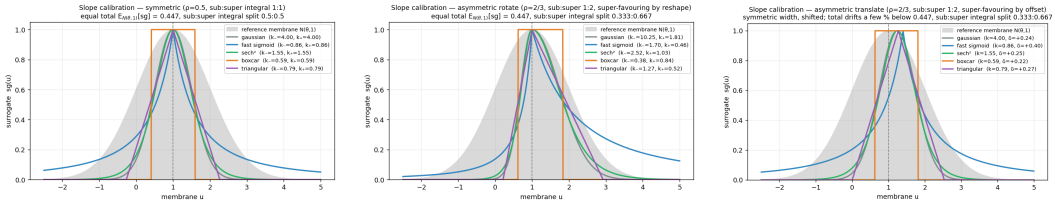


Figure 6: Sampled distributions at μ_u=1. *Rows:* input current I (top; Gaussian, mean = median = μ_I), membrane potential U (middle; stationary Gaussian $\approx\mathcal{N}(1,1)$), downstream gradient δ (bottom; base domain, right-skewed log-normal, mean $e^{1/2}\approx1.65$ and median 1); *columns:* $\beta \in \{0, 0.5, 0.9\}$. Each panel is a density histogram with mean (red) and median (blue, dashed). The gradient is β-independent; calibration holds $U\approx\mathcal{N}(1,1)$ across the slice (subsection B.1).



(a) Symmetric ($\rho = 0.5$) (b) Rotate (1:2) (c) Translate (1:2)

Figure 7: Slope calibration for the three sub-studies; in each, the five surrogates enclose the same expected overlap $\mathbb{E}_{\mathcal{N}(\theta,1)}[\mathrm{sg}]$=0.447 (Gaussian k=4 reference). *(a)* Symmetric: equal $k_- = k_+$, 50:50 split. *(b)* Rotate: reshaped flanks (narrow sub, wide super, peak on θ), 1:2 split. *(c)* Translate: the symmetric shape shifted into super, same 1:2 split.

Table 4: Calibrated slopes for the five surrogates in each sub-study, solving equal participation ($\mathbb{E}_{\mathcal{N}(\theta,1)}[\mathrm{sg}] = I_{\mathrm{ref}} = 0.447$, the Gaussian k=4 reference; subsection B.3). *Symmetric* reports the single slope k ($k_- = k_+$); *rotate* the per-side slopes k_-/k_+; *translate* the symmetric width k and the shift offset (in membrane-potential units).

	Symmetric ($k_- = k_+$)	Rotate (k_-/k_+)	Translate (k, offset)
Gaussian	4.00	10.25/1.81	4.00, +0.24
Fast-sigmoid	0.86	1.70/0.46	0.86, +0.40
Sech^2	1.55	2.52/1.03	1.55, +0.25
Boxcar	0.59	0.38/0.84	0.59, +0.22
Triangular	0.79	1.27/0.52	0.79, +0.27

B.4 PART 2 — ASYMMETRIC VARIANTS: ROTATION AND TRANSLATION

The symmetric effective surrogate is sub-heavy: the leak biases delivered credit below threshold (mass 71:29 at $\beta = 0.9$, Table 1) and the value jump ΔG favours the sub side. A natural fix is to bias the *applied* surrogate toward the super-threshold side—the side whose credit the discontinuity drops—and measure how much survives the adjoint. Under the identical procedure and input distributions of Part 1, we use the two super-favouring calibrations of subsection B.3—*rotation* (reshaped flanks) and *translation* (a rigid shift), both at the 1:2 split. The applied and effective curves are shown in Figure 8 and Figure 9, and scored in Table 5 and Table 6.

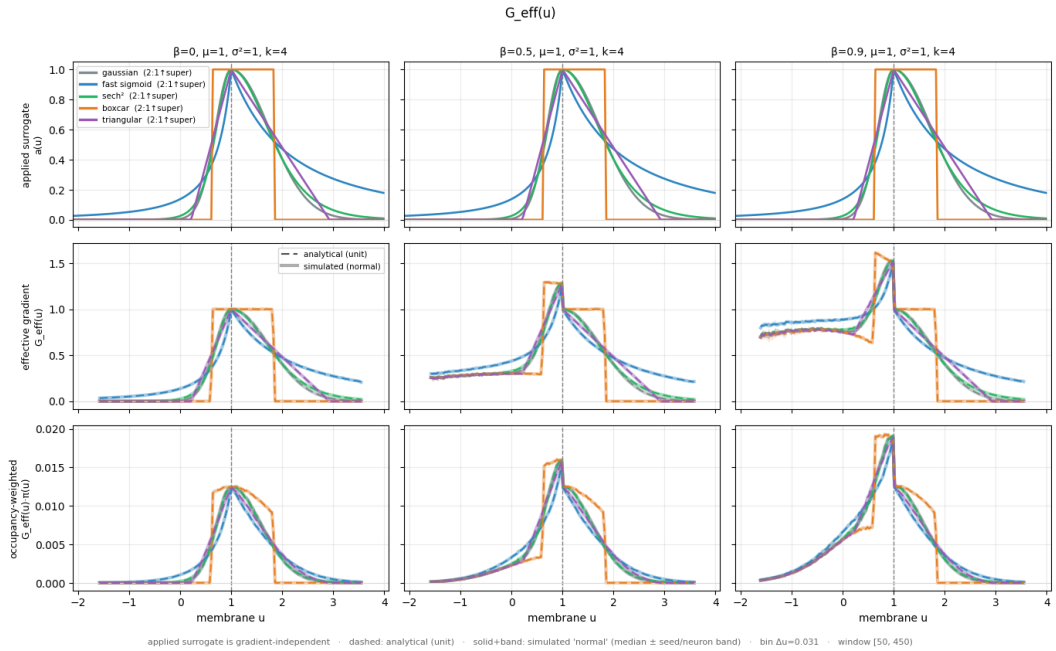


Figure 8: *Asymmetric rotate*, five calibrated surrogates overlaid. *Rows:* applied (top), effective G_{eff} (middle), occupancy-weighted $G_{\text{eff}}\pi$ (bottom); *columns:* $\beta \in \{0, 0.5, 0.9\}$. Flanks reshaped to a 1:2 sub:super split, peak pinned on θ (subsection B.3). Curves at unit gradient; the log-normal effective surrogate has the same shape, rescaled by $\approx$1.649.

Table 5: *Asymmetric rotate* shape scores (log-normal gradient): flanks reshaped to a 1:2 sub:super split, peak pinned on θ; slopes per side, k_-/k_+. Columns as in Table 1. The super-mode stays on θ; the leak erodes the super-bias (33:67 $\rightarrow$ 50:50 $\rightarrow$ 62:38) and ΔG stays positive and grows.

	$\beta = 0$			$\beta = 0.5$			$\beta = 0.9$		
Surrogate (k_-/k_+)	ΔG [−]	$u^\star_-/u^\star_+$ [−]	mass [%]	ΔG [−]	$u^\star_-/u^\star_+$ [−]	mass [%]	ΔG [−]	$u^\star_-/u^\star_+$ [−]	mass [%]
Gaussian (10.25/1.81)	+0.15	−0.02/+0.02	33/67	+0.59	−0.02/+0.08	50/50	+0.99	−0.02/+0.05	62/38
Fast-sigmoid (1.70/0.46)	−0.18	−0.02/+0.02	33/67	+0.28	−0.02/+0.02	50/50	+0.73	−0.02/+0.02	62/38
Sech^2 (2.52/1.03)	+0.13	−0.02/+0.02	33/67	+0.57	−0.02/+0.08	50/50	+0.98	−0.02/+0.05	62/38
Boxcar (0.38/0.84)	+0.24	−0.17/+0.77	33/67	+0.68	−0.27/+0.67	50/50	+1.05	−0.33/+0.39	61/39
Triangular (1.27/0.52)	+0.01	−0.02/+0.02	33/67	+0.46	−0.02/+0.02	50/50	+0.88	−0.02/+0.05	62/38

Both mechanisms deliver the same super-threshold mass gain: the effective split moves from 71:29 (symmetric) to $\approx$ 62:38 at $\beta = 0.9$, whether the applied bias is produced by reshaping or by shifting. They differ only in where the credit sits—rotation keeps the effective super-mode on θ, while translation carries it into the super domain (Table 5, Table 6). In both, the leak erodes the applied super-bias: the effective super-share falls $67\% \rightarrow 50\% \rightarrow 38\%$ as β runs $0 \rightarrow 0.5 \rightarrow 0.9$, so a fixed asymmetry over-corrects at low leak and under-corrects at high leak. The threshold jump follows the same pattern and changes sign across the slice—for the translated Gaussian ΔG runs $-0.07 \rightarrow +0.36 \rightarrow +0.73$, and for the rotated fast-sigmoid $-0.18 \rightarrow +0.28 \rightarrow +0.73$—so a bias tuned to cancel the jump at one β misses it at the others.

A static asymmetry can thus cancel the discontinuity at a single operating point, but not at every cell at once, and no closed form sets the per-cell parameters. Even a single-point correction does not persist: a fixed surrogate holds the leak constant, so it cannot adapt during training, and the membrane distribution shifts as training proceeds, moving the operating point away from the one the asymmetry was calibrated for. This appears to call for a correction that tracks the dynamics—but even granting the per-cell parameters in closed form, the choice is not thereby settled; it becomes a trade-off between bias and variance. A correction re-fitted to the moving distribution follows the discontinuity, yet it redefines the credit-assignment target from step to step and so adds variance to the optimization; a correction fixed in advance keeps that target stable but grows mismatched as the

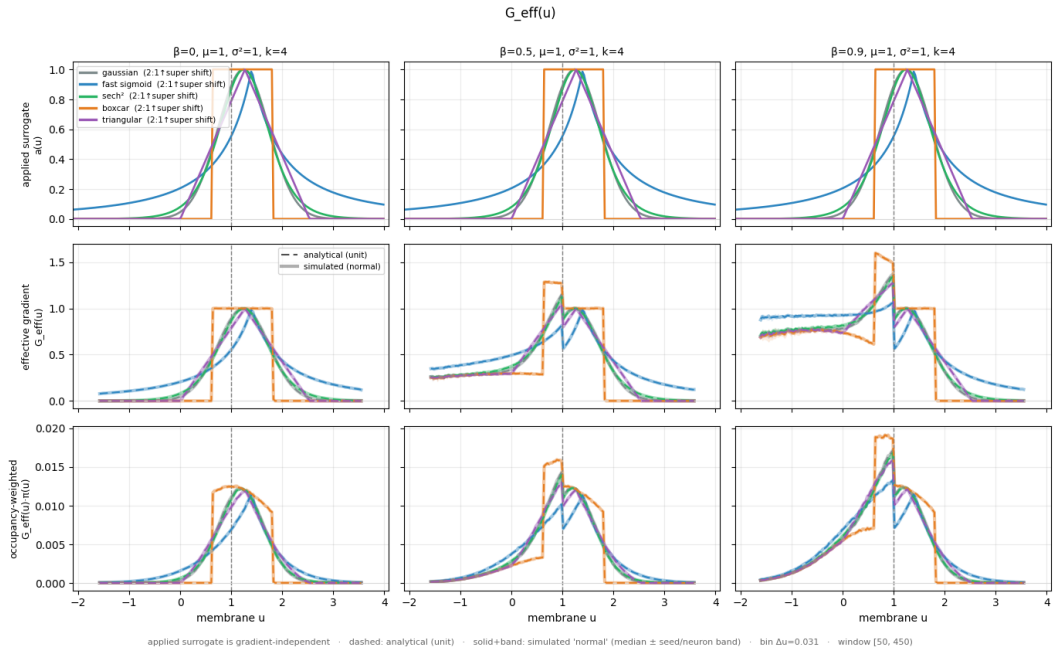


Figure 9: *Asymmetric translate*, five calibrated surrogates overlaid. *Rows:* applied (top), effective G_{eff} (middle), occupancy-weighted $G_{\text{eff}}\pi$ (bottom); *columns:* $\beta \in \{0, 0.5, 0.9\}$. Same 1:2 split reached by rigidly shifting the symmetric shape into super (subsection B.3); the effective super-mode is carried off θ. Curves at unit gradient; the log-normal effective surrogate has the same shape, rescaled by $\approx$1.649.

Table 6: *Asymmetric translate* shape scores (log-normal gradient): the symmetric shape (width k) rigidly shifted by an offset (membrane-potential units, peak at θ + offset) to the same 1:2 split. Columns as in Table 1. Unlike rotate, the super-mode is carried into the super domain ($u_+^\star > 0$); the mass split matches rotate (33:67 $\to$ 50:50 $\to$ 62:38), but ΔG flips sign with the leak (super-favoured at β=0, sub-favoured at β=0.9).

	$\beta = 0$			$\beta = 0.5$			$\beta = 0.9$		
Surrogate (k, offset)	$\Delta G\,[-]$	$u_-^\star/u_+^\star\,[-]$	mass [%]	$\Delta G\,[-]$	$u_-^\star/u_+^\star\,[-]$	mass [%]	$\Delta G\,[-]$	$u_-^\star/u_+^\star\,[-]$	mass [%]
Gaussian (4.00, +0.24)	−0.07	−0.02/+0.27	33/67	+0.36	−0.02/+0.30	50/50	+0.73	−0.02/+0.27	62/38
Fast-sigmoid (0.86, +0.40)	+0.04	−0.02/+0.39	33/67	+0.48	−0.02/+0.39	50/50	+0.89	−0.02/+0.39	62/38
Sech^2 (1.55, +0.25)	−0.11	−0.02/+0.27	33/67	+0.32	−0.02/+0.30	50/50	+0.70	−0.02/+0.27	62/38
Boxcar (0.59, +0.22)	+0.32	−0.17/+0.77	33/67	+0.74	−0.27/+0.67	50/50	+1.11	−0.33/+0.39	61/39
Triangular (0.79, +0.27)	−0.10	−0.02/+0.27	33/67	+0.33	−0.02/+0.30	50/50	+0.71	−0.02/+0.27	62/38

distribution drifts, and so imposes bias. Which term dominates is case-dependent, but under typical conditions the stable, higher-bias correction is generally preferable, with a continually shifting one warranted only where the mismatch is large enough to offset its added variance. It is this trade-off—not the availability of an asymmetry—that a principled correction must navigate, and it motivates the surrogate design introduced in the methodology.

B.5 Depth propagation of cdx-t*

The main comparison is read at a single node; because cdx-t* is the only adjoint that removes the single-layer discrepancy (Figure 3), we test whether the correction survives across depth. We stack $L = 4$ single-neuron layers, indexed $l = 1, \ldots, L$, each the hard-reset LIF of Equation 2 driven to the same stationary membrane $U^l \sim \mathcal{N}(\mu_u, \sigma_u^2)$. Layer 1 is driven by the input-current distribution of subsection B.1; the downstream gradient is injected at layer L (the log-normal δ of Figure 6); the LIF temporal adjoint carries the cotangent back through the stack, and the effective surrogate is read at every layer.

Forward: input current across layers. A layer's input current is its predecessor's spike train plus a Gaussian innovation,

$$I^l_{t,i} \;=\; S^{l-1}_{t,i} + \varepsilon^l_{t,i}, \qquad \varepsilon^l_{t,i} \sim \mathcal{N}\big(\mu_I - p,\; \sigma_I^2 - p(1-p)\big). \tag{22}$$

Because the upstream spike is Bernoulli, $S^{\,l-1} \sim \mathrm{Ber}(p)$ with mean p and variance $p(1-p)$, the innovation supplies exactly the mean and variance it does not, so I^l carries the calibrated Gaussian moments (μ_I, σ_I^2) of subsection B.1 and the membrane stays stationary at every depth.

Backward: straight-through across layers. To propagate the cotangent through the additive input current, the map $S^{\,l-1} \to I^l$ is treated as a straight-through estimator: the cotangent arriving at I^l is passed to the upstream spike $S^{\,l-1}$ with unit gain and then transformed by that layer's adjoint. The per-layer adjoints compose, and the gradient injected at layer L reaches the effective surrogate at every layer.

We show cdx-t* alone across depth, as it is the only adjoint that resolves the single-layer discrepancy. Its effective surrogate is depth-stable: the mode stays on θ and the threshold jump remains closed at every layer (Table 7), and the occupancy-weighted curves nearly overlie across depth—the optimal-gate correction does not re-open upstream.

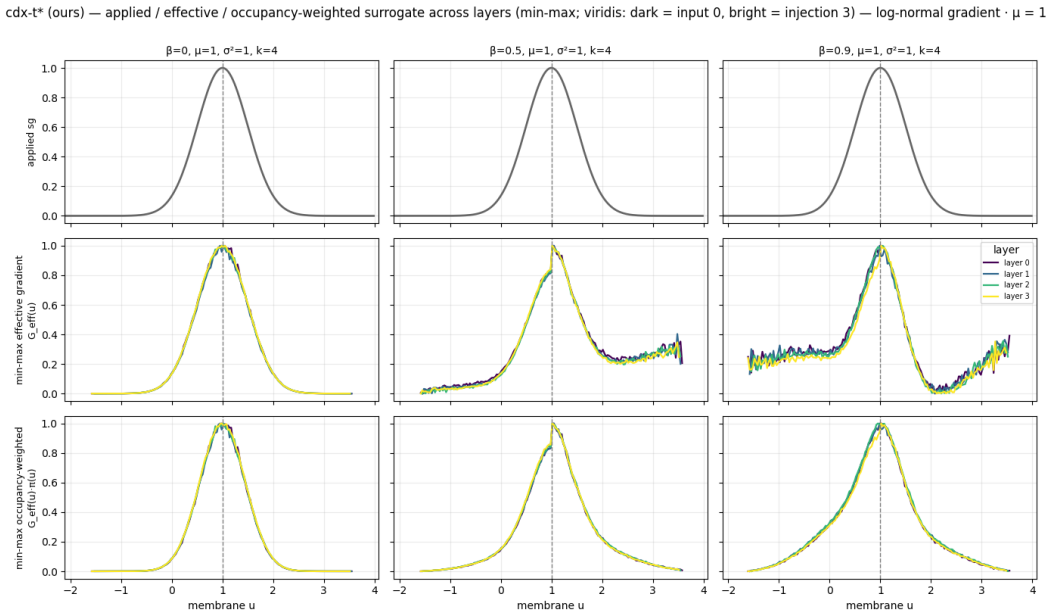


Figure 10: Depth propagation of cdx-t* (Gaussian surrogate, k=4; min–max normalised). *Rows:* applied (top), effective G_{eff} (middle), occupancy-weighted $G_{\text{eff}}\pi$ (bottom); *columns:* $\beta \in \{0, 0.5, 0.9\}$; overlaid: the four stacked layers (dark = input layer 1, bright = injection layer L). The corrected shape is depth-stable.

Table 7: cdx-t* effective-surrogate shape scores across depth (Gaussian surrogate, k=4). Rows are the four stacked layers (layer 1 input $\to$ layer L=4 injection). Columns as in Table 1, with ΔG read on the min–max-normalised curves of Figure 10. The input-layer row matches the cdx-t* entry of Table 2; the jump stays closed and the mass split is invariant across depth.

	$\beta = 0$			$\beta = 0.5$			$\beta = 0.9$		
Layer	$\Delta G\,[-]$	$u_-^\star/u_+^\star\,[-]$	mass [%]	$\Delta G\,[-]$	$u_-^\star/u_+^\star\,[-]$	mass [%]	$\Delta G\,[-]$	$u_-^\star/u_+^\star\,[-]$	mass [%]
Layer 1 (input)	0.02	$-0.02/+0.05$	50/50	0.16	$-0.02/+0.05$	44/56	0.02	$-0.02/+0.02$	52/48
Layer 2	0.02	$-0.05/+0.05$	50/50	0.17	$-0.05/+0.05$	44/56	0.02	$-0.02/+0.08$	52/48
Layer 3	0.01	$-0.02/+0.02$	50/50	0.17	$-0.05/+0.02$	44/56	0.02	$-0.05/+0.05$	52/48
Layer 4 (injection)	0.00	$-0.02/+0.05$	50/50	0.15	$-0.02/+0.02$	44/56	0.09	$-0.02/+0.02$	51/49

C Appendix - Empirical evaluation

The experiments of subsection 3.2 use three neuromorphic classification benchmarks. Each is used as its encoding produces it: we do not augment the data, normalise sequence lengths, or standardise individual channels. The three are summarised in Table 8 and detailed below; they vary in modality, input dimensionality, sequence length, and label granularity (whole sequence or single step).

ECG. Electrocardiogram beat classification (Laguna et al., 1997). Inputs are the level-crossing spike trains released by Yin et al. (2021): two leads $\times$ two polarities (4 channels) at 1 ms resolution, with a fixed length of 1300 steps. Unlike the other two tasks, ECG has step-wise labels, and the beat class may change mid-sequence; the task is 6-way, split $618/141$ with 8 training batches.

SHD (Spiking Heidelberg Digits). Spoken digits recorded as event streams from a model cochlea (Cramer et al., 2020). Events are binned into 5 ms frames (Lenz et al., 2021) with one entry per cochlear channel, giving 700 inputs per step. Sequences span 46–273 steps; the task is 20-way with one label per sequence, split $8156/2264$ with 16 training batches.

DVS Gesture. Hand gestures recorded with an event camera (Amir et al., 2017). We halve the spatial resolution and accumulate events into 50 ms frames (Lenz et al., 2021), keeping ON and

Table 8: Benchmark overview: input dimension, sequence length (fixed, or the observed $[T_{\min}, T_{\max}]$ over both splits), classes, and train/test sizes. *Binning* is the time window over which events are accumulated into one frame.

Dataset	Input dim	Steps	Classes	Train/test	Binning
ECG	4	1300 (fixed)	6	$618/141$	1 ms (level-crossing)
SHD	700	46–273	20	$8156/2264$	5 ms
DVS Gesture	8192	34–369	11	$1077/264$	50 ms

OFF events separate, so each step holds $64 \times 64 \times 2$ values (8192 inputs). The only denoising is the refractory filter Tonic applies by default. Sequences span 34–369 steps; the task is 11-way with one label per sequence, split $1077/264$ with 16 training batches.

C.1 Training-set curves and tallies

We report the training-set counterparts of the out-of-sample results of subsection 3.2: Figure 11 plots the training accuracy over epochs and Table 9 the final training tallies, under the same grid, adjoints, and fixed applied surrogate. The adjoint ranking matches the test split; the train–test gap widens with hidden width and depth, most visibly on DVS Gesture, where the conventional and cdx-t* adjoints reach near-perfect training accuracy while their test accuracy plateaus lower.

C.2 Steeper applied surrogate (k=8)

As a robustness check on the applied-surrogate slope, we repeat the study of subsection 3.2 with a steeper Gaussian surrogate (k=8), holding everything else fixed. The picture matches k=4: the cdx-t* adjoint attains the best out-of-sample accuracy in most cells, the conventional adjoint is second, and Sparse SG is weakest and collapses to a near-chance floor at the deeper stacks; accuracy again rises with hidden width. Out-of-sample curves and tallies are in Figure 12 and Table 10; the training-set counterparts in Figure 13 and Table 11.

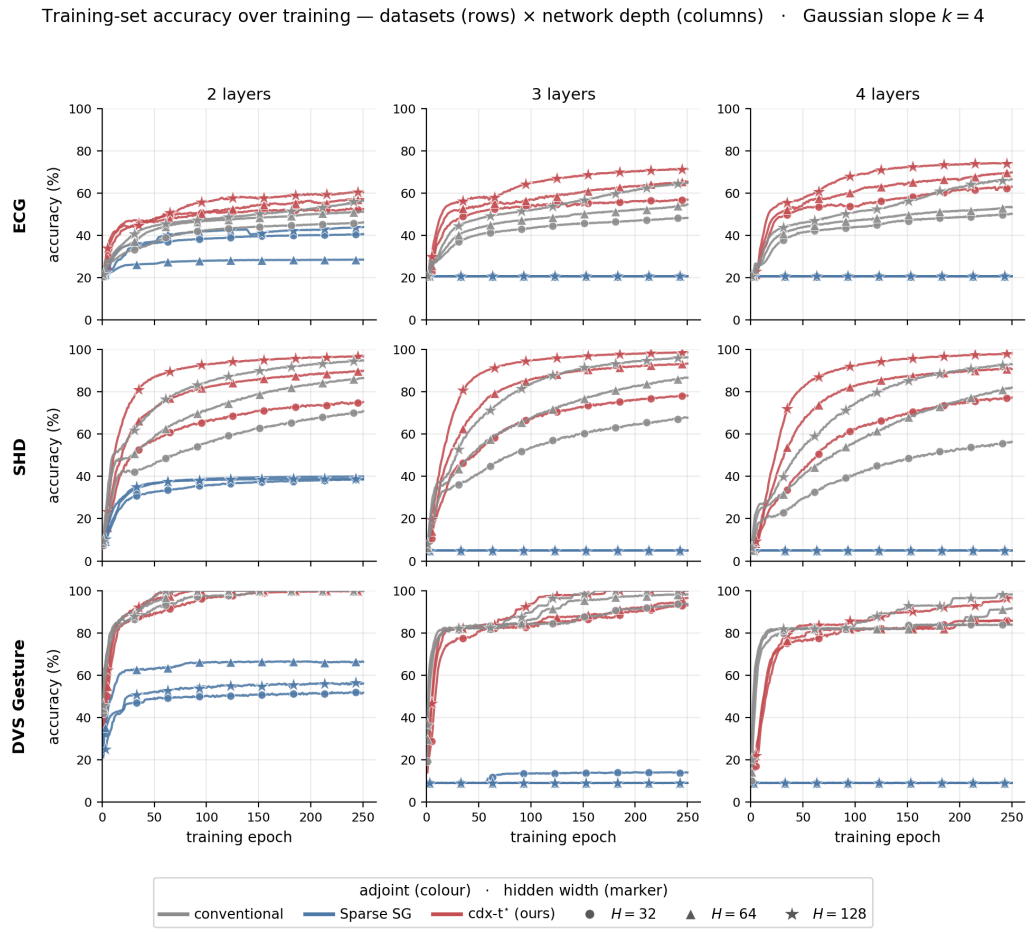


Figure 11: Training-set accuracy by epoch, laid out as Figure 4. *Rows:* datasets (ECG, SHD, DVS Gesture); *columns:* depth (2–4 LIF layers). Overlaid per panel: conventional (grey), Sparse SG (blue), cdx-t* (ours, red), at widths $H \in \{32, 64, 128\}$ (circle/triangle/star); lines are the seed mean. Tallies in Table 9; test curves in Figure 4.

Table 9: Final training-set accuracy (%) at epoch 250, mean$_{\pm\text{std}}$ over five seeds, across datasets, network depth, and hidden width, for the conventional, Sparse SG, and cdx-t* (ours) adjoints at a fixed applied surrogate (Gaussian slope k=4). The best adjoint per (dataset, depth, width) is in bold. Curves in Figure 11.

		Conventional			Sparse SG			cdx-t* (ours)		
Dataset	**Depth**	32	64	128	32	64	128	32	64	128
ECG	2 layers	$46.1_{\pm 2.5}$	$51.2_{\pm 0.9}$	$56.4_{\pm 2.9}$	$40.6_{\pm 18.5}$	$28.5_{\pm 9.9}$	$43.9_{\pm 21.3}$	$\mathbf{52.5}_{\pm 4.4}$	$\mathbf{57.2}_{\pm 5.0}$	$\mathbf{60.4}_{\pm 4.7}$
	3 layers	$48.3_{\pm 2.6}$	$54.6_{\pm 2.9}$	$64.7_{\pm 3.1}$	$20.7_{\pm 0.0}$	$20.7_{\pm 0.0}$	$20.7_{\pm 0.0}$	$\mathbf{56.9}_{\pm 1.3}$	$\mathbf{65.4}_{\pm 1.5}$	$\mathbf{71.5}_{\pm 2.1}$
	4 layers	$50.2_{\pm 1.5}$	$53.3_{\pm 1.8}$	$66.6_{\pm 1.5}$	$20.7_{\pm 0.0}$	$20.7_{\pm 0.0}$	$20.7_{\pm 0.0}$	$\mathbf{63.0}_{\pm 4.8}$	$\mathbf{69.8}_{\pm 4.9}$	$\mathbf{74.4}_{\pm 1.8}$
SHD	2 layers	$70.9_{\pm 2.1}$	$86.5_{\pm 0.8}$	$94.8_{\pm 0.4}$	$38.6_{\pm 10.6}$	$40.0_{\pm 6.0}$	$38.9_{\pm 13.7}$	$\mathbf{75.2}_{\pm 2.8}$	$\mathbf{89.9}_{\pm 1.1}$	$\mathbf{96.8}_{\pm 0.3}$
	3 layers	$67.8_{\pm 1.8}$	$86.7_{\pm 0.8}$	$96.3_{\pm 0.6}$	$5.0_{\pm 0.0}$	$5.0_{\pm 0.0}$	$5.0_{\pm 0.0}$	$\mathbf{78.2}_{\pm 2.5}$	$\mathbf{93.4}_{\pm 0.7}$	$\mathbf{98.7}_{\pm 0.4}$
	4 layers	$56.4_{\pm 8.2}$	$82.0_{\pm 3.4}$	$93.0_{\pm 1.5}$	$5.0_{\pm 0.0}$	$5.0_{\pm 0.0}$	$5.0_{\pm 0.0}$	$\mathbf{77.3}_{\pm 1.2}$	$\mathbf{90.7}_{\pm 2.6}$	$\mathbf{98.2}_{\pm 0.7}$
DVS Gesture	2 layers	$\mathbf{100.0}_{\pm 0.0}$	$\mathbf{100.0}_{\pm 0.0}$	$\mathbf{100.0}_{\pm 0.0}$	$51.7_{\pm 6.5}$	$66.3_{\pm 28.2}$	$56.0_{\pm 13.4}$	$99.7_{\pm 0.6}$	$100.0_{\pm 0.0}$	$100.0_{\pm 0.0}$
	3 layers	$\mathbf{93.6}_{\pm 6.1}$	$\mathbf{98.2}_{\pm 4.0}$	$\mathbf{100.0}_{\pm 0.0}$	$13.9_{\pm 11.0}$	$9.0_{\pm 0.1}$	$9.0_{\pm 0.1}$	$93.0_{\pm 4.0}$	$96.5_{\pm 4.8}$	$100.0_{\pm 0.0}$
	4 layers	$83.8_{\pm 4.0}$	$\mathbf{91.6}_{\pm 9.1}$	$\mathbf{98.2}_{\pm 4.1}$	$9.0_{\pm 0.1}$	$9.0_{\pm 0.1}$	$9.0_{\pm 0.1}$	$\mathbf{85.7}_{\pm 8.0}$	$85.7_{\pm 5.0}$	$96.4_{\pm 8.1}$

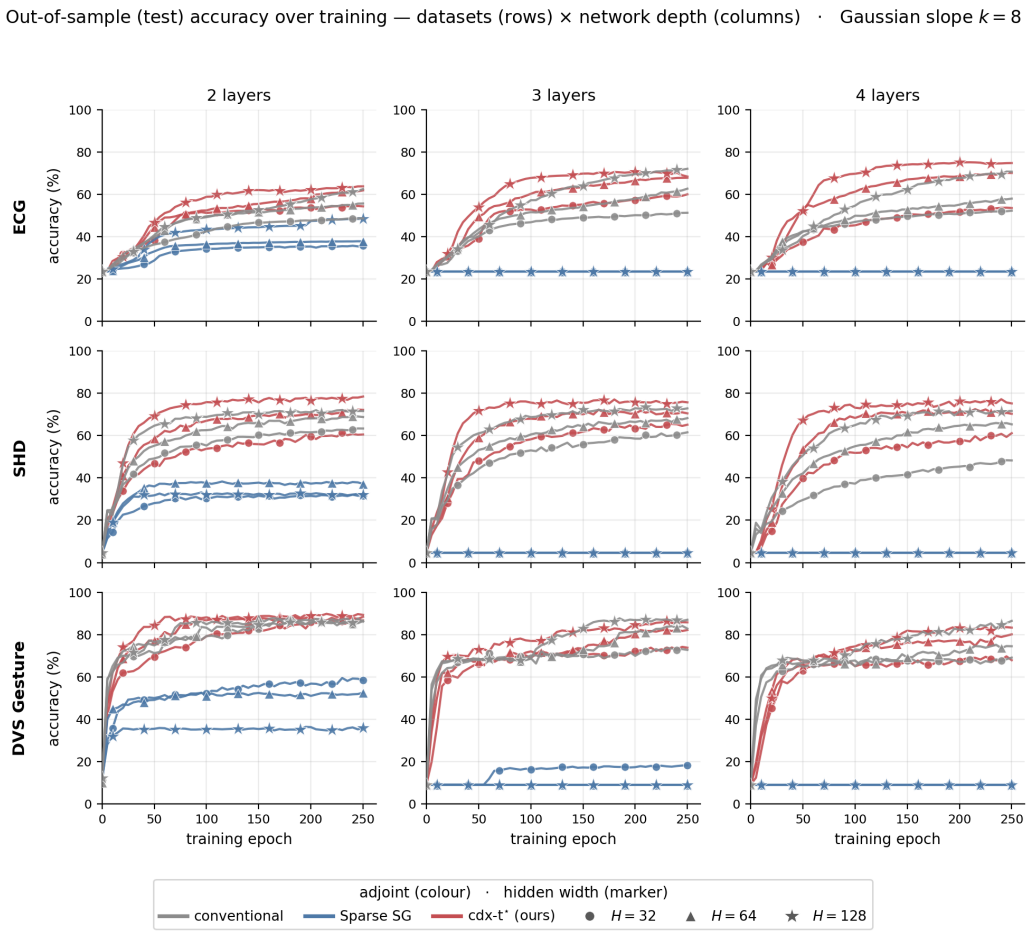


Figure 12: Out-of-sample (test) accuracy over training at applied-surrogate slope k=8, laid out as Figure 4. *Rows:* datasets (ECG, SHD, DVS Gesture); *columns:* depth (2–4 LIF layers). Overlaid per panel: conventional (grey), Sparse SG (blue), cdx-t* (ours, red), at widths $H \in \{32, 64, 128\}$ (circle/triangle/star); lines are the seed mean. Tallies in Table 10.

Table 10: Final out-of-sample (test) accuracy (%) at epoch 250, mean$_{\pm\text{std}}$ over five seeds, across datasets, network depth, and hidden width, for the conventional, Sparse SG, and cdx-t* (ours) adjoints at a fixed applied surrogate (Gaussian slope k=8). The best adjoint per (dataset, depth, width) is in bold. Curves in Figure 12.

		Conventional			Sparse SG			cdx-t* (ours)		
Dataset	**Depth**	32	64	128	32	64	128	32	64	128
ECG	2 layers	$48.6_{\pm1.6}$	$55.7_{\pm1.3}$	$62.5_{\pm3.7}$	$35.7_{\pm16.9}$	$37.8_{\pm16.8}$	$48.4_{\pm17.9}$	$\mathbf{54.3}_{\pm0.9}$	$\mathbf{61.9}_{\pm1.5}$	$\mathbf{63.8}_{\pm7.1}$
	3 layers	$51.3_{\pm3.2}$	$62.7_{\pm3.7}$	$\mathbf{72.1}_{\pm0.9}$	$23.6_{\pm0.0}$	$23.6_{\pm0.0}$	$23.6_{\pm0.0}$	$\mathbf{60.0}_{\pm3.4}$	$\mathbf{67.8}_{\pm2.1}$	$68.4_{\pm6.4}$
	4 layers	$52.3_{\pm4.5}$	$58.0_{\pm3.1}$	$70.7_{\pm2.8}$	$23.6_{\pm0.0}$	$23.6_{\pm0.0}$	$23.6_{\pm0.0}$	$\mathbf{53.5}_{\pm11.2}$	$\mathbf{70.1}_{\pm4.4}$	$\mathbf{74.8}_{\pm6.5}$
SHD	2 layers	$\mathbf{63.3}_{\pm2.8}$	$68.8_{\pm0.7}$	$71.6_{\pm1.0}$	$31.5_{\pm4.7}$	$37.2_{\pm6.6}$	$32.1_{\pm12.0}$	$60.5_{\pm2.1}$	$\mathbf{72.2}_{\pm1.9}$	$\mathbf{78.4}_{\pm1.7}$
	3 layers	$61.5_{\pm2.2}$	$68.2_{\pm1.1}$	$72.9_{\pm0.7}$	$4.5_{\pm0.0}$	$4.5_{\pm0.0}$	$4.5_{\pm0.0}$	$\mathbf{65.1}_{\pm1.3}$	$\mathbf{70.5}_{\pm1.5}$	$\mathbf{75.7}_{\pm1.6}$
	4 layers	$48.3_{\pm3.3}$	$65.3_{\pm1.2}$	$71.6_{\pm1.9}$	$4.5_{\pm0.0}$	$4.5_{\pm0.0}$	$4.5_{\pm0.0}$	$\mathbf{61.1}_{\pm4.8}$	$\mathbf{70.3}_{\pm1.0}$	$\mathbf{75.2}_{\pm2.6}$
DVS Gesture	2 layers	$86.4_{\pm1.1}$	$86.1_{\pm1.2}$	$86.4_{\pm5.2}$	$58.3_{\pm15.3}$	$52.5_{\pm14.2}$	$35.9_{\pm18.8}$	$\mathbf{86.7}_{\pm0.9}$	$\mathbf{88.4}_{\pm0.6}$	$\mathbf{89.3}_{\pm1.7}$
	3 layers	$72.7_{\pm5.1}$	$\mathbf{82.8}_{\pm4.7}$	$\mathbf{87.0}_{\pm2.0}$	$18.2_{\pm20.3}$	$9.1_{\pm0.0}$	$9.1_{\pm0.0}$	$\mathbf{73.8}_{\pm6.1}$	$82.1_{\pm4.8}$	$85.8_{\pm4.0}$
	4 layers	$\mathbf{69.1}_{\pm4.4}$	$74.5_{\pm7.4}$	$\mathbf{86.4}_{\pm1.5}$	$9.1_{\pm0.0}$	$9.1_{\pm0.0}$	$9.1_{\pm0.0}$	$67.9_{\pm4.7}$	$\mathbf{80.1}_{\pm4.2}$	$83.3_{\pm5.7}$

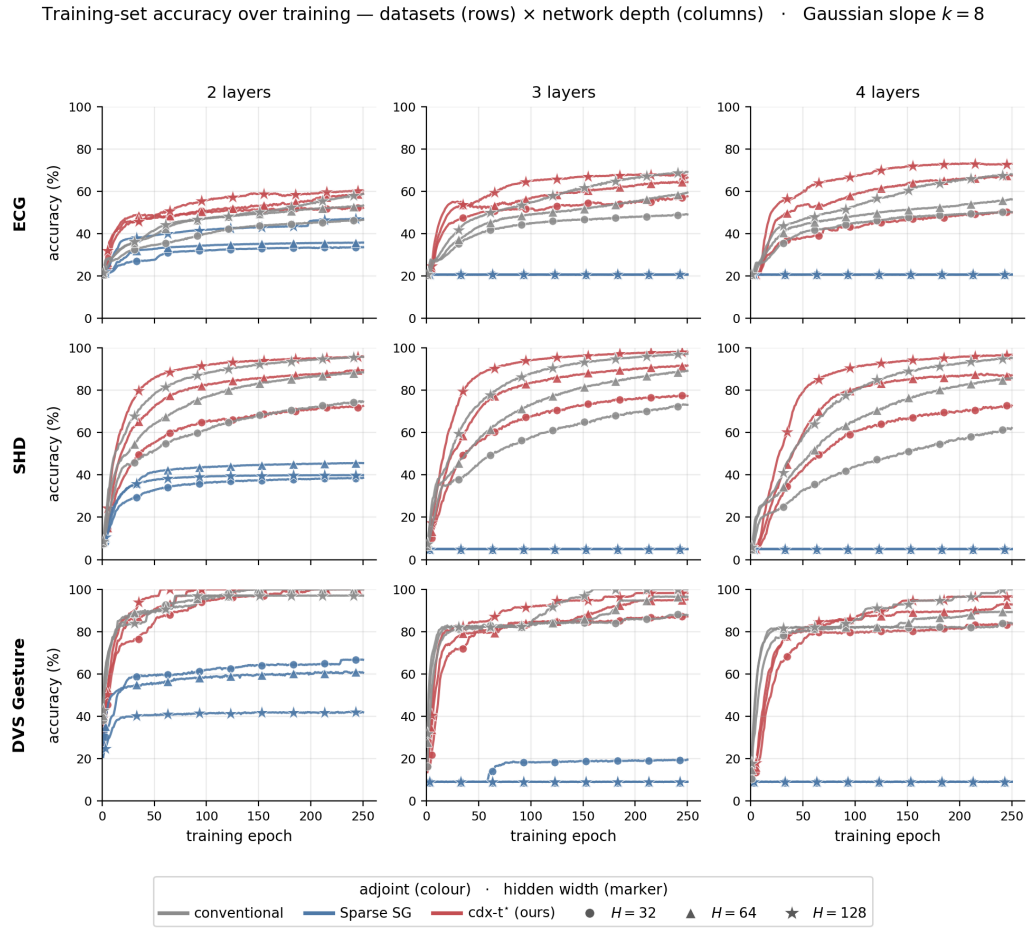


Figure 13: Training-set accuracy over training at applied-surrogate slope k=8, laid out as Figure 4. *Rows:* datasets (ECG, SHD, DVS Gesture); *columns:* depth (2–4 LIF layers); conventional (grey), Sparse SG (blue), cdx-t* (ours, red), at widths $H \in \{32, 64, 128\}$ (circle/triangle/star); lines are the seed mean. Tallies in Table 11; test curves in Figure 12.

Table 11: Final training-set accuracy (%) at epoch 250, mean$_{\pm\text{std}}$ over five seeds, across datasets, network depth, and hidden width, for the conventional, Sparse SG, and cdx-t* (ours) adjoints at a fixed applied surrogate (Gaussian slope k=8). The best adjoint per (dataset, depth, width) is in bold. Curves in Figure 13.

		Conventional			Sparse SG			cdx-t* (ours)		
Dataset	**Depth**	32	64	128	32	64	128	32	64	128
ECG	2 layers	$46.5_{\pm1.7}$	$53.3_{\pm1.2}$	$58.8_{\pm3.5}$	$33.6_{\pm18.0}$	$35.8_{\pm17.7}$	$47.1_{\pm18.8}$	$\mathbf{52.1}_{\pm1.2}$	$\mathbf{58.9}_{\pm1.4}$	$\mathbf{60.5}_{\pm7.4}$
	3 layers	$49.1_{\pm3.4}$	$59.4_{\pm4.3}$	$\mathbf{69.1}_{\pm1.5}$	$20.7_{\pm0.0}$	$20.7_{\pm0.0}$	$20.7_{\pm0.0}$	$\mathbf{57.6}_{\pm3.0}$	$\mathbf{64.4}_{\pm2.1}$	$66.5_{\pm6.6}$
	4 layers	$\mathbf{50.2}_{\pm4.4}$	$56.2_{\pm2.8}$	$68.0_{\pm2.0}$	$20.7_{\pm0.0}$	$20.7_{\pm0.0}$	$20.7_{\pm0.0}$	$50.0_{\pm11.1}$	$\mathbf{67.5}_{\pm4.5}$	$\mathbf{73.0}_{\pm6.7}$
SHD	2 layers	$\mathbf{74.6}_{\pm1.1}$	$88.5_{\pm1.0}$	$95.8_{\pm0.5}$	$38.6_{\pm7.2}$	$45.6_{\pm12.6}$	$39.9_{\pm15.6}$	$72.6_{\pm2.1}$	$\mathbf{89.4}_{\pm0.8}$	$\mathbf{95.8}_{\pm0.3}$
	3 layers	$73.1_{\pm2.1}$	$89.2_{\pm0.8}$	$97.2_{\pm0.3}$	$5.0_{\pm0.0}$	$5.0_{\pm0.0}$	$5.0_{\pm0.0}$	$\mathbf{77.3}_{\pm1.2}$	$\mathbf{91.7}_{\pm1.0}$	$\mathbf{98.2}_{\pm0.7}$
	4 layers	$62.2_{\pm3.1}$	$85.7_{\pm2.4}$	$95.2_{\pm0.8}$	$5.0_{\pm0.0}$	$5.0_{\pm0.0}$	$5.0_{\pm0.0}$	$\mathbf{72.8}_{\pm2.4}$	$\mathbf{87.1}_{\pm1.6}$	$\mathbf{96.7}_{\pm1.7}$
DVS Gesture	2 layers	$\mathbf{100.0}_{\pm0.0}$	$\mathbf{100.0}_{\pm0.0}$	$97.0_{\pm5.3}$	$66.7_{\pm19.0}$	$60.6_{\pm17.0}$	$41.9_{\pm22.8}$	$99.9_{\pm0.1}$	$100.0_{\pm0.0}$	$\mathbf{100.0}_{\pm0.0}$
	3 layers	$\mathbf{87.8}_{\pm4.5}$	$\mathbf{96.4}_{\pm4.9}$	$\mathbf{100.0}_{\pm0.0}$	$19.5_{\pm23.4}$	$9.0_{\pm0.1}$	$9.0_{\pm0.1}$	$87.1_{\pm5.0}$	$95.1_{\pm4.5}$	$98.2_{\pm4.0}$
	4 layers	$\mathbf{83.9}_{\pm4.3}$	$89.2_{\pm7.5}$	$\mathbf{100.0}_{\pm0.0}$	$9.0_{\pm0.1}$	$9.0_{\pm0.1}$	$9.0_{\pm0.1}$	$83.4_{\pm7.1}$	$\mathbf{92.8}_{\pm3.9}$	$96.5_{\pm4.8}$