ACTIVATION-AWARE INITIALIZATION FOR ODD-SIGMOID NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation functions and weight initialization are tightly coupled, especially in deep networks with bounded sigmoidal nonlinearities. Standard zero-mean Gaussian initializations, including Xavier, He, and edge-of-chaos variants, often become brittle in deep and narrow fully connected networks, where forward activations collapse and backpropagated gradients decay. We study this problem for a class of bounded, odd, monotone activations whose slopes decrease away from the origin, which we call odd-sigmoid activations. For an activation f, our initialization sets a diagonal gain using the critical slope $\omega = 1/f'(0)$ and adds Gaussian noise whose scale is calibrated by a terminal sign-flip surrogate. The scalar analysis gives a closed-form relation between the surrogate sign-flip target, depth, and noise scale, while an empirical depth schedule transfers this calibration to finite-width networks. Across deep fully connected networks on image benchmarks and PINN benchmarks, the proposed diagonal-plus-noise initialization is substantially less sensitive to depth, width, activation scaling, and learning-rate choice than Gaussian i.i.d. baselines. The results suggest that sign-flip calibration is a practical tool for training deep odd-sigmoid networks, while also highlighting the need for structured initialization beyond variance-only Gaussian schemes.

1 INTRODUCTION

Activation functions and weight initialization are tightly coupled: the scale and shape of the nonlinearity determine how forward signals and backpropagated gradients evolve with depth, and an initialization not tuned to the chosen activation can drive activations into saturation or cause gradients to vanish or explode (He et al., 2015; Lee et al., 2024). This coupling is especially pronounced for bounded sigmoidal activations in deep, narrow networks — a regime that remains practically dominant in physics-informed neural networks (PINNs) and scientific regression, where bounded outputs and smooth derivatives are required for PDE residuals (Raissi et al., 2019; Karniadakis et al., 2021; Rathore et al., 2024), as well as in FFN/MLP blocks embedded in larger models. We focus on this regime — not as a competitor to modern ReLU/GELU-based vision and language architectures (LeCun et al., 2015; He et al., 2016), but as the setting where odd-sigmoid activations are commonly used and where Xavier, He, and EOC — the standard initializers — become brittle in deep, narrow, and activation-rescaled regimes. The image-classification benchmarks used in this paper serve as controlled testbeds for initialization quality, while PINNs provide an application setting where odd-sigmoid activations are widely used.

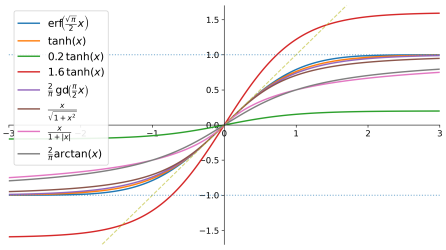


Figure 1: Odd–sigmoid activations.

Standard zero-mean Gaussian (Gaussian) i.i.d. initializations (Xavier (Glorot & Bengio, 2010), He (He et al., 2015), EOC (Hayou et al., 2019)) are derived under wide or infinite width assumptions and choose a single variance parameter to preserve signal propagation. In practice, we find that this design breaks down in deep and relatively narrow networks and under activation rescaling: small deviations in the variance can trigger saturation, or extreme learning rate sensitivity. Motivated by these observations, we propose an activation aware initialization for fully feedforward neural

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.

networks (FFNNs) with odd–sigmoid activations in the class $\mathcal{F}$ (Figure 1). For any $f \in \mathcal{F}$ with $\omega := 1/f'(0)$, we study the scalar dynamics of $x \mapsto f(ax)$ and choose the layerwise gains a from a distribution with mean ω so that trajectories neither collapse to zero nor saturate. This gain design is then implemented via a simple initial weight matrix whose effective gain statistics match the desired behavior. Unlike Gaussian i.i.d. schemes, whose trainability is highly sensitive to the exact variance, the proposed initialization keeps both forward activations and backpropagated gradients in a well conditioned range across depth (Raghu et al., 2017). A finite-width analysis of the layerwise Jacobian diagnostic χ_ℓ (Appendix A.3) separates the deterministic gain ω^2 from the noise contribution, and numerically χ_ℓ stays close to 1 over a wider band of noise scales than for Gaussian i.i.d. baselines (Figure 25).

We evaluate the proposed initialization in extensive experiments with odd–sigmoid activations, covering both the regime $\omega \approx 1$ and the more challenging regime $\omega \gg 1$ or $\omega \ll 1$. On standard image classification benchmarks, Gaussian initializations often fail or degrade sharply in deep, narrow, or rescaled-activation networks, whereas our scheme trains stably and is more data-efficient than these Gaussian baselines. We also study scaled and composite activations (e.g., $\alpha f(x)$, $f(\alpha x)$, and positive linear combinations) and observe that our initialization remains stable even for extreme scales (up to $\alpha \approx 10^9$). In PINN settings, appropriately scaled odd–sigmoid activations combined with our initialization consistently achieve lower losses than EOC-based baselines and do so without batch normalization (BN) (Ioffe & Szegedy, 2015), indicating that diagonal–plus–noise weights provide a practical alternative when Gaussian schemes become brittle. Our main contributions are as follows:

- We define the odd-sigmoid class $\mathcal{F}$, characterize its pitchfork fixed-point structure under $x \mapsto f(ax)$, and use this to motivate a diagonal-plus-noise initialization parameterized by $\omega := 1/f'(0)$ (Section 4.1, 4.2).
- We calibrate the noise scale by the terminal sign-flip rate: a closed-form scale in a scalar surrogate, and a depth schedule whose exponential form follows from the critical scaling $\sigma_z^2 \propto 1/L$ (Section 4.2).
- We show empirically that the proposed scheme trains where Xavier/He/EOC fail, including activation rescalings up to $10^9 \tanh(x)$, and lowers the relative L^2 error of deep PINNs on four PDE benchmarks by 14–$45\times$ over EOC (Section 5).

2 Related Work

Weight initialization and activation choice jointly determine signal propagation and trainability (Sutskever et al., 2013; Hayou et al., 2019). Zero-mean Gaussian i.i.d. schemes (Xavier (Glorot & Bengio, 2010), He (He et al., 2015), EOC (Poole et al., 2016; Schoenholz et al., 2016; Hayou et al., 2019)) fix a single variance from wide-network theory; for sigmoidal networks, nonzero weight means (Yilmaz & Poli, 2022) and mutual-information criteria (Qiao et al., 2016) have also been proposed. Beyond variance matching, prior work studies feature diversity (Blumenfeld et al., 2020), residual gating (Bachlechner et al., 2021), and maximal initial learning rates (Iyer et al., 2023). Orthogonal and isometry-based initializations (Saxe et al., 2014; Pennington et al., 2017; Xiao et al., 2018) control the spectrum of the input–output Jacobian and form a separate family; our comparison is restricted to Gaussian i.i.d. schemes, whose brittleness in deep, narrow, and activation-rescaled networks is the subject of this paper. Closest to our work, Lee et al. (2025) initialize $\tanh$ networks with an identity-plus-noise matrix and an empirically chosen noise scale. We extend this construction to the odd-sigmoid class $\mathcal{F}$ through the critical gain $\omega = 1/f'(0)$, which covers scaled and composite activations, and we replace the hand-tuned noise scale by a terminal sign-flip target with a closed-form surrogate solution and a depth schedule. PINNs embed PDE residuals in the loss (Raissi et al., 2019; Karniadakis et al., 2021) and are known to be hard to optimize (Rathore et al., 2024); adaptive slopes and learned activation combinations have been proposed as remedies (Jagtap et al., 2020; Wang et al., 2023). We keep the activation fixed and address the same difficulty through initialization alone.

3 Preliminaries

We introduce the notation used throughout the paper, review a simplified elementwise analysis of signal propagation (Lee et al., 2025), and recall the classical edge of chaos mean field theory (Schoenholz et al., 2016) and its associated Gaussian i.i.d. EOC initialization (Hayou et al., 2019).

Notation. We consider a feedforward neural network with L layers. The dataset comprises K pairs $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^K$, where $\boldsymbol{x}_i \in \mathbb{R}^{N_x}$ is the input and $\boldsymbol{y}_i \in \mathbb{R}^{N_y}$ is the corresponding target. For $\ell = 1, \ldots, L$, the layerwise update is $\boldsymbol{h}^\ell = \mathbf{W}^\ell \boldsymbol{x}^{\ell-1} + \mathbf{b}^\ell, \qquad \boldsymbol{x}^\ell = f(\boldsymbol{h}^\ell) \in \mathbb{R}^{N_\ell}$, where $\mathbf{W}^\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ is the weight matrix, $\mathbf{b}^\ell \in \mathbb{R}^{N_\ell}$ is the bias vector, and $f(\cdot)$ denotes an activation function. We write $\mathbf{W}^\ell = [w_{ij}^\ell]$.

Signal Propagation Analysis. Following the simplified framework of Lee et al. (2025), we analyze signal propagation in feedforward networks with an elementwise activation f. For convenience, all layers, including input and output, are assumed to have dimension n (i.e., $N_\ell = n$ for all ℓ). A key observation of this approach is that the usual matrix–vector multiplication $\mathbf{W}^\ell \boldsymbol{x}^{\ell-1}$ can be reformulated elementwise, such that each coordinate is expressed in terms of an effective self-scaling factor. Formally, for $\ell = 0, \ldots, L-1$ and $i = 1, \ldots, n$,

$$x_i^{\ell+1} = f\big(a_i^{\ell+1} x_i^\ell\big), \qquad a_i^{\ell+1} = w_{ii}^{\ell+1} + \sum_{j \neq i} \frac{w_{ij}^{\ell+1} x_j^\ell}{x_i^\ell}, \tag{1}$$

assuming that $\mathbf{b}^\ell = \mathbf{0}$. This elementwise decomposition highlights the role of $a_i^{\ell+1}$ as an effective gain that combines both diagonal and off-diagonal terms and provides a tractable basis for studying signal propagation in deep networks.

Edge of chaos. For fully connected networks with Gaussian i.i.d. weights, infinite-width mean-field theory describes signal and gradient propagation (Poole et al., 2016; Schoenholz et al., 2016). Weights and biases are sampled as $w_{ij}^\ell \sim \mathcal{N}(0, \sigma_w^2/N_{\ell-1})$ and $b_i^\ell \sim \mathcal{N}(0, \sigma_b^2)$, where σ_w and σ_b control the weight and bias scales, and $N_{\ell-1}$ is the preceding layer's width. At a fixed point q of the preactivation variance, the mean-field squared-gradient amplification factor is $\chi = \sigma_w^2 \mathbb{E}[f'(h)^2]$, where $h \sim \mathcal{N}(0, q)$. The edge of chaos (EOC), defined by $\chi = 1$, separates the ordered ($\chi < 1$) and chaotic ($\chi > 1$) regimes, associated with vanishing and exploding gradients, respectively. Suitable parameter choices on this boundary improve information propagation and accelerate training (Hayou et al., 2019). We use EOC initialization to denote Gaussian i.i.d. initialization with variance parameters (σ_w^2, σ_b^2) chosen on this curve.

4 METHODOLOGY

Section 4.1 defines the odd–sigmoid class, Section 4.2 proposes the initialization, and Section 4.3 contrasts it with Gaussian i.i.d. schemes; proofs are in Appendix A and additional simulations in Appendix B.

4.1 ODD-SIGMOID FUNCTION CLASS

In practical neural networks with sigmoidal activations, both forward signals and backpropagated gradients tend to saturate or vanish as the depth increases. Our goal in this work is to move beyond a single sigmoid and consider a broad class of odd–sigmoid activation functions, and to design a weight initialization that prevents such vanishing for any activation in this class. To this end, Section 4.1 introduces the odd–sigmoid function class and establishes its basic properties, which will be used in our subsequent analysis. Formal definitions of the activation functions used in this paper are provided in Appendix B.1.

Definition 4.1. *A function $f : \mathbb{R} \to \mathbb{R}$ is an **odd-sigmoid function** if it satisfies the following:*

(i) ***Regularity:*** $f \in C^1(\mathbb{R})$.

(ii) ***Odd symmetry:*** $f(-x) = -f(x)$ *for all* $x \in \mathbb{R}$.

(iii) ***Boundedness:*** $\sup_{x \in \mathbb{R}} |f(x)| < \infty$.

(iv) ***Strict monotonicity:*** $f'(x) > 0$ *for all* $x \in \mathbb{R}$.

(v) ***Slope decay:*** f' *is strictly decreasing on* $[0, \infty)$.

Denote the class of all odd-sigmoid functions as $\mathcal{F}$. In the following, we establish several basic properties of functions in $\mathcal{F}$.

Pitchfork Bifurcation. Recall that $x^* \in \mathbb{R}$ is a fixed point of $f : \mathbb{R} \to \mathbb{R}$ if $f(x^*) = x^*$. For $f \in \mathcal{F}$ define $\omega_f := 1/f'(0) > 0$ and, unless stated otherwise, write ω simply. Consider $\phi_a(x) := f(ax)$ for $a > 0$. We say that ϕ_a has a pitchfork bifurcation at $a = \omega$ if it has exactly one fixed point $\{0\}$ for $0 < a < \omega$ and exactly three fixed points $\{0, \pm\xi_a\}$ for $a > \omega$, with the nonzero points occurring as a symmetric pair. We show that this holds for every $f \in \mathcal{F}$.

Proposition 4.2. *Suppose $f \in \mathcal{F}$ with $\omega := 1/f'(0)$, and for a fixed $a > 0$ define $\phi_a(x) := f(ax)$. Then*

(i) If $0 < a \leq \omega$, then $\phi_a(x)$ has a unique fixed point $x^* = 0$.

(ii) If $a > \omega$, then $\phi_a(x)$ has three distinct fixed points: $x^* = -\xi_a,\ 0,\ \xi_a$ such that $\xi_a > 0$.

The following theorem establishes convergence properties of $x_{n+1} = f(ax_n)$ for all $x_0 > 0$.

Theorem 4.3. *Suppose $f \in \mathcal{F}$ with $\omega := 1/f'(0)$, and for a fixed $a > 0$ define $x_0 > 0,\ x_{n+1} = \phi_a(x_n),\ n = 0, 1, 2, \ldots$. Then the sequence $\{x_n\}$ converges for every $x_0 > 0$. Furthermore,*

(1) *if $0 < a \leq \omega$, then $x_n \to 0$ as $n \to \infty$.*

(2) *if $a > \omega$, then $x_n \to \xi_a$ as $n \to \infty$.*

According to Theorem 4.3, when $a > \omega$, for any initial value $x_0 > 0$ (resp. $x_0 < 0$), the sequence defined by $x_{n+1} = f(ax_n)$ converges to ξ_a (resp. $-\xi_a$) as $n \to \infty$. From a signal propagation viewpoint in FFNN (Equation 1), this implies that activations do not vanish as depth increases. See Figures 8 and 9 for the convergence of the iterates $x_{n+1} = \phi_a(x_n)$ as a function of a and the initial value x_0. Proposition A.2 and Corollary A.3 analyze the iteration with a coefficient sequence $\{a_m\}$ that may vary across steps. They show that excessively large or small a_m drives the dynamics into saturation, highlighting the importance of appropriately scaling $\{a_m\}$ when designing the initialization (Figure 10).

4.2 Proposed Weight Initialization Method

We build on the theoretical results in Section 4.1 to design an initialization scheme that keeps forward signals well dispersed and avoids saturation even in very deep networks. We then compare the proposed initialization with standard Gaussian i.i.d. schemes (e.g., Xavier, He, EOC) from both forward and backward pass perspectives, studying signal and gradient propagation through theoretical analysis and experiments.

Corollary 4.4. *Let $f_1, f_2 \in \mathcal{F}$ and let $c_1, c_2 \geq 0$ with $(c_1, c_2) \neq (0, 0)$. If $g = c_1 f_1 + c_2 f_2$, then $g \in \mathcal{F}$. Furthermore, it holds that $1/\omega_g = c_1/\omega_{f_1} + c_2/\omega_{f_2}$.*

Since $\mathcal{F}$ is closed under addition, more generally, any finite sum $\sum_{j=1}^{M} c_j f_j \in \mathcal{F}$ for all $c_j \geq 0$ with $(c_1, \ldots, c_M) \neq \mathbf{0}$ and $f_j \in \mathcal{F}$.

Proposed Weight Initialization. Consider $f \in \mathcal{F}$ with $\omega := 1/f'(0) > 0$ and a target depth L. We initialize each layer as $\mathbf{W}^\ell = \mathbf{D}^\ell + \mathbf{Z}^\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$, where $(\mathbf{D}^\ell)_{ij} = \omega$ if $i \equiv j \pmod{N_{\ell-1}}$ and 0 otherwise. The noise matrix $\mathbf{Z}^\ell$ is sampled with noise scale σ_z as $(\mathbf{Z}^\ell)_{ij} \sim \mathcal{N}\Big(0,\ \sigma_z^2/N_{\ell-1}\Big)$. Throughout, σ_z denotes this pre-normalization scale; the entrywise standard deviation is $\sigma_z/\sqrt{N_{\ell-1}}$.

The noise scale σ_z is set via a calibrated scalar surrogate. Let $p_{\text{real}} = 0.4$ be the target FFNN-level sign-flip rate at depth L (for all classification experiments; Fig. 6 and App. B.7 justify this choice as a robust operating point), and let $p_{\text{sur}}(L)$ denote the surrogate target driving the closed-form calibration

$$\sigma^*(p, L, \omega) \ = \ -\frac{\omega}{\Phi^{-1}\big(\frac{1-(1-2p)^{1/L}}{2}\big)},$$

where Φ is the standard normal CDF.

We do not derive $p_{\text{sur}}(L)$ from theory. Instead, we sweep depths at moderate L and read off, for each target π, the surrogate value that aligns the empirical FFNN-level sign-flip rate with π at depth L. Figure 2(b) shows that on semi-log axes the calibrated $p_{\text{sur}}(L, \pi)$ falls on a straight line for every π,

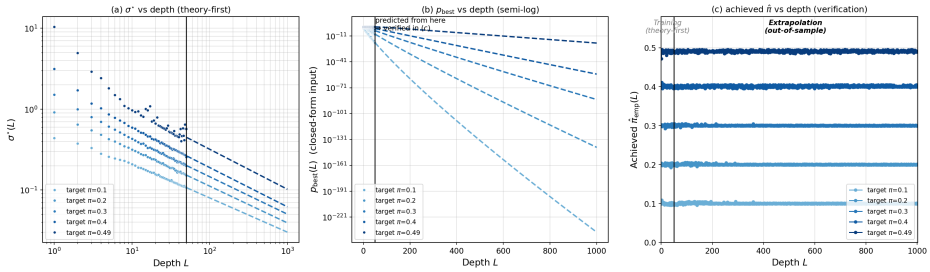


Figure 2: For five FFNN-level sign-flip targets $\pi \in \{0.1, 0.2, 0.3, 0.4, 0.49\}$ and depths L up to 10^3, we calibrate the surrogate target $p_{\mathrm{sur}}(L, \pi)$ as the value plugged into the closed-form $\sigma^*(p, L, \omega)$ that produces the target π empirically at depth L. **(a)** The resulting $\sigma^*(L)$ on log–log axes; each π yields a roughly straight line. **(b)** On semi-log axes the calibrated $p_{\mathrm{sur}}(L, \pi)$ is well approximated by an exponential $p_{\mathrm{sur}}(L, \pi) = c_\pi e^{-\kappa_\pi L}$. We fit $\log \sigma^*$ linearly in $\log L$ left of the vertical line at $L = 50$ and extrapolate beyond (dashed). **(c)** Verification: with the fitted schedule plugged into σ^*, the achieved sign-flip rate $\hat{\pi}_{\mathrm{emp}}(L)$ sits within a narrow band of its target π across all depths, including the out-of-sample region.

so we model it as $p_{\mathrm{sur}}(L, \pi) = c_\pi e^{-\kappa_\pi L}$ and fit (c_π, κ_π) by least squares. For our operating point $\pi = 0.4$, the fit on $L \in [10, 100]$ (Figure 13) yields $c \approx 2.05$ and $\kappa \approx 0.133$. We adopt the schedule $p(L) := \min\{\, p_{\mathrm{real}},\ c\, e^{-\kappa L} \,\}$, which equals p_{real} for $L \le 12$ and follows the fitted decay beyond, and set $\sigma_z := \sigma^*(p(L), L, \omega)$. The exponential form is what the critical scaling $L\sigma_z^2 \approx s$ predicts: under this scaling $p_{\mathrm{sur}}(L) \sim C\sqrt{L}\, e^{-\kappa L}$ with $\kappa = \omega^2/(2s)$ (Proposition B.1), and the fitted schedule indeed satisfies $L(\sigma_z/\omega)^2 \in [3.30, 3.69]$ for $20 \le L \le 10^3$ (Appendix B.4), i.e. $\sigma_z^2 \approx s\,\omega^2/L$ with a single empirical constant s. Figure 2(c) verifies that this single fitted schedule keeps the achieved $\hat{\pi}_{\mathrm{emp}}(L)$ within a narrow band of the target π across depths up to $L = 10^3$, far beyond the band used to fit it. We use this fitted schedule as the default classification calibration. For PINNs, we examine five target sign-flip rates and numerically calibrate the noise scale for each target. For the learning rate we use a band proportional to ω; with Adam (Kingma & Ba, 2014), $\eta \in [10^{-5}\omega,\ 10^{-3}\omega]$.

Derivation of the initialization. Using the elementwise formulation (1), the usual matrix–vector product $\mathbf{W}^\ell \mathbf{x}^{\ell-1}$ can be rewritten as $x_i^{\ell+1} = f\big(a_i^{\ell+1}\, x_i^\ell\big)$, $a_i^{\ell+1} = w_{ii}^{\ell+1} + \sum_{j \ne i} \frac{w_{ij}^{\ell+1} x_j^\ell}{x_i^\ell}$, so that $a_i^{\ell+1}$ plays the role of an effective gain for neuron i in layer $\ell + 1$. Under our proposed initialization, Lemma 4.5 shows that, conditional on $\mathbf{x}^\ell$ with $x_i^\ell \ne 0$, the effective gain $a_i^{\ell+1}$ is exactly Gaussian with conditional mean ω and a state-dependent variance.

Lemma 4.5. *Using the elementwise formulation in* (1) *and employing the proposed weight initialization, fix an arbitrary layer ℓ and index i such that $x_i^\ell \ne 0$. Then, conditionally on x^ℓ,*

$$a_i^{\ell+1} \ \sim\ \mathcal{N}\!\left(\omega,\ \frac{\sigma_z^2}{N_\ell}\Big(1 + \sum_{j \ne i}\Big(\frac{x_j^\ell}{x_i^\ell}\Big)^2\Big)\right). \tag{2}$$

Moreover, if $|x_j^\ell| \le M$ for all j and $|x_i^\ell| \ge \varepsilon > 0$, then $\frac{\sigma_z^2}{N_\ell} \ \le\ \mathrm{Var}\big(a_i^{\ell+1} \,\big|\, x^\ell\big) \ \le\ \sigma_z^2\, \frac{M^2}{\varepsilon^2}$.

Lemma 4.5 specifies a conditional Gaussian law. Averaging over random activations generally yields a Gaussian mixture; the fixed-variance scalar model introduced below is a surrogate. In the network, the conditional gain variance depends on the current coordinate magnitudes, so the surrogate target need not equal the FFNN-level sign-flip rate. The closed form parameterizes the noise scale, while the fitted depth schedule supplies the empirical transfer between these two targets. We next vary the deterministic diagonal gain a_0, which sets the location of the conditional Gaussian distribution. Theorem 4.6 examines the supercritical choice $a_0 > \omega$.

Theorem 4.6. *Let $f \in \mathcal{F}$, $\omega = 1/f'(0)$, $a_0 > \omega$, and $\mathbf{x}^0 \in (\mathbb{R} \setminus \{0\})^n$ be fixed. Consider an equal-width network with zero biases and independent weights $\mathbf{W}^\ell = a_0 \mathbf{I}_n + (\sigma_z/\sqrt{n})\mathbf{G}^\ell$, where the entries of $\mathbf{G}^\ell$ are independent standard Gaussians. For every $\gamma \in (0, 1)$ there exists an integer*

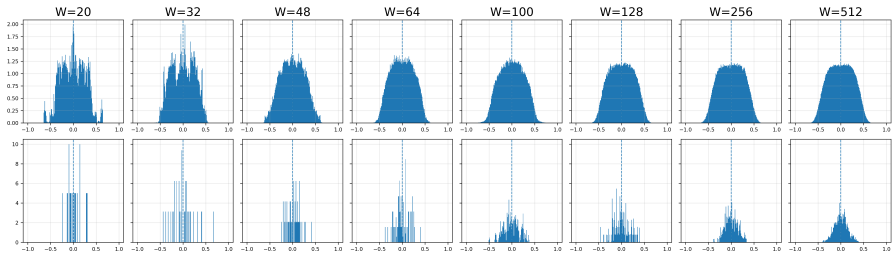


Figure 3: Last layer activation histograms for tanh networks with depth $L = 1000$ and varying width under the proposed initialization **(top row)** and the EOC initialization **(bottom row)**. Each column corresponds to a different hidden width.

L_0 such that, for each finite $L > L_0$, there exists $\sigma_0 > 0$ for which

$$\mathbb{P}\big((1-\gamma)\sigma_z^2 \le \mathrm{Var}(a_i^{\ell+1} \mid \mathbf{x}^\ell) \le (1+\gamma)\sigma_z^2 \quad \text{for all } L_0 \le \ell < L,\ 1 \le i \le n\big) \ge 1-\gamma$$

for every $0 < \sigma_z \le \sigma_0$. The threshold σ_0 may depend on L.

The data dependency of the variance term can be characterized as follows. For each layer ℓ and neuron i with $x_i^\ell(\sigma_z) \neq 0$, define $R_\ell(\sigma_z; i) := \frac{\|\mathbf{x}^\ell(\sigma_z)\|_2^2}{N_\ell\,(x_i^\ell(\sigma_z))^2}$. Here, $x_i^\ell(\sigma_z)$ denotes the activation at layer ℓ under noise scale σ_z. Theorem 4.6 shows that, for a supercritical diagonal gain $a_0 > \omega$, sufficiently small noise yields $R_\ell(\sigma_z; i) \approx 1$ with high probability for $L_0 \le \ell < L$ on each fixed finite horizon $L > L_0$. The admissible noise threshold may depend on L. In the noiseless dynamics, nonzero coordinates approach the common magnitude ξ_{a_0} while retaining their input signs (Figure 8); their limiting magnitudes therefore lose dependence on the initial magnitudes. For $0 < a_0 < \omega$, the noiseless iterates instead contract toward zero (Theorem 4.3 and Proposition A.2). These behaviors motivate choosing $a_0 = \omega$, where $\phi'_\omega(0) = \omega f'(0) = 1$. Although noiseless iterates still converge to zero at this boundary, the choice avoids strict linear contraction at the origin and the nonzero attracting fixed points of the supercritical regime. Mean-field studies likewise highlight the importance of criticality for deep signal propagation (Poole et al., 2016; Schoenholz et al., 2016). With the diagonal gain fixed at $a_0 = \omega$, we select a depth-dependent noise scale using the sign-flip calibration introduced below.

Proposition A.2 and Corollary A.3 suggest that, for most gains a_i^ℓ across layers, if the variance of the gains is either too small or too large, the activations tend to saturate. We define the sign flip rate at depth L as $\pi_L(\sigma) := \mathbb{P}\big(x_L < 0 \,\big|\, x_0 > 0\big)$, i.e., the probability that a neuron that starts with a positive activation ends up with a negative sign at layer L. Under the surrogate model introduced below, it suffices to track a single neuron with $x_i^1 > 0$, since for any $f \in \mathcal{F}$ the activation is sign-preserving and odd. Consequently, the probability that a positive entry becomes negative at layer L is equal to the probability that a negative entry becomes positive, and does not depend on the particular choice of x_i^1.

To handle this dependence on the current activations, we approximate the gain distribution by a scalar surrogate model, $a_i^{\ell+1} \sim \mathcal{N}(\omega, \sigma_z^2)$, and estimate the sign flip rate from this Gaussian approximation. Our initialization has a nonzero mean gain ω, which makes sign changes along a given coordinate well defined across layers. We interpret frequent sign flips during forward propagation as a form of information loss and therefore use the surrogate calibration to control the sign flip rate at the final layer. In practice, we set the noise level so that the empirical FFNN-level sign-flip rate at depth L is close to $p_{\text{real}} = 0.4$. Figure 6 supports this choice: moderate targets ($p = 0.3$–0.4) achieve best or near-best early training accuracy, while the noise-free baseline ($\sigma_z = 0$, $p = 0$) performs comparably or slightly worse in the tested settings.

Theorem 4.7. *Consider the scalar surrogate $X_0 = x_0 > 0$ and $X_j = f(A_j X_{j-1})$, where $f \in \mathcal{F}$ and $A_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\omega, \sigma^2)$. Fix $p \in (0, \frac{1}{2})$, an integer $\ell \ge 1$, and $\omega > 0$. There exists a unique $\sigma^\star > 0$ such that $\mathbb{P}(X_\ell < 0) = p$, given by*

$$\sigma^\star(p, \ell, \omega) = -\frac{\omega}{\Phi^{-1}\left(\dfrac{1-(1-2p)^{1/\ell}}{2}\right)}. \tag{3}$$

At $p = 0$, define $\sigma^\star(0, \ell, \omega) = 0$ by continuity.

Figure 11 shows how the sign flip rate varies with L and p. Figure 12 reports a layerwise mean-square Jacobian diagnostic at the scalar-surrogate scales $\sigma^*(p, L, \omega)$ for $p = 0.01$ and $p = 0.49$. These local gains are close to one over the tested depths. This is a local diagnostic rather than a guarantee of stable end-to-end gradients, which also depend on the accumulation of gains and their correlations. We report direct gradient measurements in Figure 4. The relationship between the FFNN-level sign flip rate and our scalar surrogate, as well as the corresponding validation experiments, is detailed in Appendix B.4.

4.3 Comparative Analysis of Gaussian and Proposed Initializations

Gaussian i.i.d. schemes such as Xavier, He, and EOC all draw weights independently from $\mathcal{N}(0, \sigma_w^2/N_\ell)$, differing only in the choice of σ_w. In this section, we contrast our method with EOC, examining their forward and backward signal propagation both theoretically and empirically.

Forward Pass. As shown in Figures 20, 21, and 22, our proposed initialization keeps the activations well dispersed even at depth 10,000, whereas the EOC initialization drives them toward saturation near zero. Moreover, Figures 3, 23, and 24 show that, for EOC, signal propagation degrades as the width decreases, reflecting that this initialization is derived under an infinite width (or sufficiently wide) assumption. In contrast, our initialization is obtained by directly controlling the effective gain a_i^ℓ and does not rely on any large width approximation, which leads to stable forward propagation even in deep and relatively narrow networks. The theoretical motivation for forward signal propagation with odd-sigmoid activations is discussed in Sections 4.1 and 4.2.

Backward Pass. For equal-width hidden layers of width n, let $J^{\ell+1} = \mathrm{diag}(f'(h^{\ell+1}))W^{\ell+1}$ and define the mean-square Jacobian diagnostic $\chi_{\ell+1} = n^{-1}\mathbb{E}\|J^{\ell+1}\|_F^2$. Appendix A.3 establishes a finite-width error bound for $\chi_{\ell+1} \approx (\omega^2 + \sigma_z^2)\bar{d}_{\ell+1}$, $\bar{d}_{\ell+1} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[f'(h_i^{\ell+1})^2]$. For fixed ω and σ_z, the absolute approximation error is of order $n^{-1/2}$, without assuming independence between weights and preactivations. The diagnostic gives the ratio of expected squared norms for an independent isotropic probe propagated through $(J^{\ell+1})^\top$. Its use in a recursion for actual loss gradients requires the additional conditional second-moment approximation specified in Appendix A.3. We complement this local analysis with numerical measurements of backpropagated gradient norms in Figure 4.

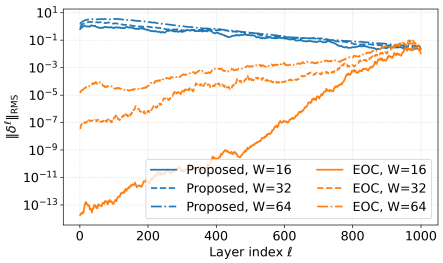


Figure 4: Layerwise gradient norms at initialization for a tanh FFNN of depth $L = 1000$. We set $\boldsymbol{\delta}^\ell := \partial L/\partial \mathbf{h}^\ell$.

The deterministic gain contributes the term ω^2 to the local Jacobian diagnostic. The resulting dependence on the noise scale also involves the derivative moment $\bar{d}_{\ell+1}$. Figure 25 numerically examines this dependence. In the configurations shown in Figure 5, the proposed initialization supports high validation accuracy over a broader region of depths and noise scales than the EOC baseline.

Figure 4 reports the backpropagated preactivation-gradient norms $\|\boldsymbol{\delta}^\ell\|_2$, where $\boldsymbol{\delta}^\ell = \partial\mathcal{L}/\partial h^\ell$, for the quadratic loss $\mathcal{L} = \frac{1}{2}\|\mathbf{y}\|_2^2$. The reported measurements show improved gradient preservation under the proposed initialization. Together, these experiments support robustness across the tested depths and noise scales, complementing the local Jacobian analysis in Appendix A.3.

5 Experiments

Section 5.1 studies activations with $\omega \approx 1$ (data efficiency, depth and width dependence) and Section 5.2 activations with $\omega \not\approx 1$ (BN-free training, learning-rate sensitivity, activation scaling, PINNs), against Xavier, He, and EOC. Settings are given in Appendix C.

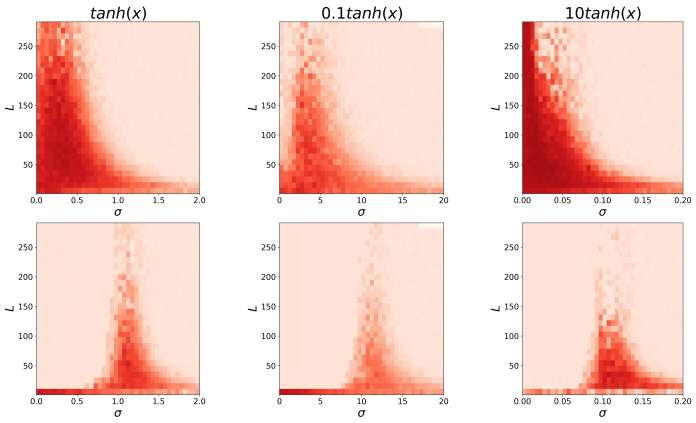


Figure 5: Heatmaps of validation accuracy on MNIST as a function of depth L and noise scale σ for three activations $\tanh(x)$, $0.1\tanh(x)$, and $10\tanh(x)$. Each model is a FFNN with width 128 trained with Adam. Each cell shows the mean validation accuracy over 3 runs. **(Top row)** Proposed. **(Bottom row)** EOC. Darker colors indicate higher validation accuracy.

5.1 Experiments with $\omega \approx 1$ Activations

Sign-flip target calibration. Before comparing against other initializers, we vary the target FFNN-level sign-flip rate p_{real} and numerically calibrate σ_z so that the empirical terminal sign-flip rate matches the target. Figure 6 shows that intermediate sign-flip rates, especially $p \in [0.3, 0.4]$, give the best or near-best early training accuracy across datasets and depths. The deterministic limit $p = 0$ is comparable or slightly worse, while the near-random regime $p = 0.49$ often degrades performance, particularly at depth $L = 100$. We therefore use $p_{\text{real}} = 0.4$ as the default target for classification; the PINN experiments separately evaluate five targets.

Dataset Efficiency. Table 1 presents the best validation accuracy within 50 epochs for a 50 layer FFNN (512 units), trained on small subsets of size 100 or 500. We compare four initializations across odd–sigmoid activations $f \in \mathcal{F}$. The proposed initialization attains the top accuracy in all settings, across both datasets and both sample regimes, with the improvements most pronounced at 100 samples. These results indicate that our method is data efficient, achieving higher validation accuracy with limited data.

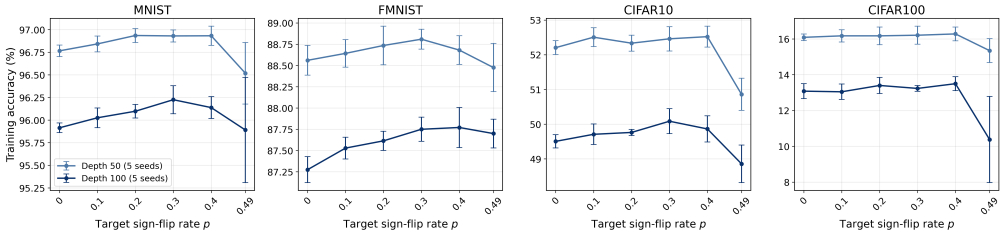


Figure 6: Effect of the target sign-flip rate p in the proposed initialization. Width-64 $\tanh$ FFNNs are initialized with σ_z calibrated by bisection so that the empirical terminal FFNN-level sign-flip rate at depth L equals the target p, and trained for 10 epochs. Curves report mean epoch-10 training accuracy over 5 seeds for depths $L = 50$ and $L = 100$. Across MNIST, Fashion-MNIST, CIFAR-10, and CIFAR-100, rates in $p \in [0, 0.4]$ perform comparably, with $p = 0.3$–0.4 best or near-best, while $p = 0.49$ degrades performance, particularly at $L = 100$.

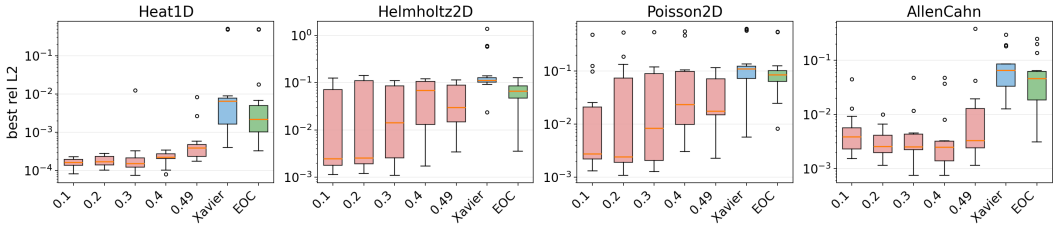


Figure 7: Best relative L^2 error across four PDE benchmarks under three initialization strategies. Depth $L = 50$, width $n = 64$, 300 Adam followed by 100 L-BFGS epochs. Each box aggregates 15 runs — 5 Odd-sigmoid activations ($\alpha \tanh(x)$ for $\alpha \in \{0.2, 0.5, 1, 1.5\}$ and erf) $\times 3$ seeds — with each run contributing its minimum relative L^2 to the reference solution over training. Red — proposed initialization calibrated to five target sign-flip rates $p_{\text{tar}} \in \{0.10, 0.20, 0.30, 0.40, 0.49\}$.

5.2 Experiments with Activations Far from $\omega \approx 1$

Batch Normalization Free Training. For $f(x) = \tanh(ax) + \text{erf}(bx) + \frac{x}{1+|cx|} + \text{gd}(dx)$, by Corollary 4.4 this f is an odd–sigmoid activation. Empirical results for various (a, b, c, d) appear in Figure 31. Across all settings, the proposed initialization achieves the highest validation accuracy even without batch normalization, outperforming Gaussian i.i.d. initializations both with and without batch normalization. We further investigate the effect of applying batch normalization, including training 800 layer networks, in Figures 32 and 33.

Learning-Rate Sensitivity. To study optimization stability, we plot learning rate versus validation accuracy curves on MNIST and Fashion MNIST for 20 layer, width 512 FFNNs with $f(x) = \frac{2}{\pi} \arctan(\alpha x)$ and $f(x) = \tanh(\alpha x)$, using scales α from 10^{-3} to 10^3 (Figures 35, 36, 37, and 38). The proposed method consistently yields a learnable, typically wider LR window across a broad range of α, demonstrating that training remains feasible across diverse ω scales. If the learning rate is chosen too large or too small relative to the ω-scaled band, the first parameter update either destroys the calibrated noise spread or becomes negligible, so that training effectively fails from the second pass onward. Motivated by this, we use the practical learning rate range $\eta \in [\, 10^{-5}\, \omega,\ 10^{-3}\, \omega\,]$.

Scale Preserving Odd–Sigmoid Activations. For $f \in \mathcal{F}$, we investigate whether the proposed and EOC initializations preserve the activation range of the scaled activation αf (Figure 39). Unlike EOC, the proposed initialization maintains an activation range proportional to α even as the network depth increases. Networks with $\alpha \tanh$, $\alpha \arctan$, and $\alpha \,\text{softsign}_2$ activations are trained on MNIST and Fashion MNIST for α ranging from 10^{-2} to 10^9. Tables 2–7 show improved trainability under many large output rescalings, including scales up to 10^9. Since an α scaled activation directly controls the scale of the output y_i, we also evaluate our initialization on regression problems using physics informed neural networks (PINNs). As shown in Figure 7, these results indicate that scaled odd–sigmoid activations are well suited for regression tasks, where controlling the output range is important. We defer detailed PINN setups and PDE formulations to Appendix C.6.

6 Conclusion

We proposed an activation-aware, diagonal-plus-noise initialization for feedforward networks with odd-sigmoid activations. Our scalar fixed-point analysis identifies the critical gain $\omega = 1/f'(0)$, and we derive a closed-form noise scale from a target terminal sign-flip rate in an i.i.d. scalar surrogate, with a depth schedule whose form follows from the critical scaling $L\sigma_z^2 \approx s$. Our initialization does not rely on a large-width approximation. Experiments show dispersed forward activations and stable gradient propagation across depths and widths, together with successful training under output rescalings up to $10^9 \tanh(x)$. The method outperforms Xavier, He, and EOC in the tested deep, narrow, and rescaled networks and improves PINN performance on four PDE benchmarks. The analysis is tailored to the odd-sigmoid class and to plain feedforward architectures; extensions to ReLU/GELU-type activations, to convolutional and attention-based models, and a fully theoretical derivation of the surrogate-to-FFNN transfer are left to future work.

AI USE STATEMENT

We used generative AI tools to check grammar and improve the readability of author-written text, and to search for related work. All references were verified manually against the original sources. We reviewed all AI-assisted content and take full responsibility for the content of this paper.

REFERENCES

Thomas Bachlechner, Bodhisattwa Prasad Majumder, Henry Mao, Gary Cottrell, and Julian McAuley. Rezero is all you need: Fast convergence at large depth. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021)*, pp. 1352–1361. PMLR, 2021.

Yaniv Blumenfeld, Dar Gilboa, and Daniel Soudry. Beyond signal propagation: is feature diversity necessary in deep neural network initialization? In *International Conference on Machine Learning*, pp. 960–969. PMLR, 2020.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

Soufiane Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of the activation function on deep neural networks training. In *International conference on machine learning*, pp. 2672–2680. PMLR, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, 2015. URL `https://proceedings.mlr.press/v37/ioffe15.html`.

Gaurav Iyer, Boris Hanin, and David Rolnick. Maximal initial learning rates in deep relu networks. In *International Conference on Machine Learning*, pp. 14500–14530. PMLR, 2023.

Ameya D. Jagtap, Kenji Kawaguchi, and George Em Karniadakis. Adaptive activation functions accelerate convergence in deep and physics-informed neural networks. *Journal of Computational Physics*, 404:109136, 2020.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Hyunwoo Lee, Yunho Kim, Seung Yeop Yang, and Hayoung Choi. Improved weight initialization for deep and narrow feedforward neural network. *Neural Networks*, 176:106362, 2024.

Hyunwoo Lee, Hayoung Choi, and Hyunju Kim. Robust weight initialization for tanh neural networks with fixed point analysis. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Advances in neural information processing systems*, 29, 2016.

Junfei Qiao, Sanyi Li, and Wenjing Li. Mutual information based weight initialization method for sigmoidal feedforward neural networks. *Neurocomputing*, 207:676–683, 2016. doi: 10.1016/j.neucom.2016.05.054.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *International Conference on Machine Learning*, pp. 2847–2854, 2017.

Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

Pratik Rathore, Weimu Lei, Zachary Frangella, Lu Lu, and Madeleine Udell. Challenges in training PINNs: A loss landscape perspective. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024.

Andrew M Saxe, James McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *arXiv preprint arXiv:1611.01232*, 2016.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pp. 1139–1147. pmlr, 2013.

Honghui Wang, Lu Lu, Shiji Song, and Gao Huang. Learning specialized activation functions for physics-informed neural networks. *arXiv preprint arXiv:2308.04073*, 2023.

Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel S. Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5393–5402. PMLR, 2018.

Ahmet Yilmaz and Riccardo Poli. Successfully and efficiently training deep multi-layer perceptrons with logistic activation function simply requires initializing the weights with an appropriate negative mean. *Neural Networks*, 153:87 – 103, 2022. URL `https://doi.org/10.1016/j.neunet.2022.05.030`.

SUPPLEMENTARY MATERIAL

The supplementary material is structured as follows.

- Appendix A provides the theoretical results and their proofs.
- Appendix B contains additional simulation results without neural networks.
- Appendix C reports extra experimental details and results with neural networks.

A THEORETICAL RESULTS

In this section, we provide proofs for the statements presented in Section 4.

A.1 THEORETICAL RESULTS FOR SECTION 4.1

Lemma A.1. *If $f \in \mathcal{F}$, then the following holds.*

(i) $f(0) = 0$.

(ii) $0 = \arg\max |f'(x)|$.

(iii) $\lim_{x\to\infty} f'(x) = 0$.

Proof. (i) It is trivial that odd symmetry implies $f(0) = 0$. (ii) Since f' is an even function, by Definition 4.1(v) $f'(x)$ has the maximum value at $x = 0$. (iii) Since f' is strictly decreasing on $[0, \infty)$ and $f'(x) > 0$ for all $x \in \mathbb{R}$, the Monotone Convergence Theorem for real functions implies that $\ell := \lim_{x\to\infty} f'(x)$ exists for some $\ell \in [0, \infty)$. Suppose that $\ell > 0$. Then there exists $R > 0$ such that

$$f'(x) \geq \frac{\ell}{2} \quad \text{for all } x \geq R.$$

Integrating from R to x, we obtain

$$f(x) - f(R) = \int_R^x f'(t)\, dt \;\geq\; \frac{\ell}{2}(x - R) \quad \text{for all } x > R.$$

Hence $f(x) \to \infty$ as $x \to \infty$, which contradicts the boundedness of f. Therefore $\ell = 0$. □

Lemma A.1 summarizes the basic regularity and saturation properties of odd-sigmoid activations in $\mathcal{F}$. In particular, it shows that every $f \in \mathcal{F}$ has a unique global slope maximum at the origin and becomes arbitrarily flat in the tails. These simple but structural features will be crucial for understanding how the gain parameter a reshapes the fixed-point structure of the scalar map $x \mapsto f(ax)$.

Proposition 4.2 Suppose $f \in \mathcal{F}$ with $\omega := 1/f'(0)$, and for a fixed $a > 0$ define $\phi_a(x) := f(ax)$. Then

(i) If $0 < a \leq \omega$, then $f(ax)$ has a unique fixed point $x^* = 0$.

(ii) If $a > \omega$, then $f(ax)$ has three distinct fixed points: $x^* = -\xi_a,\ 0,\ \xi_a$ such that $\xi_a > 0$.

Proof. For $a > 0$, consider $g(x, a) := f(ax) - x$. We have $g(0, a) = 0$ and $g'(x, a) = a\, f'(ax) - 1$.

Case (i): $0 < a \leq \omega$. For $x > 0$ we have $ax > 0$, and $f'(ax) < f'(0)$; hence

$$g'(x, a) = a\, f'(ax) - 1 < a\, f'(0) - 1 \;\leq\; 0,$$

for all $x > 0$. Thus $g(\cdot, a)$ is strictly decreasing on $(0, \infty)$, $g(0, a) = 0$, and $\lim_{x\to\infty} g(x, a) = -\infty$ by boundedness of f, so $g(x, a) < 0$ for all $x > 0$ and there is no positive root. Since $g(\cdot, a)$ is odd, there is no negative root either. Hence $x = 0$ is the unique solution.

Case (ii): $a > \omega$. Then $g'(0, a) = af'(0) - 1 > 0$. $g'(\cdot, a)$ is strictly decreasing on $[0, \infty)$, and by the Lemma A.1

$$\lim_{x\to\infty} g'(x, a) = a \lim_{x\to\infty} f'(ax) - 1 = -1 < 0.$$

By the intermediate value theorem, there exists a unique $\hat{x} > 0$ with $g'(\hat{x}, a) = 0$. Hence $g(\cdot, a)$ is strictly increasing on $[0, \hat{x}]$ and strictly decreasing on $[\hat{x}, \infty)$. Since $g(0, a) = 0$ and $g'(0, a) > 0$, we have $g(x, a) > 0$ for all $0 < x \leq \hat{x}$. Since f is bounded (Definition 4.1(iii)), $\lim_{x\to\infty} g(x, a) = -\infty$; because g is strictly decreasing on $[\hat{x}, \infty)$ and continuous, there exists a unique $\xi_a > \hat{x}$ with $g(\xi_a, a) = 0$. Thus, there is exactly one positive, nonzero root. Oddness of g yields the symmetric negative root $-\xi_a$, so the full set of real solutions is $\{-\xi_a,\ 0,\ \xi_a\}$. □

Proposition 4.2 provides a precise characterization of how the fixed points of the scalar map $x \mapsto f(ax)$ bifurcate as the gain a crosses the critical value $\omega = 1/f'(0)$. For $a \leq \omega$ the origin is the only fixed point, while for $a > \omega$ a symmetric nonzero pair $\pm\xi_a$ emerges, exhibiting a classical pitchfork structure. We next lift this static picture to the dynamical setting, showing that the iterates of $x_{n+1} = f(ax_n)$ converge to the corresponding fixed point for every positive initial condition.

Theorem 4.3 Suppose $f \in \mathcal{F}$ with $\omega := 1/f'(0)$, and for a fixed $a > 0$ define

$$x_0 > 0, \qquad x_{n+1} = \phi_a(x_n), \qquad n = 0, 1, 2, \ldots.$$

Then the sequence $\{x_n\}$ converges for every $x_0 > 0$. Furthermore,

(1) if $0 < a \leq \omega$, then $x_n \to 0$ as $n \to \infty$.

(2) if $a > \omega$, then $x_n \to \xi_a$ as $n \to \infty$.

Proof. (1) For $0 < a \leq \omega$ and $x > 0$, slope decay gives $0 < f(ax) = \int_0^{ax} f'(t)\, dt < af'(0)x \leq x$. Thus (x_n) is positive and decreasing, and its limit is a fixed point by continuity. Proposition 4.2 identifies that limit as zero.
(2) Let $0 < x_0 < \xi_a$. From the proof of Proposition 4.2, $g(x, a) = f(ax) - x > 0$ on $(0, \xi_a)$, so $x_{n+1} = \phi_a(x_n) > x_n$ whenever $0 < x_n < \xi_a$; since ϕ_a is strictly increasing and $\phi_a(\xi_a) = \xi_a$, also $x_{n+1} < \xi_a$. Hence (x_n) is increasing and bounded above by ξ_a, so it converges, and by continuity its limit is a fixed point in $(0, \xi_a]$, which must be ξ_a. If $x_0 > \xi_a$, then $g(x, a) < 0$ on (ξ_a, ∞) gives $x_{n+1} < x_n$, and monotonicity of ϕ_a gives $x_{n+1} > \xi_a$; the sequence is decreasing and bounded below by ξ_a, and its limit is again ξ_a. If $x_0 = \xi_a$ the sequence is constant. □

Corollary 4.4 Let $f_1, f_2 \in \mathcal{F}$ and let $c_1, c_2 \geq 0$ with $(c_1, c_2) \neq (0, 0)$. If $g = c_1 f_1 + c_2 f_2$, then $g \in \mathcal{F}$. Furthermore, it holds that

$$\frac{1}{\omega_g} = \frac{c_1}{\omega_{f_1}} + \frac{c_2}{\omega_{f_2}}.$$

Proof. Let $g = c_1 f_1 + c_2 f_2$. Since $c_1, c_2 \geq 0$, the linear combination preserves oddness and bounded saturation by linearity. Strict monotonicity on $\mathbb{R}$ follows from $g'(x) = c_1 f_1'(x) + c_2 f_2'(x) > 0$, and slope decay on $[0, \infty)$ is preserved because a positive linear combination of strictly decreasing functions is strictly decreasing. Thus $g \in \mathcal{F}$. Evaluating at the origin gives

$$g'(0) = c_1 f_1'(0) + c_2 f_2'(0) = \frac{c_1}{\omega_{f_1}} + \frac{c_2}{\omega_{f_2}},$$

so by definition $1/\omega_g = g'(0)$, which yields the claimed identity. □

In practice this means that we can build richer odd-sigmoid activations by mixing simpler ones without losing the pitchfork structure described above. We now move from the constant–gain setting to the more general case where the layerwise gains (a_n) are allowed to vary with depth.

Proposition A.2. *Let $f \in \mathcal{F}$ and $\{a_n\}_{n=1}^{\infty}$ be a positive real sequence, i.e., $a_n > 0$ for all $n \in \mathbb{N}$, such that only finitely many elements are greater than $\omega = 1/f'(0)$. For a positive sequence $\{a_n\}_{n\geq 1}$, set*

$$\Phi^m \;:=\; \phi_{a_m} \circ \phi_{a_{m-1}} \circ \cdots \circ \phi_{a_1}.$$

Then for any $x \in \mathbb{R}$

$$\lim_{m\to\infty} \Phi^m(x) = 0.$$

Proof. Let $N := \max\{n :\; a_n > \omega\}$ (take $N = 0$ if the set is empty) and define

$$b_n = \begin{cases} a_n, & n \leq N, \\ 0, & n > N, \end{cases} \qquad c_n = \begin{cases} a_n, & n \leq N, \\ \omega, & n > N. \end{cases}$$

Let $\hat{\Phi}^m := \phi_{b_m} \circ \cdots \circ \phi_{b_1}$ and $\tilde{\Phi}^m := \phi_{c_m} \circ \cdots \circ \phi_{c_1}$. For $x \geq 0$ and $n > N$, by definition of $\mathcal{F}$, we obtain

$$\hat{\Phi}^m(x) \;\leq\; \Phi^m(x) \;\leq\; \tilde{\Phi}^m(x).$$

Oddness yields the same in absolute value for all $x \in \mathbb{R}$:

$$|\hat{\Phi}^m(x)| \;\leq\; |\Phi^m(x)| \;\leq\; |\tilde{\Phi}^m(x)|.$$

By Theorem 4.3, $\hat{\Phi}^m(x) \to 0$ and $\tilde{\Phi}^m(x) \to 0$. The squeeze theorem gives $\Phi^m(x) \to 0$. □

Corollary A.3. *Let $\epsilon > 0$ be given and set $\omega := 1/f'(0)$. Suppose that $\{a_n\}_{n=1}^{\infty}$ be a positive real sequence such that only finitely many elements are lower than $\omega + \epsilon$. Then for any $x \in \mathbb{R} \setminus \{0\}$*

$$\liminf_{m\to\infty} |\Phi^m(x)| \;\geq\; \xi_{\omega+\epsilon}.$$

Proof. Let $N := \max\{n :\; a_n < \omega + \epsilon\}$ (take $N = 0$ if the set is empty) and define

$$b_n = \begin{cases} a_n, & n \leq N, \\ \omega + \epsilon, & n > N. \end{cases} \qquad \hat{\Phi}_m := \phi_{b_m} \circ \cdots \circ \phi_{b_1}, \qquad \Phi^m := \phi_{a_m} \circ \cdots \circ \phi_{a_1},$$

where $\phi_a(x) := f(ax)$. By definition of $\mathcal{F}$,

$$|\hat{\Phi}^m(x)| \;\leq\; |\Phi^m(x)| \qquad (\forall x \in \mathbb{R},\; \forall m).$$

Taking $\liminf$ in the inequality yields

$$\liminf_{m\to\infty} |\Phi^m(x)| \;\geq\; \xi_{\omega+\epsilon}.$$

□

Proposition A.2 and Corollary A.3 together describe two opposite extremes of how layerwise gains affect the one-dimensional dynamics. Roughly speaking, Proposition A.2 says that if, after some finite depth, all gains a_n stay below the critical value ω, then the composed map Φ^m always drives the signal back to zero, no matter what happened in the earlier layers. In contrast, Corollary A.3 shows that if the gains are eventually bounded away from ω by a fixed margin $\varepsilon > 0$, then the compositions cannot collapse to zero: for any nonzero input, the iterates stay at least as large (in absolute value) as the positive fixed point $\xi_{\omega+\varepsilon}$.

A.2 Theoretical Results for Section 4.2

Lemma 4.5 Using the elementwise formulation in (1) and employing the proposed weight initialization, fix an arbitrary layer ℓ and index i such that $x_i^\ell \neq 0$. Then, conditionally on x^ℓ,

$$a_i^{\ell+1} \;\sim\; \mathcal{N}\!\left(\omega,\; \frac{\sigma_z^2}{N_\ell}\Big(1 + \sum_{j\neq i}\Big(\frac{x_j^\ell}{x_i^\ell}\Big)^2\Big)\right).$$

Moreover, if $|x_j^\ell| \leq M$ for all j and $|x_i^\ell| \geq \varepsilon > 0$, then

$$\frac{\sigma_z^2}{N_\ell} \;\leq\; \mathrm{Var}\big(a_i^{\ell+1} \,\big|\, x^\ell\big) \;\leq\; \sigma_z^2\,\frac{M^2}{\varepsilon^2}.$$

Proof. From (1) and the proposed initialization, we write

$$W^{\ell+1} \;=\; D^{\ell+1} + Z^{\ell+1},$$

where, in the equal-width case $N_{\ell+1} = N_\ell$ assumed here, the diagonal of $D^{\ell+1}$ equals ω and $Z^{\ell+1}$ has independent entries $(Z^{\ell+1})_{ij} \sim \mathcal{N}(0, \sigma_z^2/N_\ell)$. The pre-activation at coordinate i reads

$$s_i^{\ell+1} \;=\; \sum_{j=1}^{N_\ell} \big((D^{\ell+1})_{ij} + (Z^{\ell+1})_{ij}\big)\, x_j^\ell \;=\; \omega\, x_i^\ell + (Z^{\ell+1})_{ii}\, x_i^\ell + \sum_{j\neq i} (Z^{\ell+1})_{ij}\, x_j^\ell.$$

On the event $x_i^\ell \neq 0$, we define the effective gain

$$a_i^{\ell+1} := \frac{s_i^{\ell+1}}{x_i^\ell} \;=\; \omega + (Z^{\ell+1})_{ii} + \sum_{j\neq i} (Z^{\ell+1})_{ij}\, \frac{x_j^\ell}{x_i^\ell}.$$

Conditionally on x^ℓ, the coefficients $\{x_j^\ell/x_i^\ell\}$ are deterministic, whereas the random variables $\{(Z^{\ell+1})_{ij}\}_{j=1}^{N_\ell}$ are independent, centered Gaussians with variance σ_z^2/N_ℓ. Therefore $a_i^{\ell+1} \mid x^\ell$ is a linear combination of independent Gaussians, hence Gaussian:

$$a_i^{\ell+1} \,\big|\, x^\ell \;\sim\; \mathcal{N}\!\left(\omega,\; \frac{\sigma_z^2}{N_\ell}\Big(1 + \sum_{j\neq i}\Big(\frac{x_j^\ell}{x_i^\ell}\Big)^2\Big)\right).$$

This yields the conditional variance

$$\mathrm{Var}\big(a_i^{\ell+1} \mid x^\ell\big) = \frac{\sigma_z^2}{N_\ell}\Big(1 + \sum_{j\neq i}\Big(\frac{x_j^\ell}{x_i^\ell}\Big)^2\Big).$$

Since the summation term is nonnegative, we obtain the conditional lower bound

$$\mathrm{Var}(a_i^{\ell+1} \mid x^\ell) \geq \sigma_z^2/N_\ell.$$

For the upper bound, note that

$$1 + \sum_{j\neq i}\Big(\frac{x_j^\ell}{x_i^\ell}\Big)^2 \;=\; \frac{(x_i^\ell)^2 + \sum_{j\neq i}(x_j^\ell)^2}{(x_i^\ell)^2} \;=\; \frac{\|x^\ell\|_2^2}{(x_i^\ell)^2},$$

whence

$$\mathrm{Var}\big(a_i^{\ell+1} \mid x^\ell\big) = \frac{\sigma_z^2}{N_\ell}\,\frac{\|x^\ell\|_2^2}{(x_i^\ell)^2}.$$

If $|x_j^\ell| \leq M$ for all j, then $\|x^\ell\|_2^2 = \sum_{j=1}^{N_\ell}(x_j^\ell)^2 \leq N_\ell M^2$, and if $|x_i^\ell| \geq \varepsilon > 0$, we obtain

$$\frac{\|x^\ell\|_2^2}{(x_i^\ell)^2} \;\leq\; \frac{N_\ell M^2}{\varepsilon^2},$$

which implies

$$\mathrm{Var}\big(a_i^{\ell+1} \mid x^\ell\big) \;\leq\; \frac{\sigma_z^2}{N_\ell}\,\frac{N_\ell M^2}{\varepsilon^2} \;=\; \sigma_z^2\,\frac{M^2}{\varepsilon^2}.$$

□

Lemma 4.5 gives the exact conditional Gaussian law of the effective gain. Its conditional variance equals σ_z^2 times a state-dependent factor and is at least σ_z^2/N_ℓ. For bounded activations, the conditional upper bound applies when the denominator coordinate is bounded away from zero.

Theorem 4.6 Let $f \in \mathcal{F}$, $\omega = 1/f'(0)$, $a_0 > \omega$, and $\mathbf{x}^0 \in (\mathbb{R} \setminus \{0\})^n$ be fixed. Consider an equal-width network with zero biases and independent weights $\mathbf{W}^\ell = a_0\mathbf{I}_n + (\sigma_z/\sqrt{n})\mathbf{G}^\ell$, where the entries of $\mathbf{G}^\ell$ are independent standard Gaussians. For every $\gamma \in (0,1)$ there exists an integer L_0 such that, for each finite $L > L_0$, there exists $\sigma_0 > 0$ for which

$$\mathbb{P}\big((1-\gamma)\sigma_z^2 \le \mathrm{Var}(a_i^{\ell+1} \mid \mathbf{x}^\ell) \le (1+\gamma)\sigma_z^2 \quad \text{for all } L_0 \le \ell < L,\ 1 \le i \le n\big) \ge 1-\gamma$$

for every $0 < \sigma_z \le \sigma_0$. The threshold σ_0 may depend on L.

Proof. Write $\mathbf{y}^\ell = \mathbf{x}^\ell(0)$. By Theorem 4.3, $|y_i^\ell| \to \xi_{a_0} > 0$ for every i. Consequently,

$$R_\ell(0;i) := \frac{\|\mathbf{y}^\ell\|_2^2}{n(y_i^\ell)^2} \to 1.$$

Choose L_0 so that $|R_\ell(0;i) - 1| \le \gamma/2$ for every $\ell \ge L_0$ and i. Fix a finite $L > L_0$ and couple all noise scales using the same Gaussian matrices $\mathbf{G}^1, \dots, \mathbf{G}^L$. Continuity of f implies, by induction, that each $\mathbf{x}^\ell(\sigma_z)$ is continuous at $\sigma_z = 0$ for every fixed realization of these matrices. Since every y_i^ℓ is nonzero,

$$\max_{L_0 \le \ell < L} \max_{1 \le i \le n} |R_\ell(\sigma_z;i) - R_\ell(0;i)| \longrightarrow 0 \quad \text{almost surely as } \sigma_z \downarrow 0.$$

Hence the probability that this maximum is at most $\gamma/2$ tends to one. Choose $\sigma_0 > 0$ so that this probability is at least $1-\gamma$ whenever $0 < \sigma_z \le \sigma_0$. Lemma 4.5, with a_0 replacing ω, gives $\mathrm{Var}(a_i^{\ell+1} \mid \mathbf{x}^\ell) = \sigma_z^2 R_\ell(\sigma_z;i)$, completing the proof. □

The previous results describe how the magnitude of the effective gains behaves across layers (Figures 8 and 9). To control the sign dynamics, we now turn to a scalar surrogate model in which the gains are i.i.d. Gaussian with mean ω and variance σ^2. In this simplified setting, the only quantity that matters is the probability that the scalar iterate becomes negative at a given depth. The next lemma provides a closed-form recursion for this sign flip rate.

Lemma A.4. *Let $f \in \mathcal{F}$ and $x_0 > 0$, for every $j \ge 1$,*

$$\pi_j = \tfrac{1}{2}\Big(1 - (1-2p_-)^j\Big), \qquad p_- = \mathbb{P}(A_1 < 0) = \Phi\Big(-\frac{\omega}{\sigma}\Big). \tag{4}$$

Proof. Since f is odd and strictly increasing, $\mathrm{sign}(f(u)) = \mathrm{sign}(u)$ for all u, hence

$$\{X_j < 0\} = \{A_j X_{j-1} < 0\} = \big(\{A_j < 0\} \cap \{X_{j-1} > 0\}\big) \cup \big(\{A_j > 0\} \cap \{X_{j-1} < 0\}\big),$$

up to null sets (because $\mathbb{P}(A_j = 0) = 0$ for Gaussian A_j). Independence of A_j and X_{j-1} yields

$$\pi_j = \mathbb{P}(A_j < 0)\,\mathbb{P}(X_{j-1} > 0) + \mathbb{P}(A_j > 0)\,\mathbb{P}(X_{j-1} < 0) = p_-(1-\pi_{j-1}) + (1-p_-)\pi_{j-1}.$$

Thus $\pi_j = (1-2p_-)\pi_{j-1} + p_-$. With $\pi_0 = \mathbb{P}(X_0 < 0) = 0$, the first-order linear recursion solves to

$$\pi_j - \tfrac{1}{2} = (1-2p_-)\big(\pi_{j-1} - \tfrac{1}{2}\big) \implies \pi_j - \tfrac{1}{2} = (1-2p_-)^j\big(\pi_0 - \tfrac{1}{2}\big) = -\tfrac{1}{2}(1-2p_-)^j.$$

The value of p_- follows from $A_1 \sim \mathcal{N}(\omega, \sigma^2)$. □

Theorem 4.7 $p \in (0, \frac{1}{2})$, a depth $\ell \in \mathbb{N}$, and $\omega > 0$. There exists a unique $\sigma^\star = \sigma^\star(p, \ell, \omega) > 0$ such that $\pi_\ell(\sigma^\star) = p$, and it is given by

$$\sigma^\star(p, \ell, \omega) \;=\; -\,\frac{\omega}{\Phi^{-1}\!\left(\dfrac{1-(1-2p)^{1/\ell}}{2}\right)}. \tag{5}$$

Proof. $\pi_\ell(\cdot)$ is continuous and strictly increasing from 0 (at $\sigma \downarrow 0$) to $1/2$ (as $\sigma \uparrow \infty$). Hence for any $p \in (0, 1/2)$ there exists a unique $\sigma^\star > 0$ with $\pi_\ell(\sigma^\star) = p$.

To obtain the explicit form, set $q := p_-(\sigma^\star) = \Phi(-\omega/\sigma^\star) \in (0, 1/2)$. From Lemma A.4 we have

$$p = \pi_\ell(\sigma^\star) = \tfrac{1}{2}\Big(1 - (1-2q)^\ell\Big) \quad \Longrightarrow \quad q = \frac{1-(1-2p)^{1/\ell}}{2}.$$

Applying the inverse CDF Φ^{-1} to $q = \Phi(-\omega/\sigma^\star)$ yields

$$-\frac{\omega}{\sigma^\star} = \Phi^{-1}(q) \quad \Longrightarrow \quad \sigma^\star = -\frac{\omega}{\Phi^{-1}(q)}.$$

Since $q \in (0, 1/2)$, the quantile $\Phi^{-1}(q) < 0$, so the right-hand side is positive. □

A.3 Theoretical Results for Section 4.3

We give a finite-width characterization of a local Jacobian diagnostic for the proposed initialization. Let $f \in \mathcal{F}$ and $\omega = 1/f'(0)$. Consider equal-width hidden layers with $N_\ell = N_{\ell+1} = n$ and zero biases:

$$W^{\ell+1} = \omega I_n + Z^{\ell+1}, \qquad Z_{ij}^{\ell+1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_z^2/n).$$

Write

$$h^{\ell+1} = W^{\ell+1}x^\ell, \qquad J^{\ell+1} = \operatorname{diag}(f'(h^{\ell+1}))W^{\ell+1},$$

and define

$$\chi_{\ell+1} := \frac{1}{n}\mathbb{E}\|J^{\ell+1}\|_F^2, \qquad \bar{d}_{\ell+1} := \frac{1}{n}\sum_{i=1}^n \mathbb{E}[f'(h_i^{\ell+1})^2].$$

Interpretation of the diagnostic. Let $v \sim \mathcal{N}(0, I_n)$ be independent of the network. Conditioning on $J^{\ell+1}$ yields

$$\mathbb{E}_v\big[\|(J^{\ell+1})^\top v\|_2^2 \,\big|\, J^{\ell+1}\big] = \|J^{\ell+1}\|_F^2.$$

Since $\mathbb{E}\|v\|_2^2 = n$, it follows that

$$\frac{\mathbb{E}\|(J^{\ell+1})^\top v\|_2^2}{\mathbb{E}\|v\|_2^2} = \chi_{\ell+1}.$$

Thus, $\chi_{\ell+1}$ is the ratio of expected squared norms for an independent isotropic probe propagated through the transposed Jacobian.

Finite-width error bound. Suppress the layer index and define the squared row norm $S_i = \sum_{j=1}^n W_{ij}^2$. Its exact expansion is

$$S_i = \omega^2 + 2\omega Z_{ii} + \sum_{j=1}^n Z_{ij}^2.$$

The Gaussian moments give

$$\begin{aligned}\mathbb{E}S_i &= \omega^2 + \sigma_z^2,\\ \operatorname{Var}(S_i) &= 4\omega^2 \operatorname{Var}(Z_{ii}) + \sum_{j=1}^n \operatorname{Var}(Z_{ij}^2)\\ &= \frac{4\omega^2\sigma_z^2 + 2\sigma_z^4}{n}.\end{aligned}$$

Here, the covariance between Z_{ii} and $\sum_j Z_{ij}^2$ vanishes by independence of the entries and the zero third moment of a centered Gaussian variable.

By the definition of the Frobenius norm,

$$\chi_{\ell+1} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[f'(h_i^{\ell+1})^2 S_i].$$

Using $0 \le f'(x) \le f'(0)$ for $f \in \mathcal{F}$ and Cauchy–Schwarz, we obtain

$$\begin{aligned}
&\left|\chi_{\ell+1} - (\omega^2 + \sigma_z^2)\bar{d}_{\ell+1}\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[f'(h_i^{\ell+1})^2\big(S_i - \mathbb{E}S_i\big)\big]\right| \\
&\le \frac{f'(0)^2}{n}\sum_{i=1}^{n} \mathbb{E}|S_i - \mathbb{E}S_i| \\
&\le \frac{f'(0)^2}{n}\sum_{i=1}^{n} \sqrt{\mathrm{Var}(S_i)} \\
&= f'(0)^2\sqrt{\frac{4\omega^2\sigma_z^2 + 2\sigma_z^4}{n}}.
\end{aligned}$$

This bound does not require independence between weights and preactivations, or exchangeability of the activation coordinates. For fixed ω and σ_z, it justifies

$$\chi_{\ell+1} \approx (\omega^2 + \sigma_z^2)\bar{d}_{\ell+1} \tag{6}$$

with an absolute error of order $n^{-1/2}$.

For comparison, consider centered i.i.d. Gaussian weights with variance σ_w^2/n. Denote the corresponding diagnostic and derivative moment by $\chi_{\ell+1}^{\mathrm{G}}$ and $\bar{d}_{\ell+1}^{\mathrm{G}}$, respectively. The same argument gives

$$\left|\chi_{\ell+1}^{\mathrm{G}} - \sigma_w^2\bar{d}_{\ell+1}^{\mathrm{G}}\right| \le f'(0)^2\sigma_w^2\sqrt{\frac{2}{n}}.$$

The derivative moments in these two initializations are evaluated under their respective preactivation distributions.

Relation to backpropagated loss gradients. For a scalar loss $\mathcal{L}$, let $g^\ell = \partial\mathcal{L}/\partial x^\ell$. The exact backpropagation identity is

$$g^\ell = (J^{\ell+1})^\top g^{\ell+1}.$$

An actual loss gradient generally depends on the network. To obtain a scalar recursion using $\chi_{\ell+1}$, a sufficient additional approximation is

$$\mathbb{E}\big[g^{\ell+1}(g^{\ell+1})^\top \,\big|\, J^{\ell+1}\big] \approx q_{\ell+1} I_n, \qquad q_{\ell+1} := \frac{1}{n}\mathbb{E}\|g^{\ell+1}\|_2^2.$$

Under this conditional second-moment approximation,

$$\frac{1}{n}\mathbb{E}\|g^\ell\|_2^2 \approx \chi_{\ell+1}\frac{1}{n}\mathbb{E}\|g^{\ell+1}\|_2^2. \tag{7}$$

The finite-width bound above establishes the local Jacobian diagnostic, while this gradient recursion requires the additional approximation just stated. We use the recursion as an auxiliary interpretation and assess backpropagated gradient norms numerically. Even under this approximation, amplification across depth depends on the product of the layerwise χ factors; values close to one at individual layers do not alone guarantee stability at arbitrary depths.

Equation (6) identifies the contribution ω^2 of the deterministic gain to the local Jacobian diagnostic. Its dependence on the noise scale also involves $\bar{d}_{\ell+1}$, which changes with the preactivation distribution. Accordingly, the wider useful range of noise scales reported in Section 4.3 is supported by the experiments, rather than being a direct consequence of this decomposition.

B Additional Experimental Results without Trained Networks

B.1 Definitions of Activation Functions

This section introduces the activation functions considered in this paper. The following functions belong to the odd-sigmoid function class.

Gudermannian function The Gudermannian function $\mathrm{gd} : \mathbb{R} \to \mathbb{R}$ is defined by

$$\mathrm{gd}(x) = \int_0^x \frac{dt}{\cosh t} = 2\arctan\bigl(\tanh\bigl(\tfrac{x}{2}\bigr)\bigr)\,.$$

Error Function. The error function $\mathrm{erf} : \mathbb{R} \to \mathbb{R}$ is defined as

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt.$$

Softsign-Type Functions. For $k \geq 1$, we define the generalized softsign function

$$\mathrm{softsign}_k(x) \;=\; \frac{x}{\bigl(1 + |x|^k\bigr)^{1/k}}, \qquad x \in \mathbb{R}.$$

This family interpolates between several commonly used smooth odd activations:

$$\mathrm{softsign}_1(x) = \frac{x}{1+|x|}, \quad \mathrm{softsign}_2(x) = \frac{x}{\sqrt{1+x^2}}, \quad \mathrm{softsign}_3(x) = \frac{x}{\bigl(1+|x|^3\bigr)^{1/3}}.$$

We also use the combined variant

$$\mathrm{softsign}_{1+2}(x) := \mathrm{softsign}_1(x) + \mathrm{softsign}_2(x).$$

Hyperbolic Tangent. The hyperbolic tangent function $\tanh : \mathbb{R} \to \mathbb{R}$ is defined by

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

Arctangent. The (scaled) arctangent function $\arctan : \mathbb{R} \to \mathbb{R}$ is given by

$$\arctan(x) = \int_0^x \frac{dt}{1+t^2}.$$

In practice, we often use the normalized form $\frac{2}{\pi}\arctan(x)$ so that its range matches that of $\tanh(x)$.

B.2 Properties of the Odd-Sigmoid Functions

This section empirically investigates the properties of the iteration $x_{n+1} = f(ax_n)$. When the gain a is fixed across all iterations, Figure 8 shows how the dynamics depend on the initial data: for any nonzero input, the iterates converge to the same nonzero fixed point ξ_a. Figure 9 further demonstrates that this limit depends only on the gain, not on the input, by comparing distinct gains $a \neq a' > 0$ and observing convergence to different fixed points ξ_a and $\xi_{a'}$. These results indicate that, by appropriately choosing a, one can control how long the signal remains informative across layers and thus preserve information up to a desired depth.

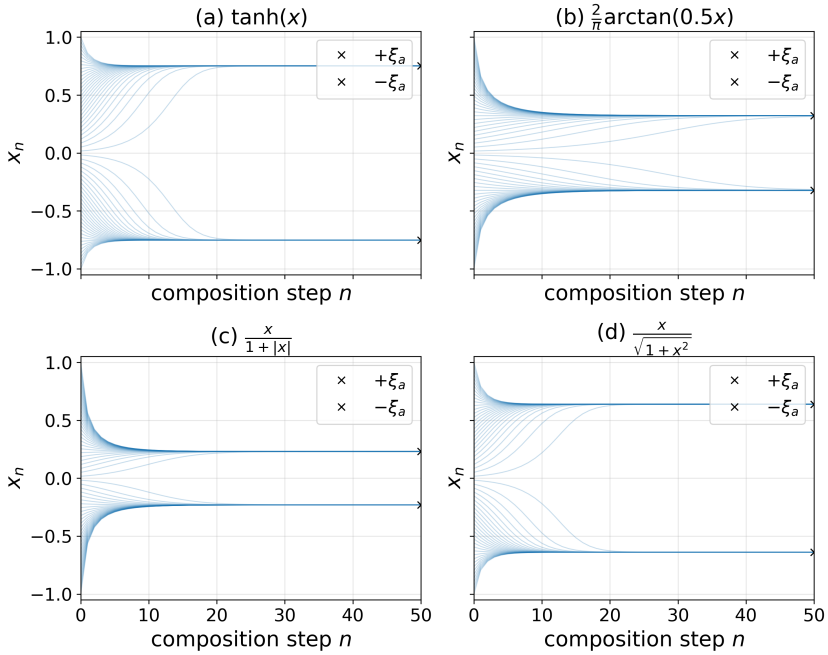


Figure 8: Iterated scalar dynamics for four odd-sigmoid activations under a fixed supercritical gain. For each activation f, we compute $\omega = 1/f'(0)$ and set $a = \omega + 0.3$. We then iterate the one–dimensional map $x_{n+1} = f(ax_n)$ for $n = 0, \ldots, 50$ starting from 60 initial values $x_0 \in [-1, 1]$. The limiting fixed points $\pm\xi_a$ satisfying $f(a\xi_a) = \xi_a$ are approximated by iterating from ± 1 and are marked at $n = 50$ with "$\times$".

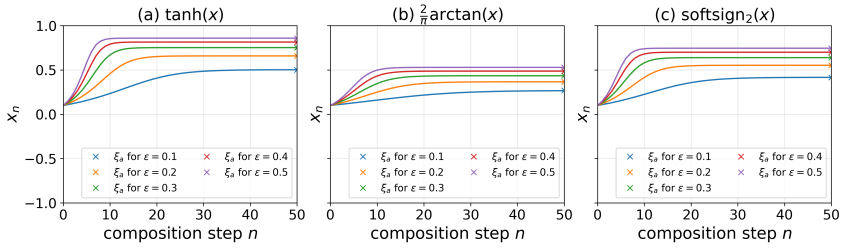


Figure 9: Convergence of the iteration $x_{n+1} = f(ax_n)$ for odd–sigmoid f with $a = \omega + \epsilon$ ($\epsilon \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$); curves show x_n up to $n = 50$ from $x_0 = 0.1$, and '$\times$' marks the positive fixed point ξ_a solving $f(a\xi_a) = \xi_a$.

B.3 Sign Flip Rate Function

We calibrate the noise scale σ_z in the surrogate model by targeting a desired sign flip rate at a specified depth. For the scalar recursion with gains $A_j \sim \mathcal{N}(\omega, \sigma^2)$ and depth L, let $\pi_L(\sigma)$ denote

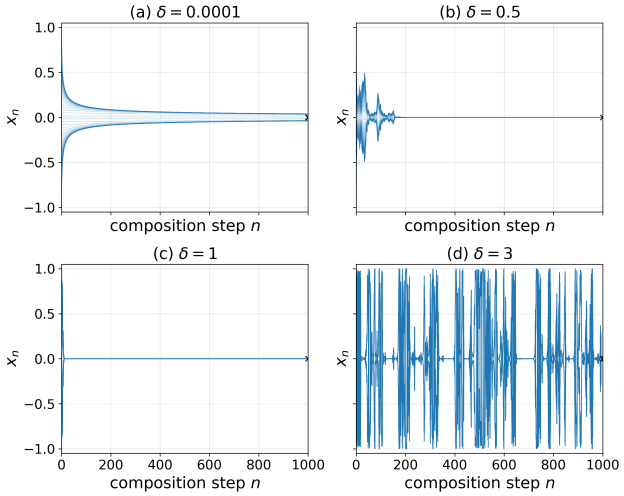


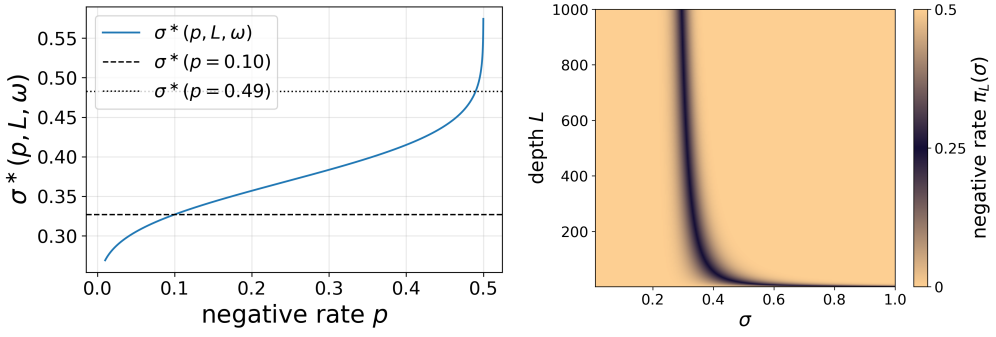
Figure 10: Iterated dynamics of the $\tanh$ activation under randomly varying gains. For each panel, we fix a sequence $(a_1, \ldots, a_{1000})$ sampled i.i.d. from $\mathcal{U}[1-\delta,\ 1+\delta]$ and iterate $x_{n+1} = \tanh(a_{n+1}x_n)$ from 60 initial points $x_0 \in [-1, 1]$. Panels (a)–(d) correspond to $\delta = 0.0001,\ 0.5,\ 1,\ 3$.

the probability that the iterate is negative at layer L. Given a target $p \in [0, \frac{1}{2})$, Theorem 4.7 yields a unique $\sigma^*(p, L, \omega)$ such that $\pi_L(\sigma^*) = p$. Figure 11 illustrates this calibration: panel (a) shows the closed-form scale $\sigma^*(p, L, \omega)$ as a function of p for $L = 100$ and $\omega = 1$, while panel (b) plots $\pi_L(\sigma)$ over (L, σ) for $\omega = 1$. As expected, $\pi_L(\sigma) \to 0$ as $\sigma \to 0$ (no sign flips) and $\pi_L(\sigma) \to \frac{1}{2}$ as $\sigma \to \infty$ (full symmetry), so the sign flip rate traces a narrow band between these two extremes.

Figure 12 evaluates the resulting calibration using the local Jacobian diagnostic of Appendix A.3. For $f(x) = \tanh(x)$ and $\omega = 1$, we choose $\sigma_z = \sigma^*(p, L, \omega)$ for target sign-flip rates $p = 0.01$ and $p = 0.49$, and estimate the local mean-square Jacobian diagnostic

$$\chi_\ell \approx (\omega^2 + \sigma_z^2)\bar{d}_\ell, \qquad \bar{d}_\ell = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f'(h_i^\ell)^2].$$

For both target rates, the estimated factors remain close to 1 over the evaluated depths, up to $L = 2 \times 10^5$.



(a) $L = 100$ (b) $L = 1, 2, \ldots, 1000$

Figure 11: Closed-form characterization of the scalar surrogate. **(a)** Closed form scale $\sigma^*(p, L, \omega)$ as a function of the target sign flip rate p for $L = 100$ and $\omega = 1$, with reference lines at $p = 0.10$ and $p = 0.49$. **(b)** Closed form scalar surrogate sign flip rate $\pi_L(\sigma)$ for $\omega = 1$, network depths $L = 1, 2, \ldots, 1000$, and $\sigma \in [0.01, 1.0]$.

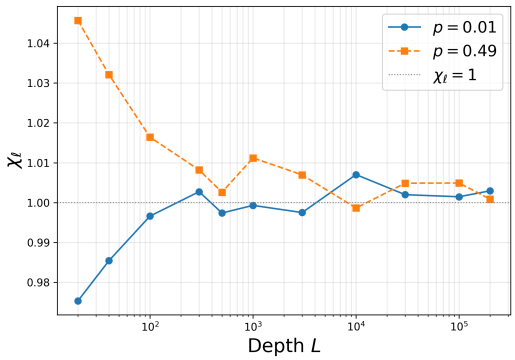


Figure 12: χ_ℓ for the proposed initialization with $f(x) = \tanh(x)$, evaluated at the closed form noise scales $\sigma^*(p, L, \omega)$ corresponding to target sign flip rates $p = 0.01$ and $p = 0.49$. For each depth ℓ, we estimate $\chi_\ell(\sigma_z) \approx (\omega^2 + \sigma_z^2)\,\mathbb{E}[f'(h^\ell)^2]$.

B.4 SIGN FLIP RATE SURROGATE ANALYSIS

We begin from the elementwise representation of a fully connected network with odd–sigmoid activation $f \in \mathcal{F}$. For each layer ℓ and neuron index i, the forward update can be written as

$$x_i^{\ell+1} = f\big(a_i^{\ell+1}\, x_i^\ell\big)\,, \qquad a_i^{\ell+1} = w_{ii}^{\ell+1} + \sum_{j\neq i} \frac{w_{ij}^{\ell+1} x_j^\ell}{x_i^\ell}, \tag{8}$$

so that $a_i^{\ell+1}$ plays the role of an effective scalar gain for neuron i in layer $\ell+1$. Under our proposed initialization $W^{\ell+1} = D^{\ell+1} + Z^{\ell+1}$ with $(D^{\ell+1})_{ij} = \omega$ on the wrapped diagonal and $(Z_{ij}^{\ell+1}) \sim \mathcal{N}(0, \sigma_z^2/N_\ell)$, Lemma 4.5 shows that, conditional on x^ℓ,

$$a_i^{\ell+1} \;\sim\; \mathcal{N}\!\left(\omega,\; \frac{\sigma_z^2}{N_\ell}\Big(1 + \sum_{j\neq i}\Big(\frac{x_j^\ell}{x_i^\ell}\Big)^2\Big)\right). \tag{9}$$

To handle this dependence on the current activations, we approximate the gain distribution by a scalar surrogate model $a_i^{\ell+1} \sim \mathcal{N}(\omega, \sigma_z^2)$ and estimate sign statistics from this Gaussian approximation.

We define the sign flip rate at depth L as

$$\pi_L(\sigma) \;:=\; \mathbb{P}\big(x_L < 0 \,\big|\, x_0 > 0\big),$$

that is, the probability that a neuron whose initial activation is positive ends up with a negative sign at layer L. Because every $f \in \mathcal{F}$ is odd and strictly increasing, the sign of x_i^ℓ is entirely controlled by the product of the effective gains along the path, and the probability that a positive entry becomes negative at depth L equals the probability that a negative entry becomes positive. It therefore suffices to track a single scalar chain starting from $x_0 > 0$.

Our initialization has a nonzero mean gain ω, which makes sign changes along a given coordinate well defined across layers. We interpret frequent sign flips during the forward pass as a form of information loss, and we therefore use the surrogate calibration to control the sign flip rate at the final layer. In the scalar surrogate, for a given depth L and gain variance σ^2 we can compute $\pi_L(\sigma)$ in closed form. We then define, for a target $p \in (0, \frac{1}{2})$, the calibrated scale $\sigma^*(p, L, \omega)$ as the unique solution to $\pi_L(\sigma^*) = p$ (see Theorem 4.7). In practice, we choose a desired "real" sign flip rate $p_{\text{real}} = 0.4$ and select σ_z so that the empirical FFNN-driven sign flip rate at depth L is close to p_{real}, which retains positive coordinate-wise sign correlation while still leaving enough randomness for learning.

To understand how the scalar surrogate relates to the actual network, we compare $\pi_L(\sigma^*)$ with the empirical sign flip rate obtained from an initialized FFNN. For each p_{target} and depth L, we first compute $\sigma^*(p_{\text{target}}, L, \omega)$, initialize an L-layer $\tanh$ network with weights $W^\ell = \omega I + (\sigma^*/\sqrt{N})G^\ell$,

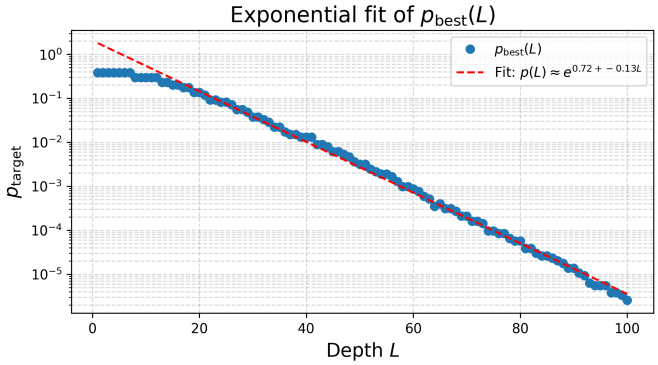
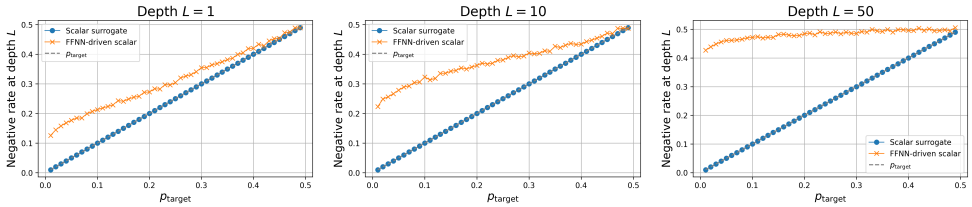


Figure 13: Calibrated surrogate target $p_{\text{sur}}(L)$ producing an FFNN-driven sign flip rate $\tilde{\pi}_L \approx 0.4$ as a function of depth L for a $\tanh$ network with width 64. The curve stays near 0.3–0.4 for shallow depths and then decays approximately exponentially with L; the dashed line is the fit $p_{\text{sur}}(L) \approx 2.05\, e^{-0.133L}$ used in Section 4.2.



(a) $L = 1$ (b) $L = 10$ (c) $L = 50$

Figure 14: sign flip rate at depth $L = 1, 10, 50$ as a function of the target sign flip rate p_{target}. The curve $\pi_L(\sigma^\star)$ denotes the scalar surrogate prediction at the calibrated scale $\sigma^\star(p_{\text{target}})$, $\tilde{\pi}_L$ is the empirical sign flip rate obtained from the FFNN-driven scalar chains, and the dashed line shows the identity p_{target}.

extract the effective gains a_i^ℓ, and form FFNN-driven scalar chains $y_0 = 1$, $y_{\ell+1} = \tanh(a^\ell y_\ell)$. Measuring $\tilde{\pi}_L = \mathbb{P}(y_L < 0)$ and comparing it to the analytic $\pi_L(\sigma^*)$ reveals that the discrepancy grows with depth (see Figure 14), as expected from the increasing influence of higher-order correlations. Nevertheless, the relationship between the surrogate target p_{target} and the FFNN-driven sign flip rate $\tilde{\pi}_L$ remains systematic.

In particular, for each depth L we can invert this relationship numerically: we search over p_{target} and find the value $p_{\text{sur}}(L)$ such that the FFNN-driven sign flip rate $\tilde{\pi}_L$ is as close as possible to $p_{\text{real}} = 0.4$ (see Section B.7).

The resulting sequence $p_{\text{sur}}(L)$ is reported in Figure 13. For small depths L, the surrogate target remains near 0.3–0.4, but for larger L it decays approximately exponentially. Fitting a simple model $\log p_{\text{sur}}(L) \approx \log c - \kappa L$ on $L \in [10, 100]$ yields

$$p_{\text{sur}}(L) \;\approx\; c\, e^{-\kappa L}, \qquad c \approx 2.05, \;\; \kappa \approx 0.133,$$

matching the schedule of Section 4.2. This empirical law expresses how aggressively the surrogate target sign flip rate must shrink with depth in order for the actual network to maintain a fixed, moderate sign flip rate at its final layer.

Origin of the exponential form. The exponential decay of $p_{\text{sur}}(L)$ has a near-critical scaling explanation. If the centered layerwise perturbations of the effective gain accumulate diffusively, a non-degenerate cumulative perturbation at depth L requires $L\sigma_z^2 \approx s$ for a depth-independent constant $s > 0$.

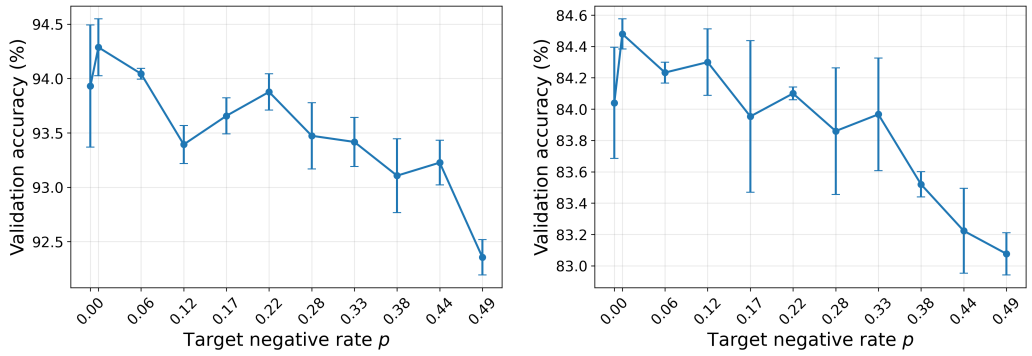
Proposition B.1. *Let $\omega > 0$ and $\sigma_z^2 = s/L$ for a constant $s > 0$. With $q_L = \Phi(-\omega/\sigma_z)$ and $p_{\mathrm{sur}}(L) = \frac{1}{2}\big(1 - (1-2q_L)^L\big)$ as in Lemma A.4, as $L \to \infty$,*

$$p_{\mathrm{sur}}(L) = C\sqrt{L}\, e^{-\kappa L}\,(1 + o(1)), \qquad \kappa = \frac{\omega^2}{2s}, \qquad C = \frac{\sqrt{s}}{\omega\sqrt{2\pi}}.$$

Proof. Set $x_L := \omega/\sigma_z = \omega\sqrt{L/s} \to \infty$. Mills' ratio gives $\Phi(-x_L) = \frac{e^{-x_L^2/2}}{x_L\sqrt{2\pi}}(1 + o(1))$, hence $q_L = \frac{\sqrt{s}}{\omega\sqrt{2\pi L}}\, e^{-\omega^2 L/(2s)}(1 + o(1))$. Since $Lq_L \to 0$, $(1-2q_L)^L = 1 - 2Lq_L + O(L^2 q_L^2)$, so $p_{\mathrm{sur}}(L) = Lq_L\,(1 + o(1))$. □

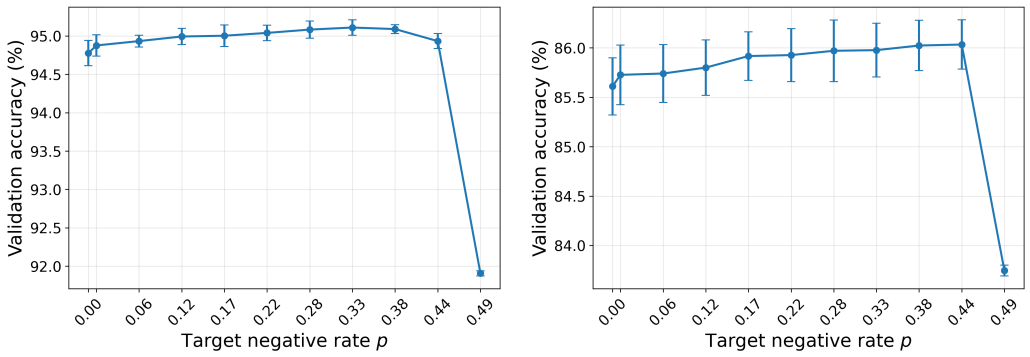
Hence $\log p_{\mathrm{sur}}(L) = -\kappa L + \frac{1}{2}\log L + O(1)$, which is the near-linear semi-log behavior of Figure 2(b); over the fitted depth range the slowly varying $\sqrt{L}$ factor is absorbed by the pure exponential fit. The fitted schedule is quantitatively consistent with this picture: for $\omega = 1$, inserting $p_{\mathrm{sur}}(L) = 2.05\, e^{-0.133L}$ into Theorem 4.7 gives $L\sigma_z^2 \in [3.30, 3.69]$ for $20 \le L \le 10^3$, close to $1/(2\kappa) \approx 3.76$. Equivalently, for $L \ge 20$ the calibration reduces to the one-constant rule $\sigma_z^2 \approx s\,\omega^2/L$. Lemma 4.5 provides the finite-width counterpart: the conditional gain variance is $\sigma_z^2 R_\ell(\sigma_z; i)$, so when the coordinates remain comparable ($R_\ell = O(1)$) the same $1/L$ scaling is preserved, while finite-width correlations enter the fitted constants. This argument explains the functional form of the schedule and reduces it to a single constant; it is not a finite-width derivation, and s remains empirically determined. For $L \gtrsim 280$ the target satisfies $p(L) < 10^{-16}$, so $(1-2p)^{1/L}$ rounds to 1 in double precision; for such p we use $q = p/L$, which approximates $\frac{1}{2}\big(1 - (1-2p)^{1/L}\big)$ with relative error $O(p)$.

These calibration insights are consistent with our training experiments. Figures 15 and 16 plot validation accuracy as a function of the surrogate target p for deep ($L = 50$) and shallow ($L = 3$) networks. For the deep case, performance is maximized when p_{sur} is on the order of 0.01, in line with the exponentially decayed target predicted by the surrogate analysis. In contrast, for shallow networks the best performance is attained near $p_{\mathrm{sur}} \approx 0.4$, matching the regime where the fitted decay has not yet taken effect. Finally, Figures 17 and 18 show that, for depths $L = 50$ and $L = 100$, validation accuracy peaks at the grid point closest to the calibrated σ^* ($\sigma^* \approx 0.26$ and 0.19, i.e. entrywise 0.032 and 0.023), with accuracy degrading at $\sigma_z = 0$ and at twice the calibrated scale. Together, these results validate that the sign-flip surrogate provides a useful, quantitatively accurate guideline for choosing the noise level σ_z across depths, and that the resulting diagonal–plus–noise initialization indeed preserves signal sign statistics in deep networks.



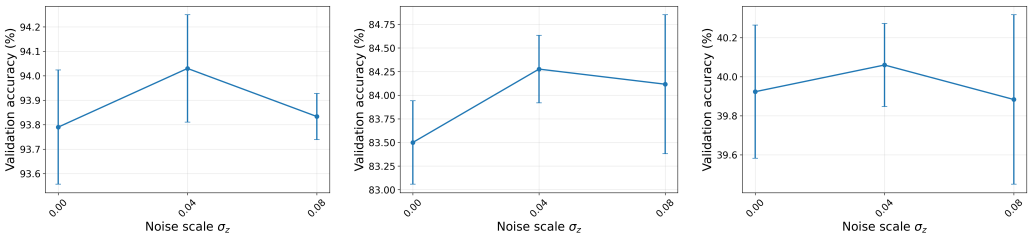
(a) MNIST (b) FMNIST

Figure 15: Validation accuracy as a function of the target sign flip rate p for a 50-layer fully connected network (width 64) with activation $f(x) = \tanh(x)$ under the proposed initialization. Each point shows the mean ± standard deviation over 5 runs, where each run is trained for 600 iterations, with $\sigma^*(p, L, \omega)$ computed from the scalar surrogate calibration.



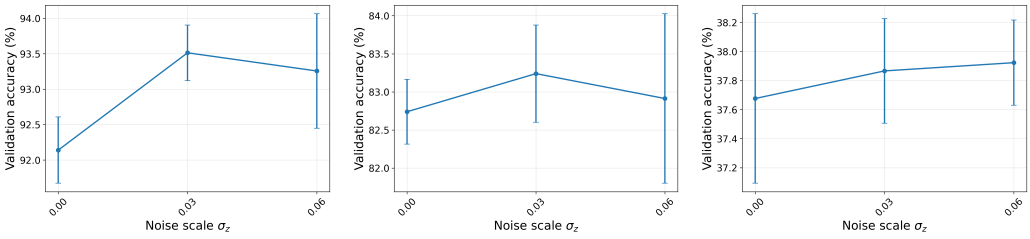
(a) MNIST (b) FMNIST

Figure 16: Validation accuracy as a function of the target sign flip rate p for a 3-layer fully connected network (width 512) with activation $f(x) = \tanh(x)$ under the proposed initialization. Each point shows the mean $\pm$ standard deviation over 5 runs, where each run is trained for 600 iterations, with $\sigma^*(p, L, \omega)$ computed from the scalar surrogate calibration.



(a) MNIST (b) FMNIST (c) CIFAR-10

Figure 17: Validation accuracy as a function of the noise scale σ_z for 50 layer fully connected networks (width 64) with activation $f(x) = \tanh(x)$ under the proposed initialization, on MNIST, Fashion MNIST, and CIFAR-10. For each $\sigma_z \in \{0, 0.32, 0.64\}$ (entrywise standard deviation $\sigma_z/\sqrt{N_\ell} \in \{0, 0.04, 0.08\}$ with $N_\ell = 64$), we report the mean $\pm$ standard deviation over 5 training runs. The schedule of Section 4.2 gives $\sigma_z \approx 0.26$ at $L = 50$.



(a) MNIST (b) FMNIST (c) CIFAR-10

Figure 18: Validation accuracy as a function of the noise scale σ_z for 100 layer fully connected networks (width 64) with activation $f(x) = \tanh(x)$ under the proposed initialization, on MNIST, Fashion MNIST, and CIFAR-10. For each $\sigma_z \in \{0, 0.24, 0.48\}$ (entrywise standard deviation $\sigma_z/\sqrt{N_\ell} \in \{0, 0.03, 0.06\}$ with $N_\ell = 64$), we report the mean $\pm$ standard deviation over 5 training runs. The schedule of Section 4.2 gives $\sigma_z \approx 0.19$ at $L = 100$.

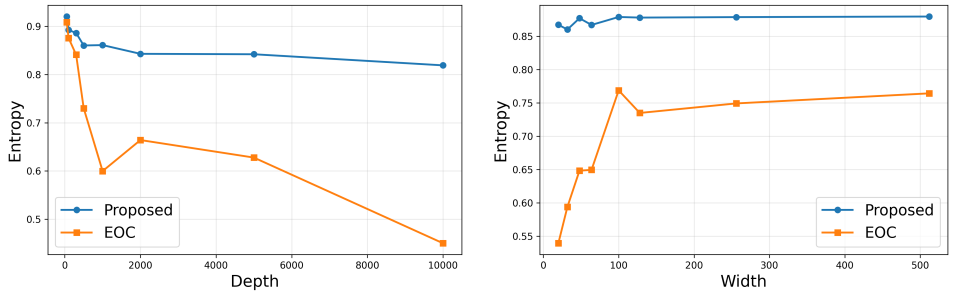
B.5 FORWARD SIGNAL PROPAGATION

In Section 4.3 we study forward signal propagation under the proposed initialization. Figure 19 (a) reports how well the last-layer activation distribution is preserved as the depth L increases (with fixed width $W = 64$), while panel (b) varies the width N_ℓ at fixed depth $L = 1000$ to test whether the distribution remains well spread even in relatively narrow networks. In both settings, the proposed initialization maintains a dispersed last-layer distribution up to depth $L = 10{,}000$ and down to width $N_\ell = 20$, whereas the EOC (Gaussian i.i.d.) initialization quickly collapses toward a more concentrated distribution.

To summarize how well the last-layer distribution is dispersed, we employ a normalized histogram-entropy score

$$\mathrm{Spread}(x) = \frac{-\sum_{i=1}^{B} p_i \log p_i}{\log B}, \qquad p_i = \int_{\text{bin } i} \hat{f}_x(t)\, dt, \quad \sum_{i=1}^{B} p_i = 1, \tag{10}$$

where $\hat{f}_x$ is the empirical density over $[-1, 1]$ using B bins. Values close to 1 indicate a highly dispersed (near-uniform) last-layer distribution, whereas values near 0 correspond to a highly concentrated distribution.



(a) Depth (b) Width

Figure 19: Entropy of the last-layer activation distribution for tanh networks under the proposed initialization and the EOC initialization. **(a)** Entropy as a function of depth L with fixed width $W = 64$. **(b)** Entropy as a function of width N_ℓ with fixed depth $L = 1000$, using widths $W \in \{20, 32, 48, 64, 100, 128, 256, 512\}$ as shown in the panels. The entropy is defined in Equation 10.

Figures 20, 21, and 22 visualize the last-layer activation histograms obtained after the initial forward pass in 1000-layer FFNNs of width 64 with activations $0.1\tanh(x)$, $\tanh(x)$, and $10\tanh(x)$, respectively. Under the proposed initialization, the activation distribution remains well dispersed in all three cases, whereas Gaussian i.i.d. initializations quickly collapse toward a narrow band around zero. Figures 23 and 24 show a similar comparison as the width is reduced: even for very narrow networks, the proposed scheme preserves a spread-out last layer distribution, while Gaussian initializations drive the last layer activations to saturate near zero.

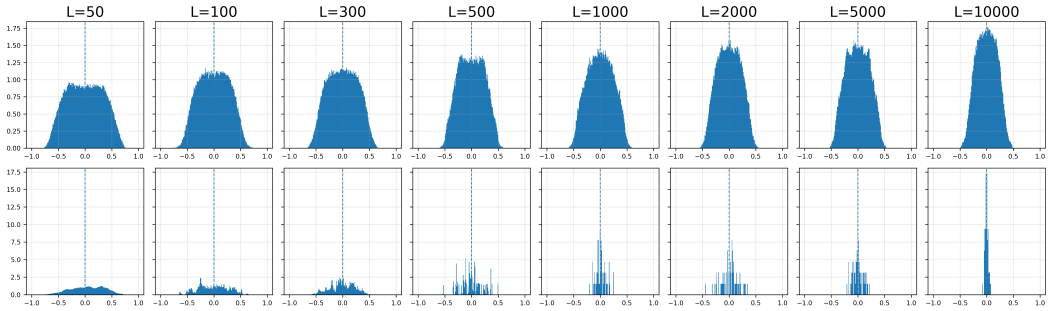


Figure 20: Last layer activation histograms for $\tanh$ networks with width $W = 64$ and varying depth under the proposed initialization (top row) and the EOC initialization (bottom row). Each column corresponds to a different depth L.

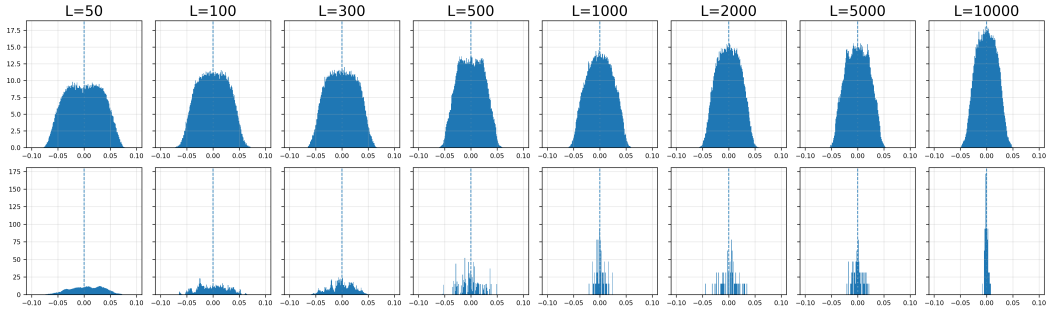


Figure 21: Last layer activation histograms for $0.1\tanh$ networks with width $W = 64$ and varying depth under the proposed initialization (top row) and the EOC initialization (bottom row). Each column corresponds to a different depth L.

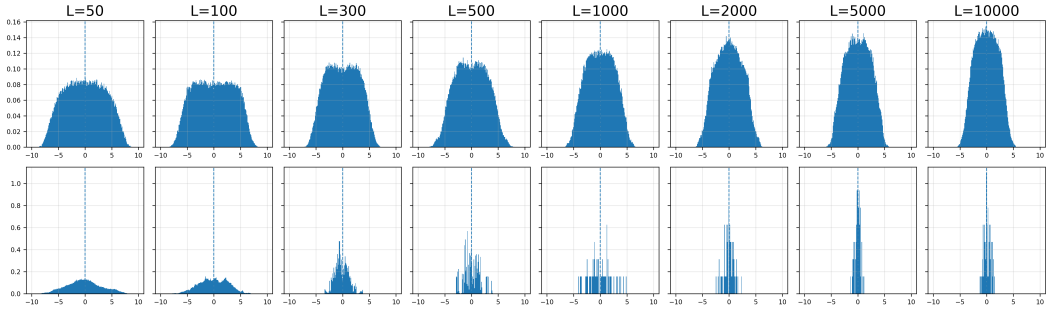


Figure 22: Last layer activation histograms for $10\tanh$ networks with width $W = 64$ and varying depth under the proposed initialization (top row) and the EOC initialization (bottom row). Each column corresponds to a different depth L.

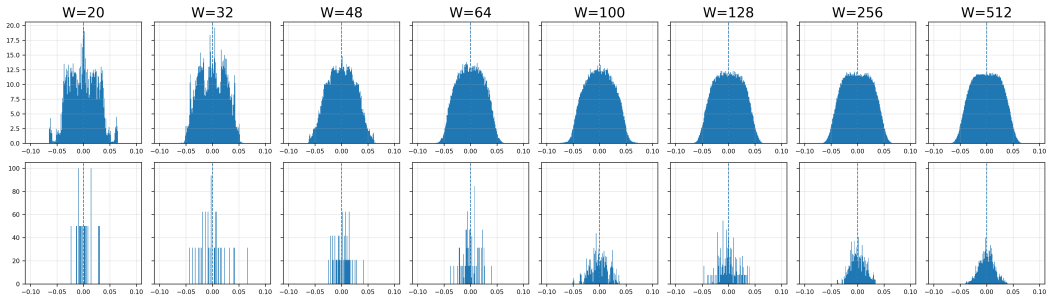


Figure 23: Last layer activation histograms for $0.1\tanh$ networks with depth $L = 1000$ and varying width under the proposed initialization (top row) and the EOC initialization (bottom row).

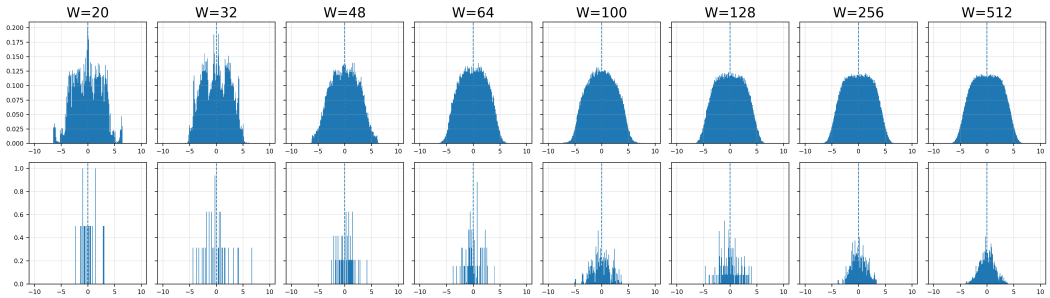


Figure 24: Last layer activation histograms for $10\tanh$ networks with depth $L = 1000$ and varying width under the proposed initialization (top row) and the EOC initialization (bottom row).

B.6 BACKWARD SIGNAL PROPAGATION

Figure 25 visualizes $|\chi_L(\sigma) - 1|$ for the proposed initialization and a Gaussian initialization with $\tanh$ activations, as a function of depth L and scale σ. For the proposed scheme, the region where $\chi_L(\sigma) \approx 1$ occupies a broad band in the (L, σ) plane, whereas for the Gaussian initialization it is confined to a narrow strip around a single variance. This indicates that our initialization keeps χ_L close to 1 over a much wider range of noise scales and depths, and is therefore more robust and better suited for stable training in deep, narrow networks.

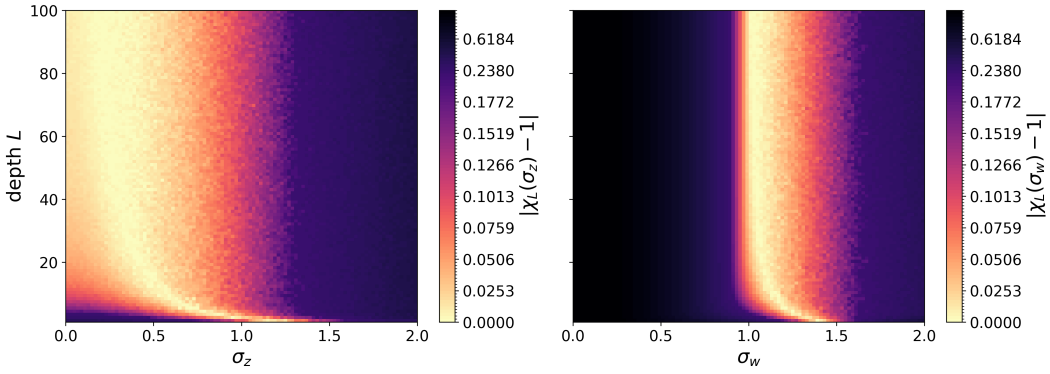


Figure 25: Heatmaps of the deviation $|\chi_L(\sigma) - 1|$ for the proposed initialization **(left)** and a Gaussian i.i.d. initialization **(right)** with $\tanh$ activations, as a function of depth L and scale σ. Brighter bands indicate near critical regimes where forward and backward signals are approximately preserved.

B.7 JUSTIFICATION OF THE TARGET SIGN-FLIP RATE $p_{\mathrm{real}} = 0.4$

Our choice $p_{\mathrm{real}} = 0.4$ is motivated by coordinate-wise sign agreement and supported by the classification experiments below; we do not claim it is information-theoretically optimal.

Sign-based intuition. The sign-flip rate $\tilde{\pi}_L$ measures the probability that the sign of a coordinate at depth L differs from that of its input. As an illustration, if input signs are balanced and their propagation is modeled as a binary symmetric channel (BSC) with crossover probability $\tilde{\pi}_L$, the mutual information between the two signs is

$$1 - H_2(\tilde{\pi}_L) \quad \text{bits per coordinate,}$$

where $H_2(p) := -p \log_2 p - (1-p)\log_2(1-p)$. This formula describes the idealized BSC, not the mutual information of the full network representations. In particular, a rate near zero means that most coordinate signs are preserved; it does not mean that the network is an identity map or lacks nonlinear expressivity. A rate near $1/2$ means that coordinatewise sign agreement is weak, but does not establish independence of the full input and output representations. At $\tilde{\pi}_L = 0.4$, approximately 60% of coordinate signs are preserved and 40% flip. We use this intermediate regime to retain a visible link to the input sign pattern while allowing substantial noise-induced sign changes. Under the idealized BSC formula its mutual information is approximately 0.029 bits per coordinate; that number is not the basis for selecting 0.4.

Empirical support. Figure 6 compares FFNN-level targets in width-64 $\tanh$ classifiers. Across the reported datasets at depths 50 and 100, targets around 0.3–0.4 give the best or near-best early training accuracy. Figures 17 and 18 separately examine validation accuracy as the noise scale varies, while Figures 15 and 16 sweep the surrogate target rather than the FFNN-level target. We therefore use $p_{\mathrm{real}} = 0.4$ as a fixed default for the reported classification settings and calibrate σ_z so that the empirical terminal FFNN sign-flip rate is close to this target. This is a practically useful operating choice in the tested settings, not a universal optimum.

C ADDITIONAL EXPERIMENTAL RESULTS WITH NEURAL NETWORKS

Experimental Setting. We evaluate the proposed initialization using the Adam optimizer with a batch size of 128 and reserve 15% of the training data for validation. All experiments are implemented in PyTorch without skip connections and without learning rate decay. Unless stated otherwise, every initializer (Proposed, Xavier, He, EOC, and their BN variants) uses the same activation-dependent rule $\eta = 10^{-4}\omega$ with $\omega = 1/f'(0)$, so that no method receives a separately tuned learning rate. For activations with $\omega \ll 1$ this rule does not disadvantage the Gaussian baselines: the full-grid sweeps in Figures 35–38 and the per-initializer grid search in Tables 2–7 show that they fail at every learning rate tested when $\omega \ll 1$.

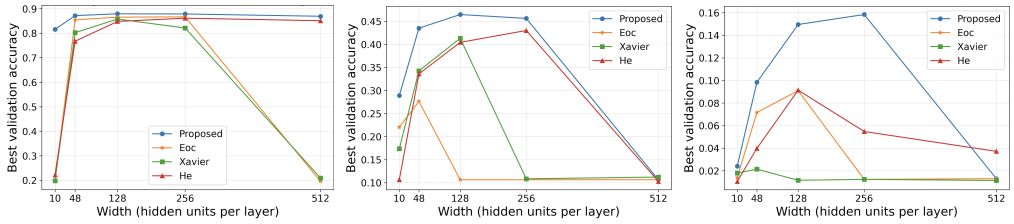
C.1 DATASET EFFICIENCY

Table 1: Mean of the best validation accuracies within 50 epochs over 10 independent runs for a 50 layer FFNN (512 units) on MNIST and Fashion MNIST, trained on 100 or 500 sample subsets.

Dataset	Method	$\tanh(x)$		$\mathrm{erf}(x)$		$\arctan(x)$		$\mathrm{gd}(x)$		$\mathrm{softsign}_2(x)$		$\mathrm{softsign}_1(x) + \mathrm{softsign}_2(x)$	
		100	500	100	500	100	500	100	500	100	500	100	500
MNIST	Xavier	64.00	79.85	66.15	84.63	64.58	81.30	60.87	84.18	60.52	81.38	32.63	53.97
	He	41.65	72.55	21.05	48.37	51.40	77.68	42.50	72.75	48.35	76.18	11.35	10.02
	EOC	59.83	82.92	64.58	82.97	64.52	83.73	62.30	84.05	66.57	82.48	59.60	71.57
	Proposed	**68.23**	**86.75**	**66.53**	**87.13**	**67.63**	**86.82**	**66.38**	**86.23**	**69.51**	**86.43**	**65.65**	**84.02**
FMNIST	Xavier	67.65	74.53	68.92	76.13	66.75	73.78	67.10	72.65	66.85	74.47	49.38	69.88
	He	61.13	74.60	44.60	65.25	62.92	76.75	62.88	76.55	64.97	75.22	11.50	11.70
	EOC	67.22	76.87	66.85	73.83	66.93	76.45	67.00	77.63	67.08	75.67	65.48	73.07
	Proposed	**70.67**	**77.43**	**70.63**	**78.17**	**71.55**	**77.92**	**68.62**	**77.80**	**68.17**	**78.33**	**67.93**	**76.43**

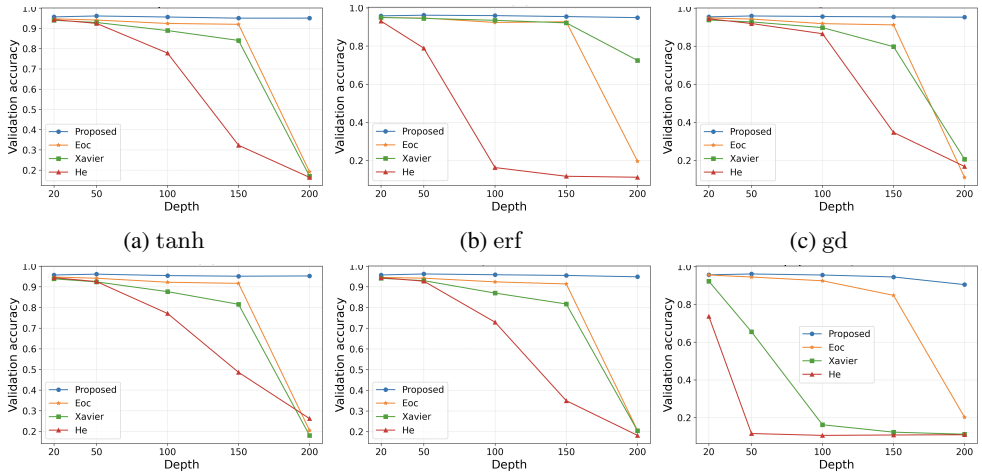
C.2 Network Size Independence

We assess how independent the proposed initialization is of depth and width in networks. Figures 27, 28, 29, and 30 show validation accuracy versus depth for FFNNs (width 64) on MNIST, Fashion MNIST, CIFAR-10, and CIFAR-100, using odd sigmoid activations and four initializations. Each panel fixes one activation and shows the best validation accuracy over 10 epochs for depths $L \in \{20, 50, 100, 150, 200\}$ (MNIST, Fashion MNIST, CIFAR-10) and $L \in \{20, 30, 50, 80, 100\}$ (CIFAR-100). Across all datasets, activations, and depths, the proposed initialization consistently achieves the highest validation accuracy among all Gaussian initializations. Figures 26 and 34 further examine the effect of width in deep networks and show that standard Gaussian initializations struggle when the network is either too narrow or too wide, whereas the proposed scheme maintains strong performance over a broad range of widths.



(a) Fashion MNIST (b) CIFAR 10 (c) CIFAR 100

Figure 26: Best validation accuracy versus width $\{10, 48, 128, 256, 512\}$ for a 100 layer $\tanh$ FFNN. Each curve shows the Proposed, EOC, Xavier, and He initialization schemes, and each point corresponds to the best validation accuracy over 20 training epochs.

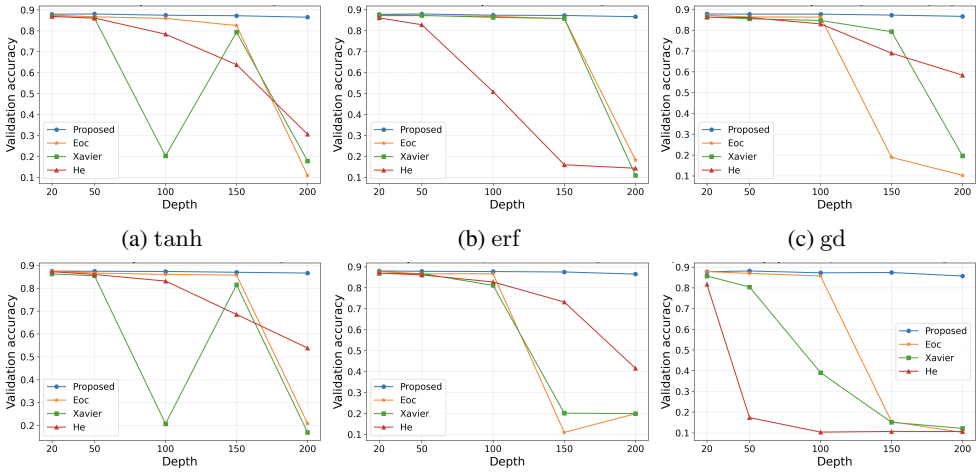


(a) $\tanh$ (b) erf (c) gd

(d) arctan (e) softsign_2 (f) $\text{softsign}_1 + \text{softsign}_2$

Figure 27: MNIST validation accuracy versus depth for FFNNs (width 64) with odd-sigmoid activations and four initializations (Proposed, EOC, Xavier, He). Each panel fixes one activation and shows the best validation accuracy over 10 epochs for depths $L \in \{20, 50, 100, 150, 200\}$.

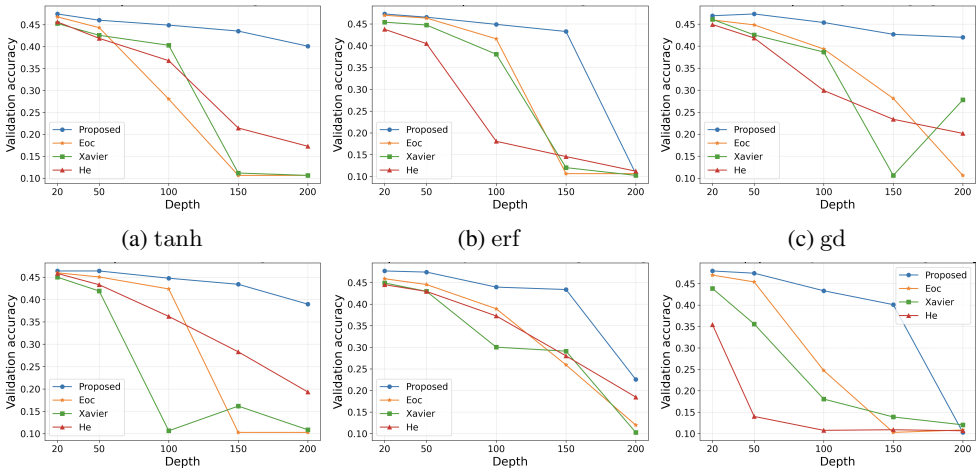
Figures 27, 28, 29, and 30 report the validation accuracy as a function of depth for networks with six different activation functions. On the more complex CIFAR-10 and CIFAR-100 datasets, even the proposed initialization exhibits some change in performance once the depth becomes sufficiently large, whereas on MNIST and Fashion-MNIST the validation accuracy is essentially preserved across all depths considered. These results indicate that, across diverse datasets and for odd–sigmoid activations with effective gain parameter ω close to one, the proposed initialization behaves substantially more depth invariant than standard Gaussian initializations.



(a) tanh (b) erf (c) gd

(d) arctan (e) softsign_2 (f) $\text{softsign}_1 + \text{softsign}_2$

Figure 28: Fashion MNIST validation accuracy versus depth for FFNNs (width 64) with odd-sigmoid activations and four initializations (Proposed, EOC, Xavier, He). Each panel fixes one activation and shows the best validation accuracy over 10 epochs for depths $L \in \{20, 50, 100, 150, 200\}$.



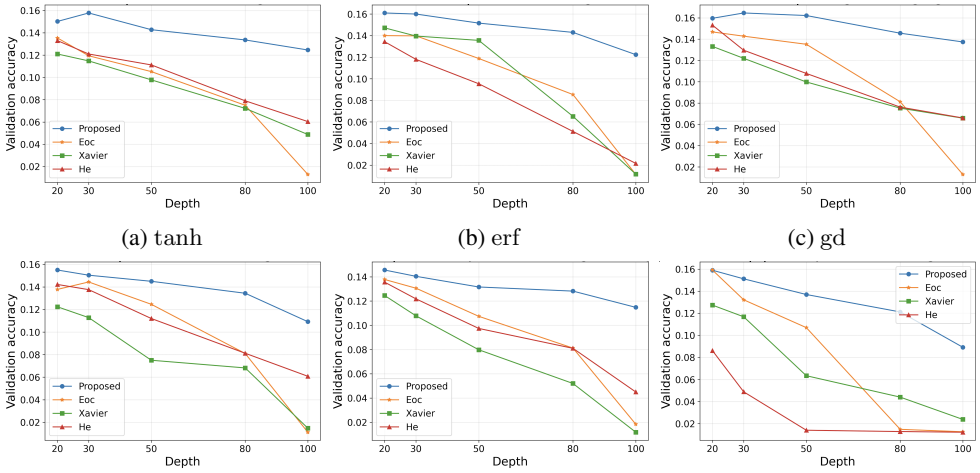
(a) tanh (b) erf (c) gd

(d) arctan (e) softsign_2 (f) $\text{softsign}_1 + \text{softsign}_2$

Figure 29: CIFAR 10 validation accuracy versus depth for FFNNs (width 64) with odd-sigmoid activations and four initializations (Proposed, EOC, Xavier, He). Each panel fixes one activation and shows the best validation accuracy over 10 epochs for depths $L \in \{20, 50, 100, 150, 200\}$.

C.3 Batch Normalization Free Training

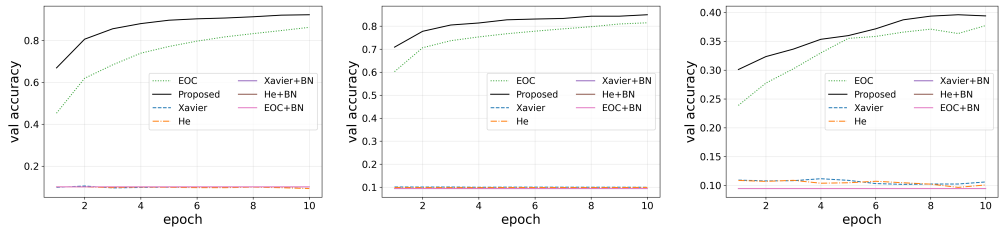
Batch normalization (BN) has made training deep networks substantially easier, but it also incurs a significant computational overhead. This raises the question of whether our initialization can match or surpass BN-based methods while avoiding this overhead. Figure 32 evaluates $\tanh$ networks and shows that, across all datasets, the proposed initialization without BN outperforms or matches competing schemes that rely on BN. Figures 31 and 33 extend this comparison to activations with $\omega = 1/f'(0)$ far from 1, including an 800 layer network with width 32. In this challenging deep and narrow regime, our method still trains reliably without BN, whereas Gaussian initializations (with or without BN) either fail to converge or achieve worse accuracy. These results suggest that the proposed initialization can substitute for BN in many settings, enabling stable training of very deep and narrow networks without the cost of normalization layers.



(a) tanh (b) erf (c) gd

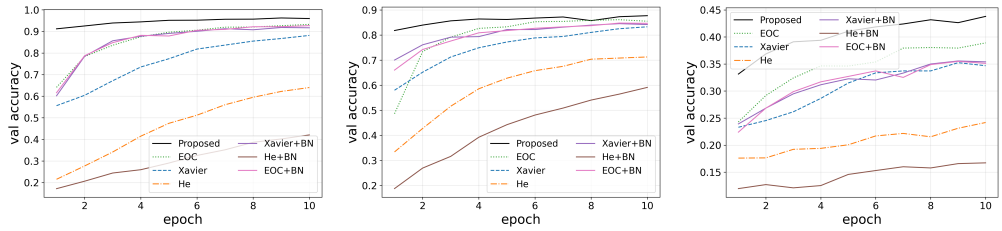
(d) arctan (e) softsign$_2$ (f) softsign$_1$ + softsign$_2$

Figure 30: CIFAR 100 validation accuracy versus depth for FFNNs (width 64) with odd-sigmoid activations and four initializations (Proposed, EOC, Xavier, He). Each panel fixes one activation and shows the best validation accuracy over 10 epochs for depths $L \in \{20, 30, 50, 80, 100\}$.



(a) MNIST (b) FMNIST (c) CIFAR10

Figure 31: Validation accuracy for the $f(x) = \tanh(ax) + \mathrm{erf}(bx) + x/(1 + |cx|) + \mathrm{gd}(dx)$ in FFNN with 100 hidden layers of width 64. We compare seven strategies: Proposed, Xavier, He, EOC, and their BN variants. Each panel uses a different dataset and a different choice of coefficients (a, b, c, d): **(a)** (10,1000,10,1), **(b)** (100,1000,10,1000), **(c)** (1000,100,0.1,0.01). Training details are given in Appendix C.



(a) MNIST (b) FMNIST (c) CIFAR 10

Figure 32: Validation accuracy over 10 epochs for fully connected 100 layer networks with $tanh$ activations and width 64. The curves compare the Proposed, EOC, Xavier, and He initializations, with and without batch normalization. Training details are given in Appendix C.

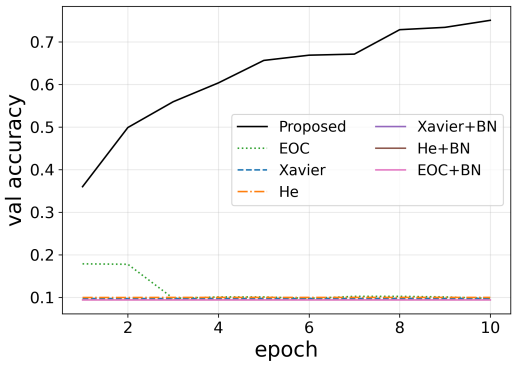


Figure 33: Validation accuracy on Fashion MNIST for a fully connected network with 800 hidden layers of width 32 and $f(x) = \tanh(ax) + \mathrm{erf}(bx) + x/(1 + |cx|) + \mathrm{gd}(dx)$, using coefficients $(a, b, c, d) = (100, 1000, 10, 1000)$. Training details are given in Appendix C.

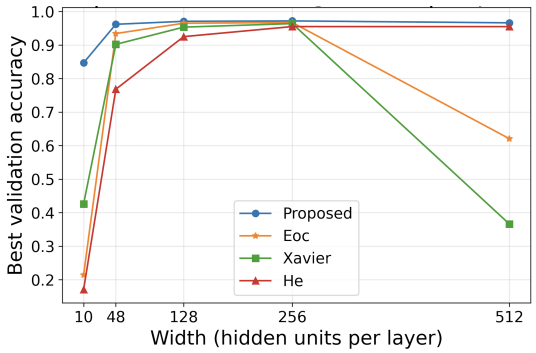
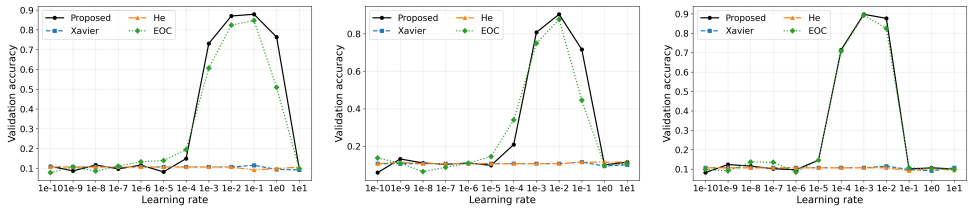


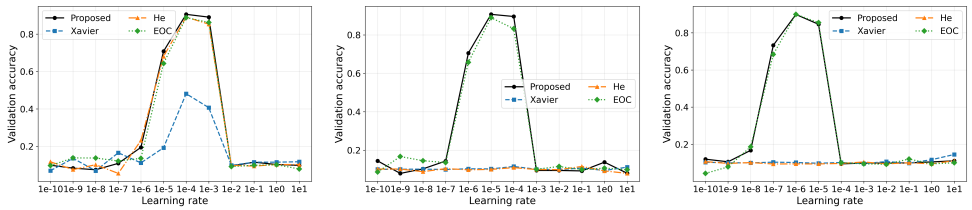
Figure 34: Best validation accuracy on MNIST versus width $\{10, 48, 128, 256, 512\}$ for 100 layer $\tanh$ FFNN. Each curve compares the Proposed, EOC, Xavier, and He initialization schemes, and each point reports the best validation accuracy over 20 training epochs.

C.4 Learnable Learning Rate

We compare the learning-rate sensitivity of Proposed, Xavier, He, and EOC in Figures 35–38. For each experimental configuration, all initializers are evaluated on the same 12-point grid $\eta \in \{10^{-10}, 10^{-9}, \ldots, 10^{0}, 10^{1}\}$, using the same architecture, optimizer, training subset, and 200-iteration budget. Since activation rescaling changes $\omega = 1/f'(0)$, we also describe the useful learning-rate range of the proposed method relative to ω. The interval $[10^{-5}\omega, 10^{-3}\omega]$ summarizes this observed operating range; all initializers are evaluated over the full grid stated above.

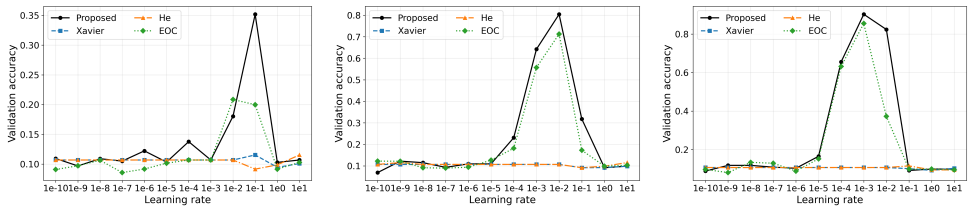


(a) $2/\pi \arctan(0.001x)$ (b) $2/\pi \arctan(0.01x)$ (c) $2/\pi \arctan(0.1x)$

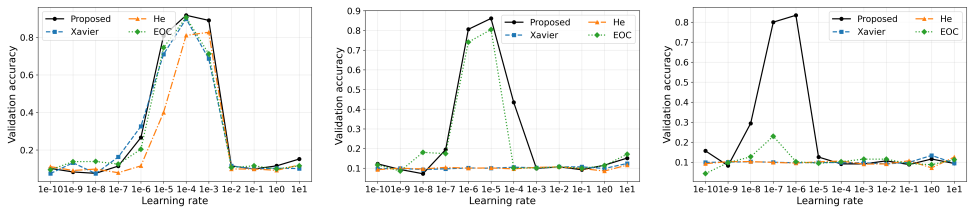


(d) $2/\pi \arctan(1x)$ (e) $2/\pi \arctan(10x)$ (f) $2/\pi \arctan(100x)$

Figure 35: Learning rate accuracy curves on MNIST for a 20 layer, width 512 feedforward network with activation $f(x) = \frac{2}{\pi}\arctan(\alpha x)$. Each panel corresponds to a different activation scale $\alpha \in \{10^2, 10^1, 1, 10^{-1}, 10^{-2}, 10^{-3}\}$. For each learning rate, we train for 200 iterations on a 10k training subset and report the validation accuracy.

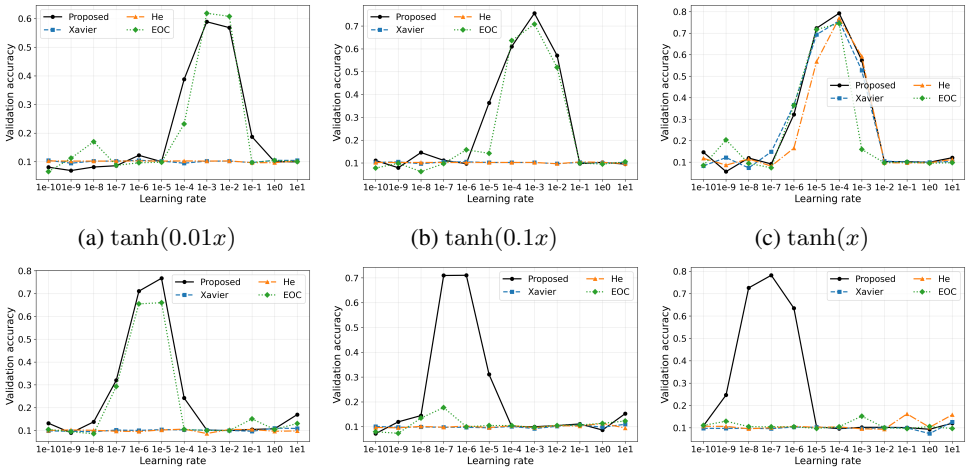


(a) $\tanh(0.001x)$ (b) $\tanh(0.01x)$ (c) $\tanh(0.1x)$



(d) $\tanh(1x)$ (e) $\tanh(10x)$ (f) $\tanh(100x)$

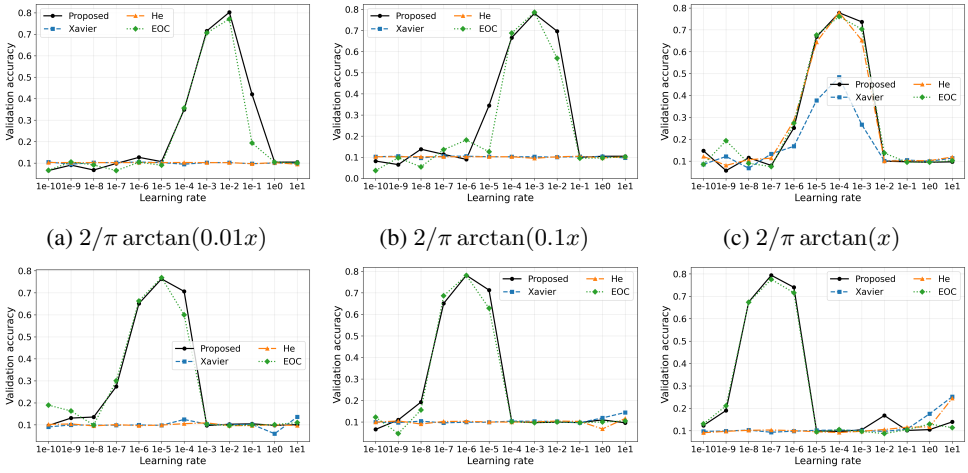
Figure 36: Learning rate accuracy curves on MNIST for a 20 layer, width 512 feedforward network with activation $f(x) = \tanh(\alpha x)$. Each panel corresponds to a different activation scale $\alpha \in \{10^2, 10^1, 1, 10^{-1}, 10^{-2}, 10^{-3}\}$. For each learning rate, we train for 200 iterations on a 1k training subset and report the validation accuracy.



(a) $\tanh(0.01x)$ (b) $\tanh(0.1x)$ (c) $\tanh(x)$

(d) $\tanh(10x)$ (e) $\tanh(100x)$ (f) $\tanh(1000x)$

Figure 37: Learning rate accuracy curves on Fashion MNIST for a 20 layer, width 512 feedforward network with activation $f(x) = \tanh(\alpha x)$. Each panel corresponds to a different activation scale $\alpha \in \{10^3, 10^2, 10^1, 1, 10^{-1}, 10^{-2}\}$. For each learning rate, we train for 200 iterations on a 10k training subset and report the validation accuracy. Curves compare four initializations: Proposed, Xavier, He, and EOC.



(a) $2/\pi \arctan(0.01x)$ (b) $2/\pi \arctan(0.1x)$ (c) $2/\pi \arctan(x)$

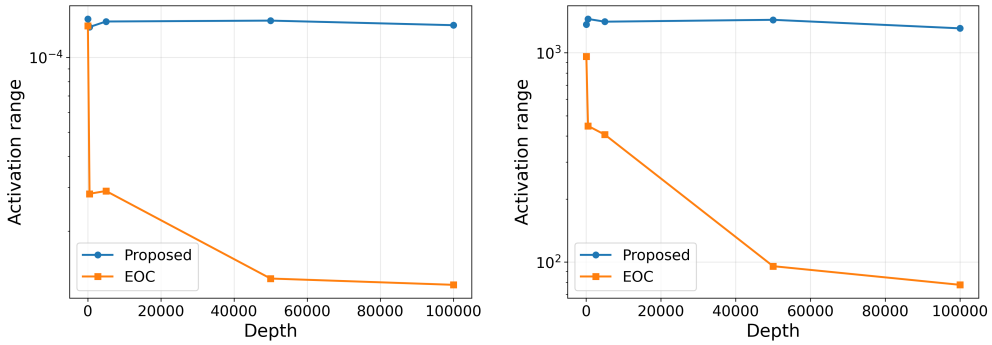
(d) $2/\pi \arctan(10x)$ (e) $2/\pi \arctan(100x)$ (f) $2/\pi \arctan(1000x)$

Figure 38: Learning rate accuracy curves on Fashion MNIST for a 20 layer, width 512 feedforward network with activation $f(x) = 2/\pi \arctan(\alpha x)$. Each panel corresponds to a different activation scale $\alpha \in \{10^3, 10^2, 10^1, 1, 10^{-1}, 10^{-2}\}$. For each learning rate, we train for 200 iterations on a 10k training subset and report the validation accuracy. Curves compare four initializations: Proposed, Xavier, He, and EOC.

C.5 Scale Preserving Odd–Sigmoid Activations.

Using a scaled activation function amounts to adjusting the effective range of both the input and output axes of the nonlinearity. For example, $\tanh(x)$ has output range $[-1, 1]$, whereas $\alpha \tanh(x)$ has range $[-\alpha, \alpha]$. With our initialization we therefore ask whether signals remain well propagated over the full interval $[-\alpha, \alpha]$ when the activation is scaled. As shown in Figure 39, for several values of α the proposed initialization keeps the last layer activations of $\alpha \tanh(x)$ spread over $[-\alpha, \alpha]$ even in very deep networks (up to $L = 10^5$), while Gaussian i.i.d. initialization rapidly drives the activations to saturate near zero. In all α-scale experiments (Tables 2–7), each initializer was given its own LR grid search. Even under this per initializer tuning, Gaussian and EOC schemes fail to train for large α, while our method remains stable.

Since the proposed initialization lets us target the output scale via the choice of α, it is natural to expect benefits on regression tasks where the range of the target y matters. To test this, we evaluate our method in physics informed neural networks (PINNs) with scaled odd–sigmoid activations on the four PDE benchmarks (Heat1D, Helmholtz2D, Poisson2D, Allen–Cahn) described in Appendix C.6.



(a) $\alpha = 0.0001$ (b) $\alpha = 1000$

Figure 39: Activation range as a function of depth $L \in \{50, 500, 5{,}000, 50{,}000, 10^5\}$ for fully connected width 64 networks with activation $\alpha \tanh(x)$ under the Proposed and EOC initializations. Panel (a) uses $\alpha = 10^{-4}$ and panel (b) uses $\alpha = 10^3$.

Table 2: Validation accuracy on MNIST (left) and Fashion MNIST (right) for a 50 layer, width 128 fully connected neural network with activation $a\tanh(x)$. Each row corresponds to a different activation scale a. For each scale, the learning rate is tuned separately for each initializer.

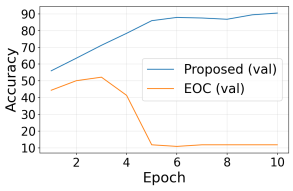
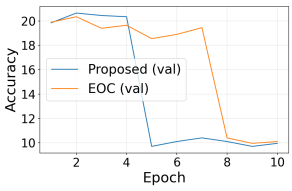
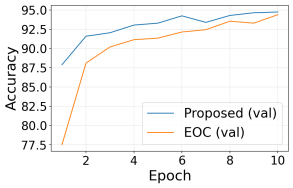
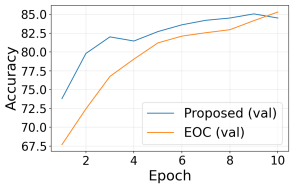
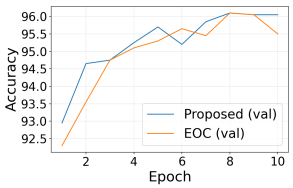
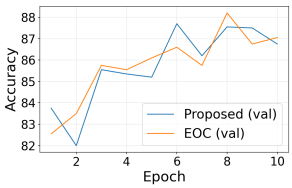
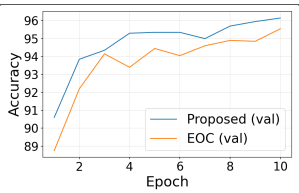
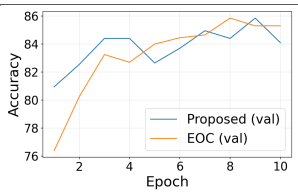
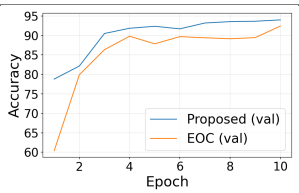
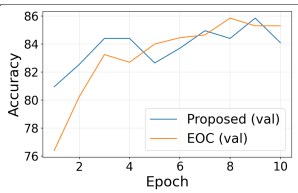
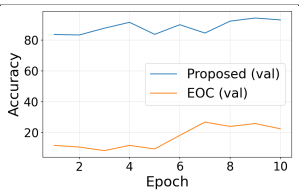
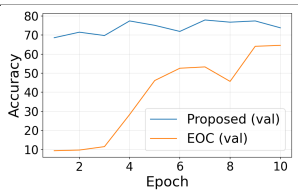
a	MNIST (accuracy vs. epoch)	Fashion-MNIST (accuracy vs. epoch)
10^{-2}		
10^{-1}		
1		
10^{1}		
10^{2}		
10^{3}		

Table 3: Validation accuracy on MNIST (left) and Fashion MNIST (right) for a 50 layer, width 128 fully connected neural network with activation $a\tanh(x)$. Each row corresponds to a different activation scale a. For each scale, the learning rate is tuned separately for each initializer.

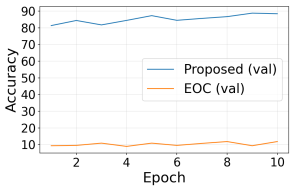
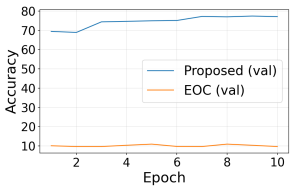
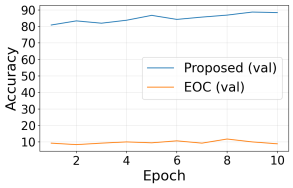
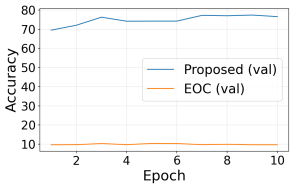
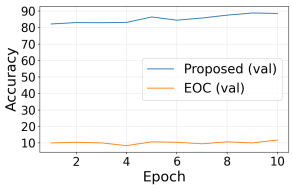
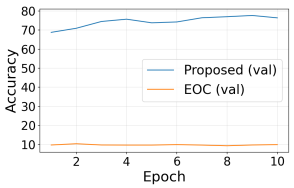
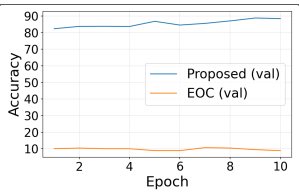
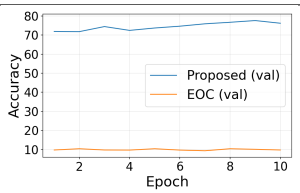
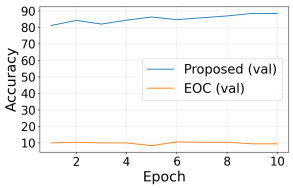
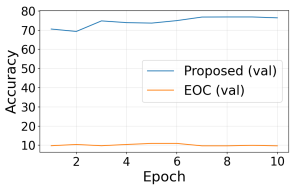
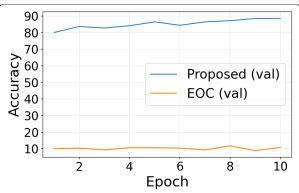
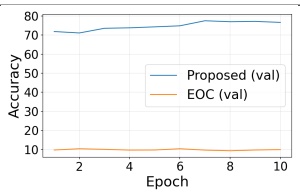
a	MNIST (accuracy vs. epoch)	Fashion-MNIST (accuracy vs. epoch)
10^4		
10^5		
10^6		
10^7		
10^8		
10^9		

Table 4: Validation accuracy on MNIST (left) and Fashion MNIST (right) for a 50 layer, width 128 fully connected neural network with activation $a \arctan(x)$. Each row corresponds to a different activation scale a. For each scale, the learning rate is tuned separately for each initializer.

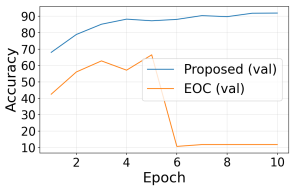
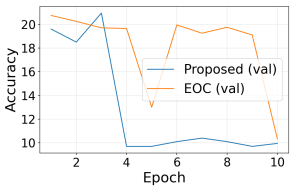
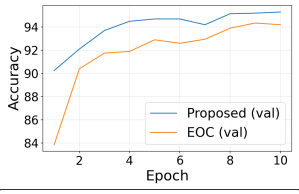
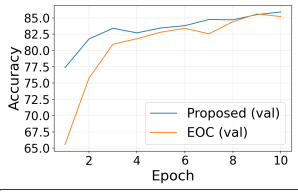
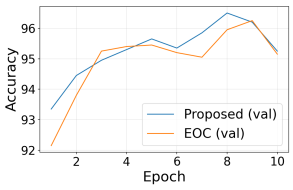
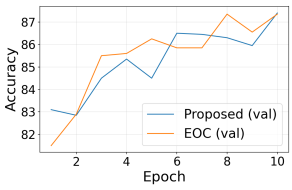
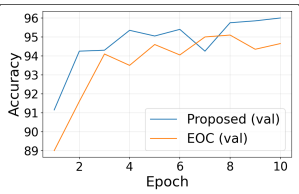
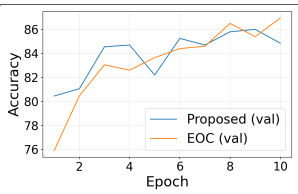
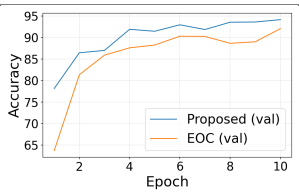
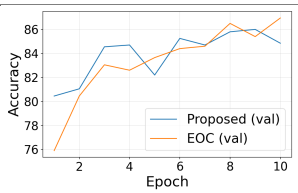
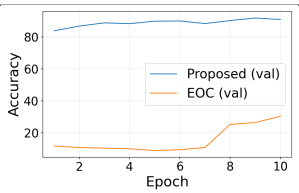
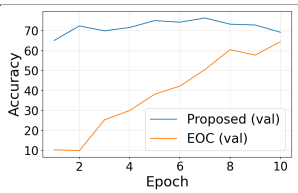
a	MNIST (accuracy vs. epoch)	Fashion-MNIST (accuracy vs. epoch)
10^{-2}		
10^{-1}		
1		
10^{1}		
10^{2}		
10^{3}		

Table 5: Validation accuracy on MNIST (left) and Fashion MNIST (right) for a 50 layer, width 128 fully connected neural network with activation $a \arctan(x)$. Each row corresponds to a different activation scale a. For each scale, the learning rate is tuned separately for each initializer.

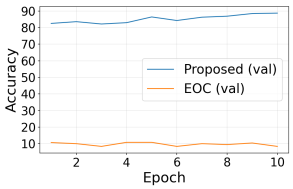
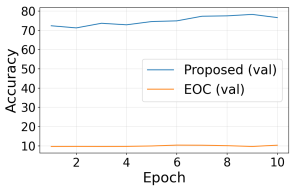
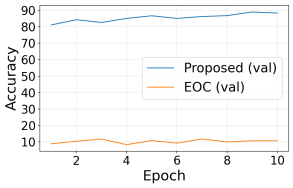
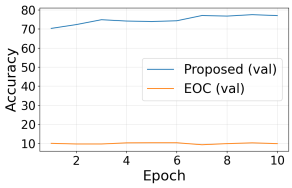
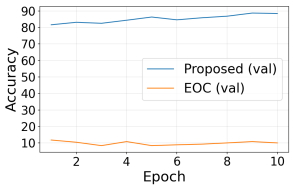
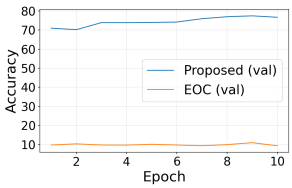
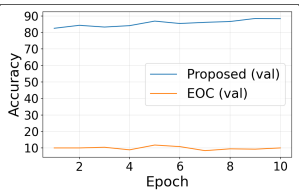
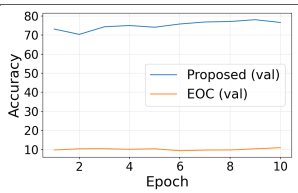
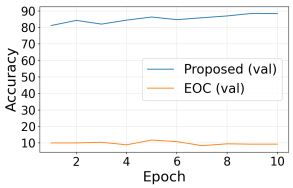
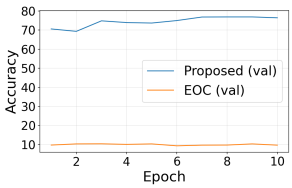
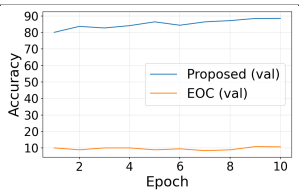
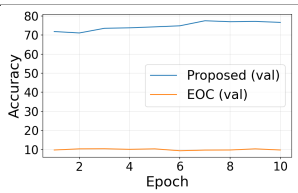
a	MNIST (accuracy vs. epoch)	Fashion-MNIST (accuracy vs. epoch)
10^4		
10^5		
10^6		
10^7		
10^8		
10^9		

Table 6: Validation accuracy on MNIST (left) and Fashion MNIST (right) for a 50 layer, width 128 fully connected neural network with activation $a\,\mathrm{softsign}_2(x)$. Each row corresponds to a different activation scale a. For each scale, the learning rate is tuned separately for each initializer.

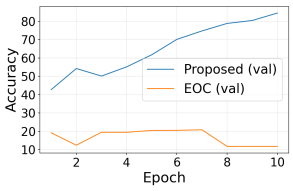
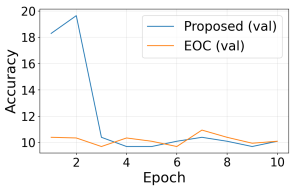
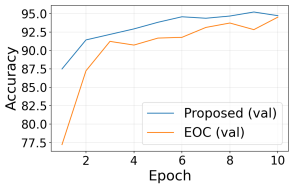
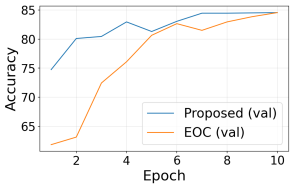
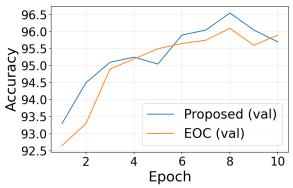
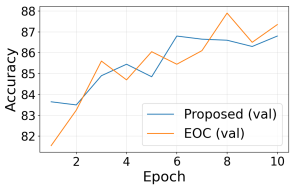
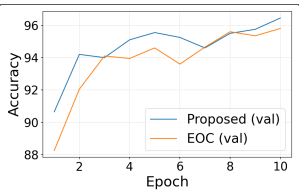
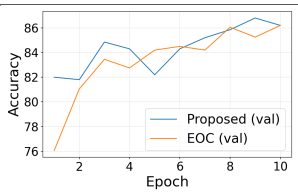
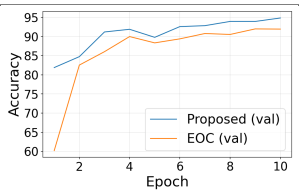
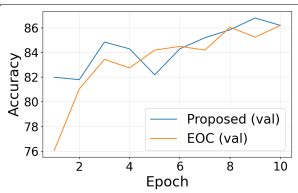
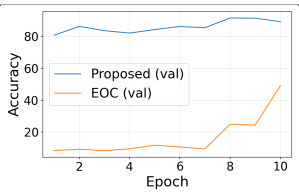
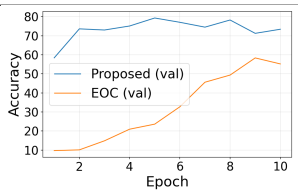
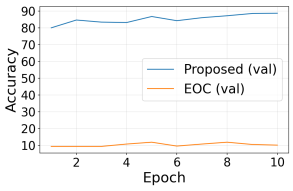
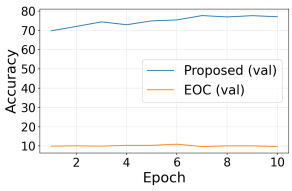
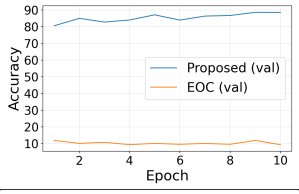
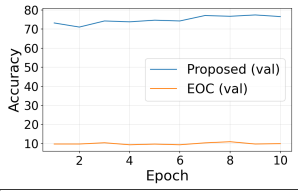
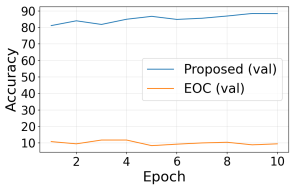
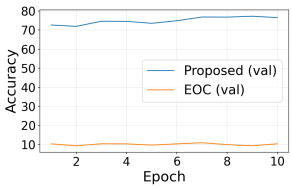
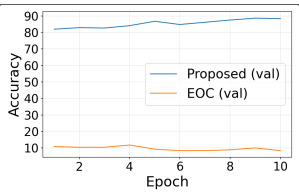
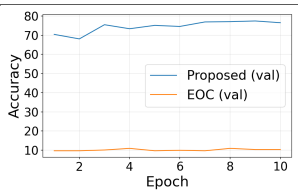
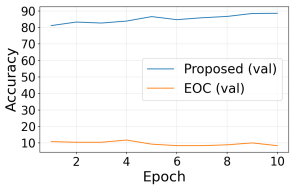
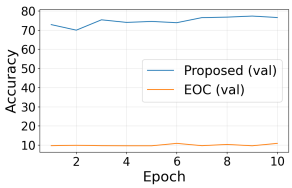
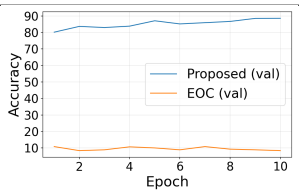
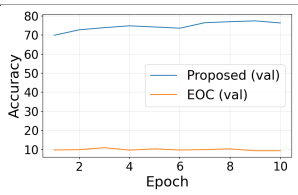
a	MNIST (accuracy vs. epoch)	Fashion-MNIST (accuracy vs. epoch)
10^{-2}		
10^{-1}		
1		
10^{1}		
10^{2}		
10^{3}		

Table 7: Validation accuracy on MNIST (left) and Fashion MNIST (right) for a 50 layer, width 128 fully connected neural network with activation $a\,\mathrm{softsign}_2(x)$. Each row corresponds to a different activation scale a. For each scale, the learning rate is tuned separately for each initializer.

a	MNIST (accuracy vs. epoch)	Fashion-MNIST (accuracy vs. epoch)
10^4		
10^5		
10^6		
10^7		
10^8		
10^9		

C.6 Common PINN Configuration

This appendix expands on the PINN experiments summarized in Section 5.2: the four reference PDEs, the network and training schedule, the three initialization strategies, and the resulting distribution of the best relative L^2 error per configuration.

Problems. We use four PDE benchmarks with known reference solutions:

- **Heat1D.** $u_t - \nu u_{xx} = 0$ on $(x,t) \in [0,1]^2$ with $\nu = 0.1$, $u(x,0) = \sin(\pi x)$, $u(0,t) = u(1,t) = 0$. Reference $u(x,t) = e^{-\nu\pi^2 t}\sin(\pi x)$.
- **Helmholtz2D.** $-\Delta u - k^2 u = (2\pi^2 - k^2)\sin(\pi x)\sin(\pi y)$ on $[0,1]^2$ with $k = 1$ and zero Dirichlet boundary; reference $u = \sin(\pi x)\sin(\pi y)$.
- **Poisson2D.** $-\Delta u = 2\pi^2 \sin(\pi x)\sin(\pi y)$ on $[0,1]^2$ with zero Dirichlet boundary; reference $u = \sin(\pi x)\sin(\pi y)$.
- **AllenCahn.** $u_t - \varepsilon^2 u_{xx} - u + u^3 = 0$ on $(x,t) \in [-1,1] \times [0,1]$ with $\varepsilon = 0.01$, $u(x,0) = x^2\cos(\pi x)$, $u(\pm 1, t) = -1$. Reference computed by an IMEX finite-difference solver on a 513-point spatial grid with 3001 time steps.

Network and training. A fully-connected network with depth $L = 50$ hidden layers and width $n = 64$, followed by a linear readout. The activation is one of five sign-preserving choices — $\alpha\tanh(x)$ for $\alpha \in \{0.2, 0.5, 1, 1.5\}$ and erf. Each epoch resamples 2048 collocation points, 256 boundary points, and (where applicable) 256 initial-condition points; the loss is $\mathcal{L} = \mathcal{L}_{\mathrm{PDE}} + 10\,\mathcal{L}_{\mathrm{BC}} + 10\,\mathcal{L}_{\mathrm{IC}}$. We train for 300 Adam epochs followed by 100 L-BFGS epochs with strong-Wolfe line search (`max_iter` $= 10$, history 50). Adam and L-BFGS learning rates are scaled by $\omega := 1/f'(0)$ for every initializer, applying the same activation-dependent scaling rule to all initialization methods. Each (problem, initialization, activation) configuration is run with 3 seeds; the per-run summary is the minimum relL2 to the reference reached across the 400 training epochs.

Initialization strategies. **Proposed:** hidden layers as $W_\ell = \omega I + \sigma Z_\ell$, $Z_{\ell,ij} \sim \mathcal{N}(0, 1/n)$, with σ calibrated by bisection on the empirical terminal sign-flip rate of a fresh forward-pass batch so that it matches a target $p \in \{0.10, 0.20, 0.30, 0.40, 0.49\}$. **Xavier:** hidden layers $W_\ell \sim \mathcal{N}(0, \omega^2/n)$ (standard Xavier with gain rescaled by ω); input/output layers standard Xavier. **EOC** (edge-of-chaos): Gaussian weights with variance σ_w^2 that places the iterated mean-field map on its critical line at depth L, together with a fixed-point bias σ_b found by 1D search. Each initializer thus contributes 5 activations $\times 3$ seeds $= 15$ runs per box in Fig. 7.

Observation. On every PDE, the proposed scheme at small-to-moderate p attains a substantially lower median best-relL2 than both Xavier and EOC — a factor of 14–45$\times$ over EOC and 26–46$\times$ over Xavier. The best target varies with the problem ($p^\star = 0.30, 0.10, 0.20, 0.40$ for Heat1D, Helmholtz2D, Poisson2D, AllenCahn) but always lies in the identity-dominated half of the range; pushing p to 0.49 (toward random-sign) is the worst proposed setting on three of the four problems. Both baseline boxes display heavy upper tails with several runs failing to converge (relL2 ~ 1), while the proposed scheme rarely produces such failures.

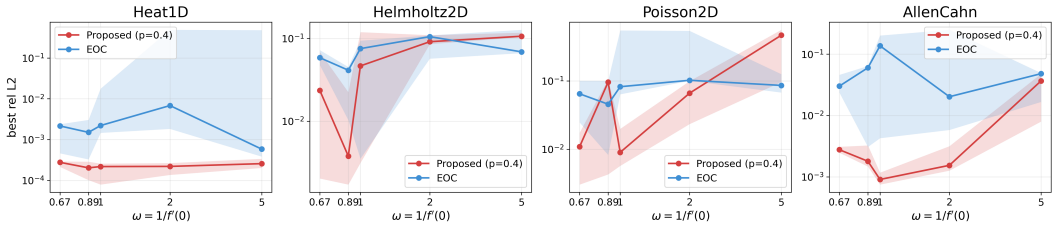


Figure 40: Best relative L^2 vs activation parameter $\omega = 1/f'(0)$ at fixed sign-flip target $p = 0.4$ (red) versus EOC (blue). Solid line: median over 3 seeds. Shaded band: min–max over seeds. Other settings as in Fig. 7.

Effect of activation at fixed p. Figure 7 marginalizes over the five activations to display the effect of p. To complement that view we fix $p = 0.4$ and trace best-relL2 as the activation parameter $\omega = 1/f'(0)$ varies over the same five-point sweep $\omega \in \{2/3, \sqrt{\pi}/2, 1, 2, 5\}$ (corresponding to $1.5\tanh(x)$, erf, $\tanh$, $0.5\tanh(x)$, $0.2\tanh(x)$), comparing against EOC. All other settings — depth, width, training schedule, ω-scaling of the learning rates — match those of Fig. 7; for each (ω, initializer) pair we plot the median over 3 seeds with a shaded min–max band.

Observation. On all four problems the proposed median sits at or below the EOC median across most of the ω range; the gap is largest at small-to-moderate $\omega \in [2/3, 1]$ (covering $\tanh$, erf, and the steepest variant we test), where it reaches one to two orders of magnitude. Both methods degrade at the right end $\omega = 5$, i.e. $0.2\tanh(x)$ with $f'(0) = 0.2$. The proposed seed band is visibly tighter than EOC's on Heat1D and AllenCahn, suggesting that calibrating σ to a sign-flip target also stabilizes runs across seeds relative to a variance-only criterion. Together with Fig. 7, this confirms that the proposed advantage at $p = 0.4$ is broad across activation choices, not artefactual to any single f.