How Long Does Infinite Width Last? Signal Propagation in Long-Range Linear Recurrences

Mariia Seleznova
Technische Universität Darmstadt
Max Planck Institute of Quantum Optics

Abstract

We study signal propagation in linear recurrent models at finite width. While existing signal propagation theory relies predominantly on the infinite-width limit, it remains unclear for how long that approximation remains accurate when recurrent depth t grows jointly with width n. This question is especially relevant for modern recurrent sequence models, whose natural operating regime involves long input sequences, i.e., large t. We derive exact finite-width formulas for the hidden state signal energies in linear recurrences under complex Gaussian initialization. Using these formulas, we identify the joint depth–width scaling regimes that govern signal propagation: (i) a *subcritical regime* $t = o(\sqrt{n})$, in which the infinite-width approximation remains valid; (ii) a *critical regime* $t \sim c\sqrt{n}$, in which non-negligible deviations from infinite-width predictions appear and a nontrivial joint scaling limit emerges; and (iii) a *supercritical regime* $t \gg \sqrt{n}$, in which finite-width effects dominate. Thus, our results pinpoint the precise recurrent depth scale at which infinite-width theory breaks down in long-range linear recurrences. In turn, this shows when standard initialization schemes, such as Glorot, become unstable. More broadly, our results show that what constitutes "large width" depends crucially on the architecture: recurrent weight sharing makes finite-width effects visible at substantially shorter depths than in feedforward networks.

1 Introduction

Understanding how signals propagate through the layers of a neural network (NN) is a foundational problem in deep learning (DL) theory. When signals are repeatedly transformed by many layers of a deep network, their magnitudes may progressively grow or decay, eventually leading to numerical failure. This phenomenon is widely known as the *vanishing/exploding signals problem*, and it poses a fundamental obstacle to training deep NNs [5, 30]. Ensuring stable signal propagation has therefore long been a central design principle in DL methods, motivating widely used initialization schemes [13, 23, 33] and architectural choices [3, 24, 25].

Despite the foundational nature and long history of this question, the theoretical understanding of signal propagation in DL models remains limited. Much of the existing literature relies on the *infinite-width limit* [31, 33, 41, 44], where the network's width n (i.e, the number of neurons per layer) tends to infinity. In this regime, mean-field-type approximations yield tractable descriptions of signal statistics and lead to elegant signal propagation theories for many DL architectures [41, 44], including recurrent models [1, 6, 12]. Realistic networks, however, operate at finite width. In this setting, even small deviations from the infinite-width limit can accumulate over many layers or recurrent steps, eventually rendering infinite-width predictions inaccurate in sufficiently deep architectures or long-range recurrences [4, 19, 21, 32, 35]. This leads to the central question of this paper:

How long does the infinite-width approximation remain accurate in recurrent models?

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

For fully-connected feedforward networks at random initialization, the answer is relatively well understood. The deviations from infinite-width behavior become significant when depth scales proportionally with width, i.e., $L \sim cn$, where L is the number of layers and n is the layer width [19–21, 35]. That is, for fully connected feedforward networks, the critical depth scale at which finite-width effects begin to qualitatively alter signal propagation is by now known.

By contrast, much less is known for recurrent networks. Although existing work suggests that finite-width corrections behave differently in recurrent and feedforward architectures [14, 34], their precise effect for long-range recurrent signal propagation remains unclear. Intuitively, finite-width effects accumulate faster in recurrences, since the same weight matrix is applied repeatedly and its dominant spectral directions are therefore reinforced across recurrent steps t. Bar et al. [4] take a first step toward formalizing this intuition in linear recurrent sequence models with Gaussian initialization, showing that accumulated finite-width effects cause signal explosion when $t \gg \sqrt{n}$. However, their argument gives only a crude lower bound on the signal size. Thus, even for linear recurrences at random initialization, the sharp depth scale at which infinite-width predictions break down has remained unknown.

This gap is especially striking from the perspective of modern long-range sequence modeling, where architectures built around *linear recurrences*—such as state-space models (SSMs) [8, 16, 17] and linear recurrent units (LRUs) [29]—are actively explored as computationally efficient alternatives to transformers for long-context tasks. These models have recently achieved competitive empirical performance while retaining the computational benefits of linear recurrence. Since they are designed for long-range tasks, their natural operating regime is that of long input sequences, or equivalently large recurrent depth t. Yet this is precisely the regime in which signal propagation theory remains weak.

This paper helps close this gap by characterizing finite-width signal propagation in randomly-initialized linear recurrences and identifying the recurrent depth scaling regimes in which the infinite-width approximation remains accurate and those in which it fails. We first specify the considered setting in Section 1.1, and then summarize the main contributions in Section 1.2.

1.1 Problem Setting

Our main object of interest is the linear recurrent unit (LRU), which has recently re-emerged as a building block for long-range sequence models [29]. On the way, our analysis also yields results for the linear recurrent network without sequential inputs. Thus, the models are:

$$\textbf{Linear recurrent unit (LRU):} \qquad \mathbf{h}_{\mathrm{LRU}}^{(t)} := \mathbf{W}\mathbf{h}_{\mathrm{LRU}}^{(t-1)} + \mathbf{x}^{(t)} = \sum_{k=0}^{t} \mathbf{W}^k \mathbf{x}^{(t-k)},\ t \geq 1, \quad (1)$$

$$\textbf{Linear recurrent network (RNN):} \qquad \mathbf{h}_{\mathrm{RNN}}^{(t)} := \mathbf{W}\mathbf{h}_{\mathrm{RNN}}^{(t-1)} = \mathbf{W}^t \mathbf{x}^{(0)},\ t \geq 1. \quad (2)$$

Here $\mathbf{h}^{(t)} \in \mathbb{C}^n$ is the hidden state vector at step $t \in \mathbb{N}$, $(\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(t)})$ is the input sequence, where $\mathbf{x}^{(i)} \in \mathbb{R}^n$ for all $0 \leq i \leq t$, and $\mathbf{W} \in \mathbb{C}^{n \times n}$ is the weight matrix. We set $\mathbf{h}^{(0)} = \mathbf{x}^{(0)}$ in both models. The weight matrix $\mathbf{W}$ is initialized with i.i.d. proper complex Gaussian entries of variance $1/n$, namely

$$\mathbf{W}_{ij} \sim \mathcal{N}_{\mathbb{C}}(0, 1/n), \quad \text{i.i.d. across } i, j = 1, \ldots, n. \quad (3)$$

The main results in this work are formulated for complex weights. This choice is natural for two reasons. First, complex-valued weights are standard in modern linear recurrent models [17, 29], since complex eigenvalues provide a convenient representation of oscillatory dynamics. Second, the complex Gaussian ensemble admits a particularly clean analysis, which allows us to derive sharp finite-width results for signal propagation. At the same time, we show that our main conclusions on the limitations of infinite-width theory remain valid for real weights (see Remark 3).

Our main quantitative objects are the normalized expected squared norms of the hidden states $\mathbf{h}_{\mathrm{LRU}}^{(t)}$ and $\mathbf{h}_{\mathrm{RNN}}^{(k)}$, which we call *signal energies*. They are defined as follows:

$$\textbf{Signal energies}: \quad S_{n,t} := \frac{1}{n}\mathbb{E}\big[\|\mathbf{h}_{\mathrm{LRU}}^{(t)}\|_2^2\big], \qquad Q_{n,k} := \frac{1}{n}\mathbb{E}\big[\|\mathbf{h}_{\mathrm{RNN}}^{(k)}\|_2^2\big], \quad (4)$$

where the expectation is taken jointly over the weights $\mathbf{W}$ and the inputs $(\mathbf{x}^{(0)}, \ldots, \mathbf{x}^{(t)})$. The same quantities are studied in classical works on signal propagation at infinite width [31, 33]. Notably,

standard Gaussian initialization schemes—such as He [23] and Glorot [13]—are designed to stabilize these signal energies across layers at infinite width, so that $Q_{\infty,k} \approx Q_{\infty,k+1}$. However, in this work we explicitly study the behavior of these quantities at *finite width*.

1.2 Contributions

In the setting described above, the main contributions of this work are as follows:

- **Exact finite-width characterization of signal propagation.** We derive exact closed-form expressions for signal energies in finite-width linear recurrences with complex Gaussian initialization. Specifically, we obtain expressions for $S_{n,t}$ and $Q_{n,k}$ (defined in (4)) for finite n, t and k. This result is given in Theorem 2.1. To the best of our knowledge, this is the first exact finite-width result for signal propagation in recurrent models, which provides a clear departure from the standard infinite-width signal propagation theory.
- **Sharp depth–width scaling laws.** By analyzing the finite-width expressions, we identify the recurrent depth–width scaling regimes that govern signal propagation. In particular, we characterize three regimes: (i) the *subcritical regime* $k, t = o(\sqrt{n})$, in which the infinite-width approximation remains accurate; (ii) the *critical regime* $k, t \sim c\sqrt{n}$, in which a nontrivial joint depth–width limit emerges; and (iii) the *supercritical regime* $k, t \gg \sqrt{n}$, in which finite-width effects dominate. The behavior of signal energies $S_{n,t}$ and $Q_{n,k}$ in these regimes is summarized in Figure 1. Together, these results yield a precise joint scaling signal propagation theory for recurrent depth and width in linear recurrences.
- **Limitations of infinite-width theory for long-range recurrent models.** The identified critical depth scale $t \sim c\sqrt{n}$, at which finite-width corrections qualitatively alter signal propagation, makes explicit when infinite-width theory breaks down in long-range recurrent models. This has two main consequences. First, standard Glorot initialization [13], while stable at infinite width, no longer ensures stable propagation from the critical regime onward. Second, recurrent and feedforward architectures behave qualitatively differently at finite width, despite being indistinguishable at infinite width. Although our theory concerns linear recurrences, numerical experiments suggest that these limitations of infinite-width theory persist in certain nonlinear models (see Figure 4).

In addition, we validate our theoretical results with numerical experiments, which confirm the predicted signal propagation behavior.

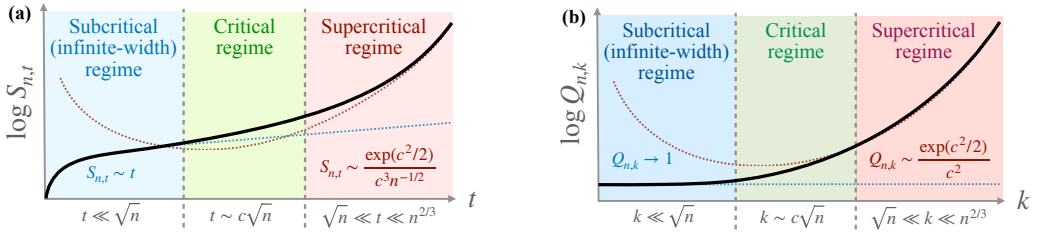

Figure 1: **Depth–width scaling laws in linear recurrences.** The behavior of the signal energies under three joint scaling regimes of recurrent depth and width: (i) the *subcritical regime* $k, t = o(\sqrt{n})$, (ii) the *critical regime* $k, t \sim c\sqrt{n}$, and (iii) the *supercritical regime* subset $\sqrt{n} \ll k, t \ll n^{2/3}$. Panel (a) shows the LRU signal energy $S_{n,t}$, and panel (b) shows the RNN signal energy $Q_{n,k}$, as defined in (4). Blue dotted lines indicate the continuation of the predicted infinite-width behavior beyond the subcritical regime, while red dotted lines indicate the continuation of the supercritical behavior beyond the displayed supercritical regime. Both quantities are shown on a logarithmic scale.

Outline. Section 2 states the main finite-width signal propagation result. Section 3 analyzes the resulting depth–width scaling laws, and Section 4 presents numerical experiments. Section 5 concludes with a discussion. Additional related work is discussed in Appendix A.

2 Recurrent Signal Propagation at Finite Width

We now state the main finite-width signal propagation theorem, proven in Appendix B.

Theorem 2.1 (Signal propagation at finite width)**.** *Suppose the weight matrix $\mathbf{W} \in \mathbb{C}^{n\times n}$ has i.i.d. proper complex Gaussian entries with variance $1/n$, as in (3). Let $(\mathbf{x}^{(0)}, \dots, \mathbf{x}^{(t)})$ be an input sequence independent of $\mathbf{W}$, and assume that $\mathbb{E}[\mathbf{x}^{(s)}\mathbf{x}^{(s)\top}] = \mathbb{I}_n$ for all $0 \le s \le t$. No independence assumptions across time are required. Then the following identities hold for the hidden state signal energies of the LRU and the linear RNN, defined in (4):*

$$S_{n,t} = \sum_{k=0}^{t} Q_{n,k}, \qquad Q_{n,k} = \frac{F^{+}_{k+2}(n) - F^{-}_{k+2}(n)}{n^{k+1}(k+1)(k+2)}. \tag{5}$$

Here $F^{+}_{p}(n)$ and $F^{-}_{p}(n)$ denote the rising and falling factorials, respectively:

$$F^{+}_{p}(n) := n(n+1)\cdots(n+p-1), \qquad F^{-}_{p}(n) := n(n-1)\cdots(n-p+1). \tag{6}$$

This theorem gives exact closed-form expressions for the signal energies of linear recurrences at finite width. Its main value is that it enables a precise analysis of signal propagation under arbitrary joint scaling of recurrent depth and width, for example in regimes of the form $t, k \sim n^{\alpha}$. Section 3 develops the corresponding depth–width scaling laws, summarized in Figure 1.

Below, we make several remarks on extensions and variants of this result.

Remark 1 (General input distributions). Theorem 2.1 is stated for whitened inputs for simplicity. However, the same argument extends immediately to arbitrary input covariances $\Sigma^{(0)}, \dots, \Sigma^{(t)}$, where $\mathbb{E}[\mathbf{x}^{(s)}\mathbf{x}^{(s)\top}] = \Sigma^{(s)}$ for all $0 \le s \le t$. In that case, the expected LRU signal energy is

$$S_{n,t} = \sum_{k=0}^{t} \frac{\operatorname{tr}\big(\Sigma^{(t-k)}\big)}{n}\, \tilde{Q}_{n,k}, \tag{7}$$

where $\tilde{Q}_{n,k}$ denotes the RNN signal energy under whitened inputs, as given in Theorem 2.1. In fact, the proof in Appendix B is formulated in this general setting. Since we impose no additional assumptions on cross-time correlations or the distribution of the input sequence $(\mathbf{x}^{(0)}, \dots, \mathbf{x}^{(t)})$, the result is fully general with respect to the inputs, apart from their independence from the weights $\mathbf{W}$.

Remark 2 (Gradient propagation). The quantities in Theorem 2.1 also govern backward (gradient) propagation through the linear recurrence. Indeed, consider the output $y = \langle \mathbf{c}, \mathbf{h}^{(T)}_{\mathrm{RNN}} \rangle$, where $\mathbf{c}$ are the output weights independent of $\mathbf{W}$. Backpropagation over t steps in this model is controlled by

$$\boldsymbol{\delta}^{(T-t)} := \frac{\partial y}{\partial \mathbf{h}^{(T-t)}_{\mathrm{RNN}}} = (\mathbf{W}^{*})^{t}\mathbf{c}, \qquad 0 \le t \le T. \tag{8}$$

This is the backward analogue of the forward recursion, and the same unitary invariance argument as in Lemma B.9 gives:

$$\frac{1}{n}\mathbb{E}\Big[\|\boldsymbol{\delta}^{(T-t)}\|_2^2\Big] = \frac{\mathbb{E}\left[\|\mathbf{c}\|_2^2\right]}{n}\, Q_{n,t}, \tag{9}$$

where $Q_{n,t}$ is the signal energy for whitened inputs, as computed in Theorem 2.1. This shows that the gradient energy obeys the same scaling laws as the forward signal energy.

Remark 3 (Real weights). While our main result is stated for complex weights, which are common in practical linear recurrences, it also directly yields a lower bound for the corresponding real Gaussian model. Let $Q^{\mathbb{R}}_{n,k}$ and $S^{\mathbb{R}}_{n,t}$ denote the analogues of $Q_{n,k}$ and $S_{n,t}$ for a real Gaussian weight matrix with i.i.d. entries of variance $1/n$. Then, with an additional assumption that cross-time input correlations satisfy $\operatorname{tr}(\mathbb{E}[\mathbf{x}^{(s)}\mathbf{x}^{(r)\top}]) \ge 0$ for all $0 \le s, r \le t$, we have:

$$Q^{\mathbb{R}}_{n,k} \ge Q_{n,k}, \qquad S^{\mathbb{R}}_{n,t} \ge S_{n,t}, \tag{10}$$

for all n, k, t. In particular, any depth–width scaling regime that leads to exploding signals in the complex model also leads to exploding signals in the real model. We prove this result in Appendix B.4.

Remark 4 (Nonlinear recurrences). While Theorem 2.1 concerns linear recurrences, the departure from infinite-width theory identified here does not appear to be restricted to the linear setting. Our empirical results for RNNs with ReLU and LeakyReLU activations show that the corresponding signal energy $Q^{\phi}_{n,t}$ grows rapidly as a function of $c = t/\sqrt{n}$, qualitatively resembling the behavior observed in the linear case (see Figure 4). Notably, the signal growth for these activations is *faster* than in the corresponding linear networks. A theoretical characterization of these finite-width effects in nonlinear RNNs remains open.

2.1 Proof Idea

The proof combines standard random matrix techniques with a reduction to a cycle counting problem for certain random permutations. The starting point is the observation that the main difficulty in computing both signal energies, $Q_{n,k}$ and $S_{n,t}$, is the evaluation of the finite-width *trace moment*

$$\frac{1}{n}\,\mathbb{E}\Big[\operatorname{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big)\Big]. \tag{11}$$

The reduction of $Q_{n,k}$ and $S_{n,t}$ to this quantity is carried out in Appendix B.3 using standard properties of Gaussian random matrices. Thus, the core of the proof is to evaluate this trace moment exactly in the given complex Gaussian setting.

To evaluate the trace moment, we use a Wick expansion argument: we first expand the trace into a sum of monomials of the form $\overline{\mathbf{W}}_{a_0a_1}\,\overline{\mathbf{W}}_{a_1a_2}\cdots\overline{\mathbf{W}}_{a_{k-1}a_k}\,\mathbf{W}_{a_0b_1}\cdots\mathbf{W}_{b_{k-1}a_k}$, and then apply Wick's theorem for Gaussian moments [26, 38] to compute their expectations. This is a standard way of converting random matrix moments into combinatorial counting problems, and we develop it in our setting in Appendix B.1. Related ideas have been used successfully in finite-width analyses of fully-connected feedforward networks [32, 40]. In the recurrent setting, however, the resulting combinatorics are substantially more intricate because the same random matrix is reused across time. The key simplification in our setting is that, for complex Gaussian recurrences, the Wick expansion takes a particularly clean form and can be indexed by permutations $\sigma \in S_k$. Each such permutation contributes according to the number of matrix indices that remain unconstrained after the corresponding contractions. Under the complex Gaussian initialization with variance $1/n$, this yields the exact representation

$$\mathbb{E}\left[\frac{1}{n}\operatorname{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big)\right] = \sum_{\sigma\in S_k} n^{-k-1+F(\sigma)}, \tag{12}$$

where $F(\sigma)$ denotes the number of *free indices*, i.e. the number of equivalence classes of indices $(a_0,\dots,a_k,b_1,\dots,b_{k-1})$ left unconstrained by the contractions associated with σ.

For a permutation σ, the quantity $F(\sigma)$ turns out to be exactly the number of cycles of the commutator $[\tau^{-1},\tilde{\sigma}^{-1}]$, where $\tau = (0,1,\dots,k)$ is a long cycle and $\tilde{\sigma}$ is the extension of σ fixing 0. This provides the link between the Wick expansion and permutation combinatorics. As we show in Appendix B.2, this identification reduces the problem to counting cycles of permutations of the form $\tau^{-1}c$, where c is a uniformly random full cycle. This quantity can be computed exactly using known results, in particular those of Dubach [9]. While the argument in Dubach [9] and ours is formulated in a different language, it is closely related to the literature on Hultman numbers and, in particular, to the corresponding closed formulas [2, 36].

Putting these ingredients together yields an exact closed-form expression for the trace moment, and therefore for $Q_{n,k}$ and $S_{n,t}$. In this way, the proof reduces the finite-width signal propagation problem to a finite combinatorial counting problem that can be solved exactly.

3 Depth–Width Scaling Regimes for $Q_{n,k}$ and $S_{n,t}$

The exact finite-width formulas in Theorem 2.1 allow to analyze signal propagation under joint scaling of recurrent depth and width. In particular, they make it possible to characterize regimes such as $k \sim n^{\alpha}$ and $t \sim n^{\alpha}$, and thus identify the depth scales at which the infinite-width approximation remains accurate and those at which it fails.

For this purpose, it is convenient to rewrite $Q_{n,k}$ as follows:

$$Q_{n,k} = \frac{n}{(k+1)(k+2)} \left[\prod_{j=0}^{k+1} \left(1 + \frac{j}{n}\right) - \prod_{j=0}^{k+1} \left(1 - \frac{j}{n}\right) \right], \tag{13}$$

since this form makes the relevant asymptotic regimes more transparent. The corresponding analysis for $Q_{n,k}$ is given in Appendix C, and for $S_{n,t}$ in Appendix D. We now describe the asymptotic behavior of $Q_{n,k}$ and $S_{n,t}$ across the three regimes illustrated in Figure 1: the *subcritical*, *critical*, and *supercritical* regimes.

3.1 Subcritical (Infinite-Width) Regime: $k = o(\sqrt{n})$ and $t = o(\sqrt{n})$

We start with the regime in which the recurrent depth remains asymptotically below the critical $\sqrt{n}$ scale. In this case, signal propagation follows the same behavior as in the infinite-width limit:

$$Q_{n,k} \to 1, \qquad S_{n,t} \sim t + 1. \tag{14}$$

Here, and throughout this section, the notation $a \sim b$ means that $a/b \to 1$ in the corresponding joint scaling limit. Precise statements and proofs for this regime are given in Theorems C.1 and D.1 in the Appendix. Below, we make several observations about the implications of these results.

Stability of Glorot initialization. The asymptotics in (14) coincide exactly with the classical infinite-width limit $n \to \infty$ at fixed k and t. In particular, $Q_{n,k}$ remains asymptotically equal to 1, uniformly over recurrent depths satisfying $k \ll \sqrt{n}$. Thus, under standard Glorot initialization (3), signals remain stable throughout the entire subcritical regime. In this sense, our result extends the classical stability of Glorot initialization at infinite width with fixed recurrent depth [6, 13] to the joint depth–width scaling regimes with $k = o(\sqrt{n})$.

Infinite-width theory does not distinguish recurrent and feedforward networks. Throughout the subcritical regime, recurrent and feedforward signal propagation remain asymptotically indistinguishable at the level of signal energies. More precisely, if one replaces the shared recurrent matrix $\mathbf{W}$ by independent matrices $\mathbf{W}^{(0)}, \ldots, \mathbf{W}^{(t)}$ across steps in the linear RNN (2) or the LRU (1), then the asymptotic behavior of $Q_{n,k}$ and $S_{n,t}$ remains unchanged under the scaling $k, t = o(\sqrt{n})$. This result also holds in existing infinite-width frameworks, such as Tensor Programs [41], and this observation is made explicit for example in Alemohammad et al. [1].

While the second observation agrees with well-established literature, it is somewhat unsatisfactory from the perspective of recurrent modeling. In practice, recurrent networks are much more prone than feedforward ones to signal instability, and mathematically one expects repeated application of the same matrix to accumulate signal growth faster than independent layer-wise multiplication. The subcritical regime therefore identifies the range in which infinite-width theory remains faithful, but it also highlights its main limitation: below the critical scale, the theory cannot distinguish recurrence from feedforward composition. One of the main contributions of this paper is to show that this distinction becomes visible precisely at the critical scaling regime, which we analyze next.

3.2 Critical Regime: $k \sim c\sqrt{n}$ and $t \sim \tilde{c}\sqrt{n}$

The finite-width corrections become non-negligible when recurrent depth reaches the critical $\sqrt{n}$ scale. In this regime, the signal energy $Q_{n,k}$ no longer follows the infinite-width prediction and instead converges to a nontrivial critical profile, illustrated in Figure 1 and given by:

$$Q_{n,k} \to \frac{2}{c^2} \sinh\left(\frac{c^2}{2}\right), \quad \text{where } k \sim c\sqrt{n}, \quad c > 0. \tag{15}$$

Thus, unlike in the subcritical regime, the signal energy now depends nontrivially on the ratio $k/\sqrt{n}$. In particular, $Q_{n,k}$ grows strictly above its infinite-width value when $c > 0$, showing systematic finite-width signal amplification over depth. This result is proven in Theorem C.2 in the Appendix.

For the cumulative LRU signal energy $S_{n,t}$, this yields the following asymptotic law:

$$S_{n,t} \sim \sqrt{n} \int_0^{\tilde{c}} \frac{2}{x^2} \sinh\left(\frac{x^2}{2}\right) dx, \quad \text{where } t \sim \tilde{c}\sqrt{n}, \quad \tilde{c} > 0. \tag{16}$$

where the integrand is continuously extended to take the value 1 at $x = 0$. This result is proven in Theorem D.2 in the Appendix. For large $\tilde{c} = t/\sqrt{n}$, this integral is dominated by its upper endpoint and behaves as $\tilde{c}^{-3}\exp(\tilde{c}^2/2)$. Thus, in the critical regime, signal growth becomes exponential in the scaling parameter $t/\sqrt{n}$. Below, we discuss the implications of the emergence of critical regime.

Instability of Glorot initialization. The transition to the critical regime shows that standard Glorot initialization (3) becomes unstable once recurrent depth reaches the $\sqrt{n}$ scale. By Remark 3, this conclusion applies to both complex and real models. This helps explain why Glorot initialization, although natural from the infinite-width perspective, typically becomes unstable in modern long-range recurrent models [17, 29]. We note that Bar et al. [4] recently showed that the scaling $t \sim \tilde{c}\sqrt{n}$ already produces exponential growth in $S_{n,t}$. However, they established only a crude lower bound, which in particular missed the $\sqrt{n}$ prefactor in (16). Our result sharpens this picture in two ways: it shows that the $\sqrt{n}$ scale is exactly the threshold at which the transition to critical behavior occurs, and identifies the full asymptotic profile in this regime. In this sense, we give a precise answer to when Glorot ceases to be stable in long-range linear recurrences.

Recurrent versus feedforward signal propagation. Starting at the critical regime, signal propagation in recurrent networks begins to differ sharply from that in feedforward networks—a distinction that is invisible in the infinite-width theory. While finite-width corrections accumulate and become non-negligible once depth reaches the $\sqrt{n}$ scale in recurrences, the relevant depth scale for fully-connected networks is $L \sim cn$ [19, 22, 35]. Moreover, even in this proportional regime, the signal energies remain stable in feedforward networks, while finite-width effects are primarily reflected in their fluctuations, which has important consequences for convergence to the kernel regime [21, 35]. Thus, the effects of finite-width corrections and the depth scales at which they first become visible are fundamentally different in recurrent and feedforward architectures. This observation helps reconcile theory with practice: recurrent models are known to be substantially more prone to signal instability, but this distinction is not captured by standard infinite-width analyses.

3.3 Supercritical sublinear regime: $\sqrt{n} \ll k \ll n$ and $\sqrt{n} \ll t \ll n$

Beyond the critical depth scale, finite-width effects dominate, and both signal energies diverge in the joint (n, k) limit. In particular, when $\sqrt{n} \ll k \ll n$ and $\sqrt{n} \ll t \ll n$, we have

$$Q_{n,k} \sim \frac{n}{(k+1)(k+2)} e^{n\Psi(k/n)}, \qquad S_{n,t} \sim \frac{n^2}{t^3}\exp\Big(n\Psi(t/n)\Big), \tag{17}$$

where $\Psi(x) := (1+x)\log(1+x) - x$. Thus, once recurrent depth exceeds the critical $\sqrt{n}$ scale, signal growth becomes exponential in depth and no finite joint scaling limit exists. These results are proven in Theorems C.3 and D.3 in the Appendix.

A particularly tractable subregime occurs when $k, t = o(n^{2/3})$. In that case, the exponent can be expanded to the leading order, yielding the asymptotic profiles illustrated in Figure 1:

$$Q_{n,k} \sim \frac{n}{k^2}\exp\Big(\frac{k^2}{2n}\Big), \qquad S_{n,t} \sim \frac{n^2}{t^3}\exp\Big(\frac{t^2}{2n}\Big). \tag{18}$$

In principle, the analysis can be extended further, beyond the proportional depth–width scaling. However, all such regimes exhibit rapid signal explosion, so we do not pursue them here.

4 Numerical Experiments

We investigate two questions numerically: whether the derived theory accurately describes empirical signal propagation behavior, and if the identified phenomena persist beyond the linear setting.[1]

Linear recurrences. Figures 2–3 compare Monte Carlo estimates of $Q_{n,k}$ and $S_{n,t}$ against the critical scaling variable $c = t/\sqrt{n}$. For complex Gaussian weights (Figure 2), the curves at different widths collapse onto the critical profiles predicted by (15)–(16). For real Gaussian weights (Figure 3), the signal energies are larger but exhibit the same qualitative amplification, consistent with Remark 3.

[1] Code and saved results available at: `https://github.com/mselezniova/long-range-rnn`.

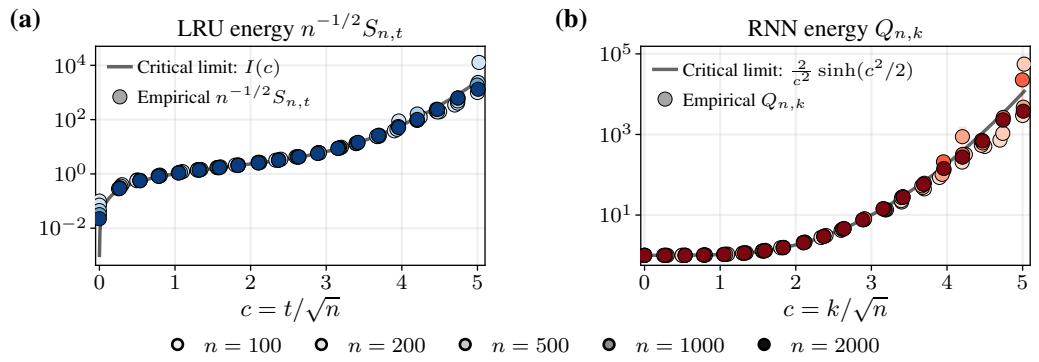

Figure 2: **Numerical validation of the critical scaling theory.** Monte Carlo estimates of $S_{n,t}$ (a) and $Q_{n,k}$ (b), based on 10^5 samples and plotted against the critical scaling variable c for several widths n. Inputs $\mathbf{x}$ are sampled uniformly on the sphere of radius $\sqrt{n}$, so that $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \mathbb{I}_n$. In both panels, the numerical estimates are compared with the corresponding theoretical critical profiles, given by (16) for $S_{n,t}$ and (15) for $Q_{n,k}$. The close agreement illustrates convergence of the signal energies to their predicted critical limits. Error bars denote Monte Carlo standard errors of the mean.

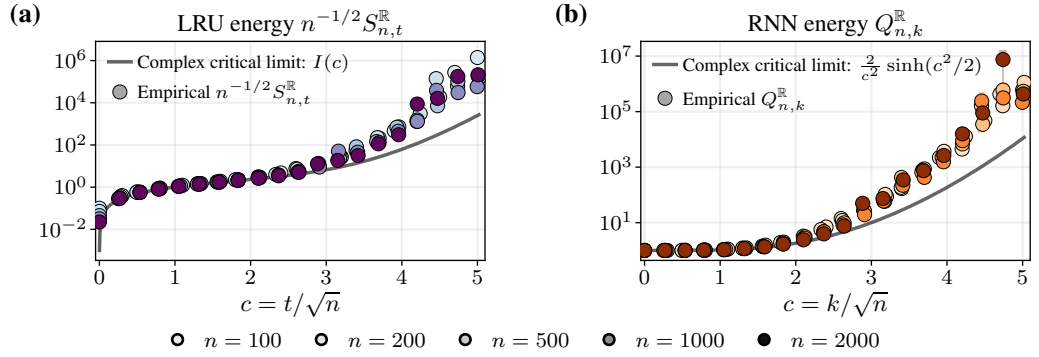

Figure 3: **Real Gaussian model.** Analogous to Figure 2, but for real Gaussian weights. In both panels, the solid curve is the complex critical profile from (16) and (15). The figure illustrates that the real model signal energies are larger than their complex counterparts, in agreement with Remark 3.

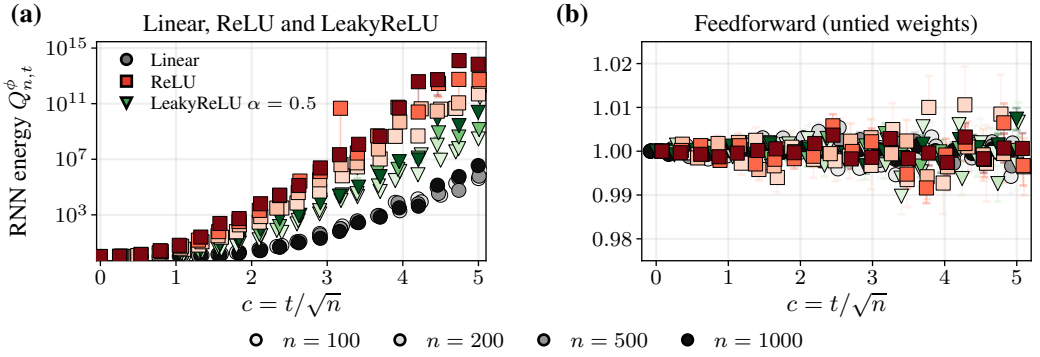

Figure 4: **Signal propagation in nonlinear RNNs.** Panel (a) compares signal energies $Q^{\phi}_{n,t}$ for linear, ReLU, and LeakyReLU (slope $\alpha = 0.5$) recurrences. Panel (b) shows the corresponding feedforward models with independent weights $\mathbf{W}^{(t)}$ at each step t. Signal energies grow rapidly in the recurrent models but remain close to one in the feedforward models.

Nonlinear recurrences. We next test whether finite-width signal amplification persists beyond the linear setting. Namely, we compare the following recurrent models:

$$\mathbf{h}_{\phi}^{(t+1)} = \phi(\mathbf{W}\mathbf{h}_{\phi}^{(t)}), \qquad \mathbf{W}_{ij} \sim \mathcal{N}(0, \sigma^2/n), \tag{19}$$

where ϕ is linear, ReLU or LeakyReLU activation, and σ^2 is chosen according to the corresponding infinite-width variance-preserving initialization. Specifically, we use $\sigma^2 = 1$ for linear, $\sigma^2 = 2$ for ReLU, and $\sigma^2 = 2/(1+\alpha^2)$ for LeakyReLU with negative slope α. Figure 4(a) compares the behavior of the signal energies $Q_{n,t}^{\phi} := n^{-1}\mathbb{E}[\|\mathbf{h}_{\phi}^{(t)}\|_2^2]$ for the three activations. ReLU and LeakyReLU exhibit rapid signal growth, mirroring the effect observed for linear networks. Figure 4(b) plots signal energies of the corresponding feedforward models, where the weight matrices are drawn independently at each step t. In this setting, the signal energies remain close to one, as predicted by infinite-width theory. Together, these experiments illustrate how recurrent weight reuse drives the observed signal amplification.

5 Discussion

Our results give a sharp characterization of when infinite-width signal propagation theory remains predictive in long-range recurrent networks. For recurrent depth $t = o(\sqrt{n})$, signal energies follow their infinite-width predictions, while at the critical scale $t \sim \sqrt{n}$ a nontrivial joint scaling limit emerges. Thus, the $\sqrt{n}$ scale marks the transition at which finite-width effects become essential for describing signal propagation.

Importantly, the identified $\sqrt{n}$ scale is fundamentally a recurrent effect. In feedforward networks, where weights are independent across layers, finite-width effects become non-negligible only at depths of order n. Thus, our results quantify how repeatedly reusing the same weights across recurrent steps accelerates the accumulation of finite-width effects. Our experiments suggest that this mechanism persists beyond the linear setting: ReLU and LeakyReLU recurrences exhibit pronounced signal growth, absent in the corresponding feedforward models.

More broadly, our results show that width alone does not determine when infinite-width theory is predictive: the relevant notion of "large width" depends on how depth grows with width and how parameters are shared across layers. Long-range recurrent models provide a particularly sharp example of this principle, and offer a starting point for understanding finite-width effects in more general models beyond fully-connected feedforward networks.

5.1 Limitations and Future Work

Our analysis gives an exact finite-width characterization of linear recurrences with complex Gaussian weights and general input sequences, providing a tractable starting point for several extensions.

A first natural direction is to broaden the class of weight distributions. Our Remark 3 already shows that the critical $\sqrt{n}$ depth scale leads to non-negligible deviations from infinite-width predictions for real Gaussian weights. Obtaining full finite-width characterization or equally sharp asymptotics in the real case would further strengthen this. More generally, it would be interesting to study other initialization schemes, such as those used in modern linear recurrent architectures, including SSMs [16, 18].

A second direction is to extend the theory beyond linear recurrences. Although many modern long-range sequence models, including LRUs and SSMs, are built around linear recurrent blocks, classical recurrent architectures typically incorporate nonlinearities [7, 11]. Our numerical experiments in Figure 4 suggest that finite-width signal amplification persists under certain nonlinear activations. Characterizing signal propagation in nonlinear models under joint depth–width scaling theoretically is an natural direction for future work.

Finally, our results characterize signal propagation at random initialization. Extending the analysis to training could clarify how the finite-width effects identified here interact with the kernel (NTK) regime known to describe training dynamics at infinite width [1, 27], and when increasing recurrent depth breaks convergence to this regime. This may also provide a route towards understanding feature learning under joint recurrent-depth–width scaling.

Acknowledgements

This work was conducted as part of the Emmy Noether project "RMT4DL: Random Matrix Theory for Deep Learning." Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — Project number 572957656. Numerical experiments for this research were conducted on the Lichtenberg high-performance computing cluster of TU Darmstadt.

References

[1] Sina Alemohammad, Zichao Wang, Randall Balestriero, and Richard Baraniuk. The recurrent neural tangent kernel. In *The International Conference on Learning Representations*, 2021.

[2] Nikita Alexeev and Peter Zograf. Hultman numbers, polygon gluings and matrix integrals. *arXiv preprint arXiv:1111.3061*, 2011.

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[4] Noga Bar, Mariia Seleznova, Yotam Alexander, Gitta Kutyniok, and Raja Giryes. Revisiting glorot initialization for long-range linear recurrences. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[5] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[6] Minmin Chen, Jeffrey Pennington, and Samuel Schoenholz. Dynamical isometry and a mean field theory of rnns: Gating enables signal propagation in recurrent neural networks. In *International Conference on Machine Learning*, pages 873–882. PMLR, 2018.

[7] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[8] Soham De, Samuel L Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427*, 2024.

[9] Guillaume Dubach. On the number of cycles in commutators of random permutations. *The Annals of Applied Probability*, 34(4):4072–4084, 2024.

[10] Daniel Y Fu, Tri Dao, Khaled Kamal Saab, Armin W Thomas, Atri Rudra, and Christopher Re. Hungry hungry hippos: Towards language modeling with state space models. In *The Eleventh International Conference on Learning Representations*, 2023.

[11] Felix A Gers and E Schmidhuber. Lstm recurrent networks learn simple context-free and context-sensitive languages. *IEEE transactions on neural networks*, 12(6):1333–1340, 2001.

[12] Dar Gilboa, Bo Chang, Minmin Chen, Greg Yang, Samuel S Schoenholz, Ed H Chi, and Jeffrey Pennington. Dynamical isometry and a mean field theory of lstms and grus. *arXiv preprint arXiv:1901.08987*, 2019.

[13] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[14] Kevin Grosvenor and Ro Jefferson. The edge of chaos: quantum field theory and deep neural networks. *SciPost Physics*, 12(3):081, 2022.

[15] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[16] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33: 1474–1487, 2020.

[17] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

[18] Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35:35971–35983, 2022.

[19] Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? *Advances in neural information processing systems*, 31, 2018.

[20] Boris Hanin. Random fully connected neural networks as perturbatively solvable hierarchies. *Journal of Machine Learning Research*, 25(267):1–58, 2024.

[21] Boris Hanin and Mihai Nica. Finite depth and width corrections to the neural tangent kernel. *arXiv preprint arXiv:1909.05989*, 2019.

[22] Boris Hanin and Mihai Nica. Products of many large random matrices and gradients in deep neural networks. *Communications in Mathematical Physics*, 376(1):287–322, 2020.

[23] 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*, pages 1026–1034, 2015.

[24] 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*, pages 770–778, 2016.

[25] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[26] Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

[27] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[28] Alexander G de G Matthews, Mark Rowland, Jiri Hron, Richard E Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

[29] Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning*, pages 26670–26698. PMLR, 2023.

[30] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. Pmlr, 2013.

[31] 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.

[32] Daniel A Roberts, Sho Yaida, and Boris Hanin. *The principles of deep learning theory*, volume 46. Cambridge University Press Cambridge, MA, USA, 2022.

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

[34] Kai Segadlo, Bastian Epping, Alexander Van Meegen, David Dahmen, Michael Krämer, and Moritz Helias. Unified field theoretical approach to deep and recurrent neuronal networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(10):103401, 2022.

[35] Mariia Seleznova and Gitta Kutyniok. Neural tangent kernel beyond the infinite-width limit: Effects of depth and initialization. In *International Conference on Machine Learning*, pages 19522–19560. PMLR, 2022.

[36] Richard P Stanley. Two enumerative results on cycles of permutations. *European Journal of Combinatorics*, 32(6):937–943, 2011.

[37] Wojciech Tarnowski, Piotr Warchol, Stanislaw Jastrzebski, Jacek Tabor, and Maciej Nowak. Dynamical isometry is achieved in residual networks in a universal way for any activation function. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2221–2230. PMLR, 2019.

[38] Gian-Carlo Wick. The evaluation of the collision matrix. *Physical review*, 80(2):268, 1950.

[39] Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of cnns: How to train 10,000-layer vanilla convolutional neural networks. In *International Conference on Machine Learning*, pages 5393–5402. PMLR, 2018.

[40] Sho Yaida. Non-gaussian processes and neural networks at finite widths. In *Mathematical and Scientific Machine Learning*, pages 165–192. PMLR, 2020.

[41] Greg Yang. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019.

[42] Greg Yang. Tensor programs iii: Neural matrix laws. *arXiv preprint arXiv:2009.10685*, 2020.

[43] Greg Yang. Tensor programs ii: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020.

[44] Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pages 11727–11737. PMLR, 2021.

A Related Work

We review related work on signal propagation theory, both at infinite width and under finite-width corrections and joint depth–width limits, and the implications for stability in recurrent models.

A.1 Signal Propagation at Infinite Width

Existing signal propagation theory primarily concerns the infinite-width limit $n \to \infty$ at fixed depth L. In fully-connected networks, the behavior of signal energies in this regime was derived by Schoenholz et al. [33] and Poole et al. [31]. These analyses rely on the fact that preactivations in infinitely-wide networks become Gaussian, as shown rigorously e.g. in Matthews et al. [28]. The same perspective was subsequently extended to other architectures, including convolutional networks [39], residual networks [37], and recurrent networks [12]. More broadly, infinite-width signal propagation theory has been formulated in a unified way for general deep learning architectures within the Tensor Programs framework [41–44].

Recurrent models. Several works explicitly analyzed signal propagation in vanilla RNNs and gated recurrent units in the infinite-width, fixed-depth setting [1, 6, 12]. Recurrent models also fall within the Tensor Programs framework. As mentioned above, a common conclusion of these works is that, at infinite width, signal propagation in recurrent models behaves as if the weights were independent across recurrent time steps. That is, at the level of infinite-width signal propagation, recurrence is indistinguishable from feedforward composition. This observation is made explicit, for example, in Theorem 3 of Alemohammad et al. [1] and also in Segadlo et al. [34].

Stable initialization schemes. One of the main practical consequences of signal propagation theory is the principled design of stable initialization schemes, which prevent signals from exploding or vanishing with depth. This leads, in particular, to the familiar variance-preserving Gaussian initializations such as He initialization [23] and Glorot initialization [13].

A.2 Finite-Width Corrections and Joint Depth–Width Limits

A growing line of work studies how finite-width networks deviate from their infinite-width limits. A central question in this literature is how such deviations accumulate with depth, which naturally leads to the study of joint depth–width scaling limits in deep learning models.

Feedforward networks. The most extensively studied setting here is fully-connected feedforward networks. In this case, finite-width corrections become non-negligible when depth scales linearly with width, i.e., in the regime $L \sim cn$. First results in this direction were obtained by Hanin [19], who analyzed how fluctuations of signal energies accumulate with depth. Follow-up works extended this picture to broader settings and studied implications for signal stability and convergence to the kernel regime [20–22, 35, 40]. A more general perspective on finite-width corrections was later developed in Roberts et al. [32]. Overall, in fully-connected feedforward networks, finite-width corrections can be analyzed layer by layer and are therefore comparatively tractable.

Recurrent models. By contrast, the repeated reuse of the same matrix across time steps in recurrent models leads to a substantially more intricate structure of finite-width corrections and of their accumulation with depth. To the best of our knowledge, the only prior work to explicitly identify the $\sqrt{n}$ scale in linear recurrences is Bar et al. [4], where the authors used finite-width corrections to the spectral radius of Gaussian matrices to show that deviations from infinite-width behavior accumulate on the $t \sim c\sqrt{n}$ scale. However, that argument yields only a crude lower bound on the signal energy and does not recover the sharp finite-width behavior derived in the present work. More broadly, several earlier works investigated fluctuations around the infinite-width limit in recurrent networks and observed that they are larger than in feedforward architectures [14, 34]. However, these works do not identify an explicit depth scale at which infinite-width predictions break down, nor do they provide a sharp characterization of the size of finite-width corrections.

Stability of Glorot initialization in recurrent sequence models. Because standard infinite-width signal propagation theory does not distinguish feedforward and recurrent models, it predicts that

Glorot initialization [13] should remain stable in long-range linear recurrences. However, in practice, recent works on recurrent sequence models have found that Glorot is unstable and have therefore introduced additional rescaling. For example, Orvieto et al. [29] used truncated Glorot, which effectively contracts the spectrum by roughly 0.8, while Gu et al. [17] explicitly report iteratively reducing the Gaussian variance until the model no longer explodes. More recently, Bar et al. [4] proposed a theoretically motivated approach to such rescaling. At the same time, state-of-the-art linear recurrent sequence models typically rely on structured non-Gaussian initializations, while rescaled variants of Glorot are used mainly as baselines [10, 15–17].

B Proof of Theorem 2.1

The central part of the proof is the exact evaluation of the trace moment $\mathbb{E}\big[\mathrm{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big)\big]$ for finite n and k, given in the following Subsections B.1 and B.2. The remaining part of the proof, which reduces the signal energies $Q_{n,k}$ and $S_{n,t}$ to the trace moments is given in Subsection B.3.

B.1 Trace moment via Wick expansion and free-index counting

In this subsection, we carry out the standard reduction of the trace moment to a combinatorial counting problem. Namely, we expand the trace into monomials in the Gaussian entries of $\mathbf{W}$, apply Wick's theorem, and then count how many indices remain free after imposing the Kronecker constraints associated with each Wick pairing.

We begin by expanding the trace as follows:

$$\mathrm{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big) = \sum_{a_0,a_k=1}^{n} (\mathbf{W}^k)^*_{a_k a_0} (\mathbf{W}^k)_{a_0 a_k} \tag{20}$$

$$= \sum_{a_0,a_k=1}^{n} \overline{(\mathbf{W}^k)_{a_0 a_k}}\, (\mathbf{W}^k)_{a_0 a_k} \tag{21}$$

$$= \sum_{\substack{a_0,\dots,a_k=1\\ b_1,\dots,b_{k-1}=1}}^{n} \overline{\mathbf{W}_{a_0 a_1} \mathbf{W}_{a_1 a_2} \cdots \mathbf{W}_{a_{k-1} a_k}}\, \mathbf{W}_{a_0 b_1} \mathbf{W}_{b_1 b_2} \cdots \mathbf{W}_{b_{k-1} a_k} \tag{22}$$

$$= \sum_{\substack{a_0,\dots,a_k=1\\ b_1,\dots,b_{k-1}=1}}^{n} \overline{\mathbf{W}}_{a_0 a_1}\, \overline{\mathbf{W}}_{a_1 a_2} \cdots \overline{\mathbf{W}}_{a_{k-1} a_k}\, \mathbf{W}_{a_0 b_1} \mathbf{W}_{b_1 b_2} \cdots \mathbf{W}_{b_{k-1} a_k}. \tag{23}$$

Thus the trace moment reduces to expectations of monomials in the entries of $\mathbf{W}$, which are centered proper complex Gaussians with variance $1/n$ under the complex Glorot initialization (3).

For such monomials, Wick's theorem takes a particularly simple form, specified in the next lemma.

Lemma B.1 (Wick theorem for proper complex Gaussians)**.** *Let $(Z_1, \dots, Z_m)$ be a centered proper complex Gaussian vector, meaning that $\mathbb{E}[Z_i Z_j] = 0$ for all $i, j \in \{1, \dots, m\}$. Let $I = (I_1, \dots, I_k)$ and $J = (J_1, \dots, J_r)$ be arbitrary index tuples with entries in $\{1, \dots, m\}$. Then*

$$\mathbb{E}\left[\prod_{a=1}^{k} \overline{Z_{I_a}} \prod_{b=1}^{r} Z_{J_b}\right] = \delta_{k,r} \sum_{\sigma \in S_k} \prod_{a=1}^{k} \mathbb{E}\big[\overline{Z_{I_a}}\, Z_{J_{\sigma(a)}}\big]\,, \tag{24}$$

where S_k denotes the set of permutations of $\{1, \dots, k\}$.

Proof. We apply the usual Wick theorem to the family obtained by listing together the conjugated and non-conjugated Gaussian variables. Define the *doubled family*

$$X_1, \dots, X_{k+r} := (\overline{Z_{I_1}}, \dots, \overline{Z_{I_k}}, Z_{J_1}, \dots, Z_{J_r}). \tag{25}$$

Applying Wick's theorem to this family yields

$$\mathbb{E}\left[\prod_{a=1}^{k} \overline{Z_{I_a}} \prod_{b=1}^{r} Z_{J_b}\right] = \sum_{\pi \in \mathcal{P}_2(k+r)} \prod_{\{p,q\} \in \pi} \mathbb{E}[X_p X_q], \tag{26}$$

where $\mathcal{P}_2(k+r)$ denotes the set of all pairings of $\{1,\dots,k+r\}$.

Since $(Z_1,\dots,Z_m)$ is proper complex Gaussian, we have

$$\mathbb{E}[Z_p Z_q] = 0, \qquad \mathbb{E}[\overline{Z_p}\,\overline{Z_q}] = \overline{\mathbb{E}[Z_p Z_q]} = 0 \tag{27}$$

for all $p,q \in \{1,\dots,m\}$. Therefore every pairing containing a pair of the form $(\overline{Z_p},\overline{Z_q})$ or (Z_p,Z_q) contributes zero to the right-hand side of (26). Hence only pairings consisting entirely of mixed pairs $(\overline{Z_{I_a}}, Z_{J_b})$ can survive.

Such pairings exist only if $k=r$. When $k=r$, every surviving pairing is uniquely determined by a permutation $\sigma \in S_k$, namely by matching

$$\overline{Z_{I_a}} \quad \text{with} \quad Z_{J_{\sigma(a)}}, \qquad a = 1,\dots,k. \tag{28}$$

Substituting these surviving pairings into (26) gives

$$\mathbb{E}\left[\prod_{a=1}^{k}\overline{Z_{I_a}}\prod_{b=1}^{r} Z_{J_b}\right] = \delta_{k,r}\sum_{\sigma\in S_k}\prod_{a=1}^{k}\mathbb{E}\left[\overline{Z_{I_a}}\, Z_{J_{\sigma(a)}}\right], \tag{29}$$

as claimed. □

Applying Lemma B.1 to the monomials appearing in the trace expansion, we see that the expectation is indexed by permutations $\sigma \in S_k$, each of which prescribes one possible pairing between the k conjugated factors and the k non-conjugated factors. The contribution of a fixed permutation σ is therefore

$$\sum_{\substack{a_0,\dots,a_k=1\\ b_1,\dots,b_{k-1}=1}}^{n}\prod_{r=1}^{k}\mathbb{E}\left[\overline{\mathbf{W}}_{a_{r-1}a_r}\,\mathbf{W}_{b_{\sigma(r)-1}b_{\sigma(r)}}\right]. \tag{30}$$

Since the entries of $\mathbf{W}$ are i.i.d. proper complex Gaussians with variance $1/n$, each covariance equals

$$\mathbb{E}\left[\overline{\mathbf{W}}_{ij}\mathbf{W}_{pq}\right] = \frac{1}{n}\,\delta_{i,p}\delta_{j,q}.$$

Hence the contribution of σ becomes

$$\sum_{\substack{a_0,\dots,a_k=1\\ b_1,\dots,b_{k-1}=1}}^{n}\prod_{r=1}^{k}\frac{1}{n}\delta_{a_{r-1},\,b_{\sigma(r)-1}}\delta_{a_r,\,b_{\sigma(r)}}. \tag{31}$$

Here we use the convention

$$b_0 := a_0, \qquad b_k := a_k,$$

so that all indices appearing in the contractions are among

$$a_0,\dots,a_k,b_1,\dots,b_{k-1}.$$

The key combinatorial quantity is the number of indices that remain unconstrained after all the Kronecker contractions have been imposed, as defined below.

Definition B.2 (Free-index count and defect of a permutation)**.** *For $\sigma \in S_k$, we define $F(\sigma)$ to be the number of **free indices**, namely the number of equivalence classes of indices that remain unconstrained after all δ-contractions associated with σ have been imposed. The corresponding **defect** is defined by*

$$r(\sigma) := k + 1 - F(\sigma). \tag{32}$$

Since each free index contributes a factor of n, whereas each covariance contributes a factor of n^{-1}, the total contribution of σ is exactly

$$n^{-k+F(\sigma)} = n^{1-r(\sigma)}. \tag{33}$$

After dividing by the outer factor $1/n$, we obtain the following representation of the normalized trace moment.

Lemma B.3 (Trace moment as a free-index sum)**.** *For every $k \geq 0$,*

$$\mathbb{E}\left[\frac{1}{n}\operatorname{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big)\right] = \sum_{\sigma\in S_k} n^{-r(\sigma)}, \tag{34}$$

where $r(\sigma) = k + 1 - F(\sigma)$ and $F(\sigma)$ is the number of free indices left after the Wick contractions corresponding to σ.

Lemma B.3 reduces the problem to understanding the free-index count $F(\sigma)$, or equivalently the defect $r(\sigma)$, associated with each permutation $\sigma \in S_k$. In the next subsection, we show that $F(\sigma)$ can be identified with the number of cycles of a certain commutator of permutations, which allows the trace moment to be evaluated exactly.

B.2 Exact evaluation of the trace moment via commutator cycle counts

By Lemma B.3, the trace moment is determined by the free-index count $F(\sigma)$ associated with each permutation $\sigma \in S_k$. We now show that this quantity admits a simple permutation-theoretic interpretation: it is exactly the number of cycles of a certain commutator. This reduces the trace computation to a cycle-counting problem, which can then be evaluated using known results on random permutations.

Lemma B.4. *For a permutation $\sigma \in S_k$, the free-index count $F(\sigma)$ is exactly the number of cycles of*

$$[\tau^{-1}, \tilde{\sigma}^{-1}] := \tau^{-1}\tilde{\sigma}^{-1}\tau\tilde{\sigma}, \tag{35}$$

where $\tau = (0, 1, \ldots, k) \in S_{k+1}$ and $\tilde{\sigma} \in S_{k+1}$ is the extension of σ fixing 0, i.e.

$$\tilde{\sigma}(0) = 0, \qquad \tilde{\sigma}(i) = \sigma(i), \quad i = 1, \ldots, k. \tag{36}$$

Proof. Recall from the previous subsection that a permutation $\sigma \in S_k$ imposes the Kronecker constraints

$$a_{i-1} = b_{\sigma(i)-1}, \qquad a_i = b_{\sigma(i)}, \qquad i = 1, \ldots, k, \tag{37}$$

together with the boundary conditions

$$b_0 = a_0, \qquad b_k = a_k.$$

The free-index count $F(\sigma)$ is the number of equivalence classes generated by these identifications.

To rewrite the constraints more symmetrically, set $j = \sigma(i)$. Then

$$a_{\sigma^{-1}(j)-1} = b_{j-1}, \qquad a_{\sigma^{-1}(j)} = b_j, \qquad j = 1, \ldots, k. \tag{38}$$

Eliminating the b-variables yields

$$a_{\sigma^{-1}(j-1)} = a_{\sigma^{-1}(j)-1}, \qquad j = 2, \ldots, k, \tag{39}$$

while the two boundary conditions become

$$a_0 = a_{\sigma^{-1}(1)-1}, \qquad a_{\sigma^{-1}(k)} = a_k.$$

These relations can be written uniformly by extending σ to $\tilde{\sigma} \in S_{k+1}$ and introducing the full cycle $\tau = (0, 1, \ldots, k)$. Then for every $j \in \{0, \ldots, k\}$,

$$a_{\tilde{\sigma}^{-1}(j)} = a_{\tau^{-1}\tilde{\sigma}^{-1}\tau(j)}. \tag{40}$$

Indeed:

- for $j = 0$, this reads $a_0 = a_{\sigma^{-1}(1)-1}$;
- for $j = 1, \ldots, k-1$, this reads $a_{\sigma^{-1}(j)} = a_{\sigma^{-1}(j+1)-1}$;
- for $j = k$, this reads $a_{\sigma^{-1}(k)} = a_k$.

Now let

$$m = \tilde{\sigma}^{-1}(j).$$

Since $j = \tilde{\sigma}(m)$, the previous identity becomes

$$a_m = a_{\tau^{-1}\tilde{\sigma}^{-1}\tau\tilde{\sigma}(m)} = a_{[\tau^{-1},\tilde{\sigma}^{-1}](m)}, \qquad m = 0, \dots, k. \tag{41}$$

Hence the equivalence classes among the indices $\{a_0, \dots, a_k\}$ are exactly the orbits of the permutation $[\tau^{-1}, \tilde{\sigma}^{-1}]$. Therefore the number of free indices is precisely the number of cycles of this commutator. □

Lemma B.4 reduces the trace problem to counting cycles of the commutator $[\tau^{-1}, \tilde{\sigma}^{-1}]$. The next step is to relate this commutator to a simpler random permutation model that has already been analyzed in the literature. Namely, we use the following known cycle-count formula, which is a reformulation of Theorem 2 in Dubach [9].

Lemma B.5. *Let $\tau = (0, \dots, M) \in S_{M+1}$, with $M \geq 0$, and let $c \in S_{M+1}$ be uniformly distributed among all $(M+1)$-cycles. Then the number of cycles of $\tau^{-1}c$, denoted $\mathcal{C}(\tau^{-1}c)$, satisfies*

$$\mathbb{E}[t^{\mathcal{C}(\tau^{-1}c)}] = \frac{1}{(M+2)!}\big(F^{+}_{M+2}(t) - F^{-}_{M+2}(t)\big), \tag{42}$$

where $F^{\pm}_n(t)$ are the rising and falling factorials, defined by

$$F^{+}_n(t) := t(t+1)\dots(t+n-1), \qquad F^{-}_n(t) := t(t-1)\dots(t-n+1). \tag{43}$$

We note that Theorem 2 in Dubach [9] is stated in terms of the number of cycles of the commutator

$$[\rho, \tau] := \rho\tau\rho^{-1}\tau^{-1},$$

where ρ is uniformly distributed over S_{M+1}, rather than in terms of $\tau^{-1}c$ as in the formulation above. However, as also observed in Dubach [9], the two formulations are equivalent. Indeed, one can write

$$[\rho, \tau] = \tau_1\tau_2, \qquad \tau_1 := \rho\tau\rho^{-1}, \qquad \tau_2 := \tau^{-1}. \tag{44}$$

Since τ_1 is conjugate to τ, it is uniformly distributed over the conjugacy class of τ, that is, over all permutations with the same cycle structure as τ. In our setting, τ is a single $(M+1)$-cycle, so τ_1 is uniform among all $(M+1)$-cycles. Therefore the distribution of the number of cycles of $[\rho, \tau]$ is exactly the same as that of $\tau^{-1}c$, where c is uniform among all $(M+1)$-cycles, which yields the stated reformulation.

It remains to show that the commutator in Lemma B.4 fits into this framework. The key observation is that conjugating the full cycle τ by a permutation fixing 0 produces a uniformly random full cycle.

Lemma B.6. *Let $\tau = (0, \dots, k)$ and let $\tilde{\sigma} \in S_{k+1}$ be uniform on the set*

$$H := \{\pi \in S_{k+1} : \pi(0) = 0\}. \tag{45}$$

Then $\tilde{\sigma}^{-1}\tau\tilde{\sigma}$ is uniform among all $(k+1)$-cycles in S_{k+1}.

Proof. The set H has cardinality $k!$, and there are also exactly $k!$ many $(k+1)$-cycles in S_{k+1}. It is therefore enough to show that the map

$$\phi_\tau : H \to \{(k+1)\text{-cycles in } S_{k+1}\}, \qquad \phi_\tau(\pi) = \pi^{-1}\tau\pi, \tag{46}$$

is injective.

Suppose $\phi_\tau(\pi_1) = \phi_\tau(\pi_2)$ for some $\pi_1, \pi_2 \in H$. Then

$$\pi_2\pi_1^{-1}\tau = \tau\pi_2\pi_1^{-1}, \tag{47}$$

so $\pi_2\pi_1^{-1}$ belongs to the centralizer of the full cycle τ. But the centralizer of τ is exactly

$$\{\tau^m : m = 0, \dots, k\}.$$

Among these elements, only the identity fixes 0. Since both π_1 and π_2 fix 0, so does $\pi_2\pi_1^{-1}$, and therefore $\pi_2\pi_1^{-1} = \mathrm{id}$. Hence $\pi_1 = \pi_2$, proving injectivity. Thus ϕ_τ is bijective, and the image of a uniform $\tilde{\sigma} \in H$ is uniform among all $(k+1)$-cycles. □

We can now combine the previous two lemmas to evaluate the cycle count appearing in Lemma B.4.

Lemma B.7. *Let $\tau = (0, \dots, k)$, let σ range uniformly over S_k, and let $\tilde{\sigma}$ be its extension to S_{k+1} fixing 0. Then*

$$\sum_{\sigma \in S_k} n^{\mathcal{C}([\tau^{-1}, \tilde{\sigma}^{-1}])} = \frac{k!}{(k+2)!}\big(F^+_{k+2}(n) - F^-_{k+2}(n)\big). \tag{48}$$

Proof. Write

$$c := \tilde{\sigma}^{-1}\tau\tilde{\sigma}.$$

Then

$$[\tau^{-1}, \tilde{\sigma}^{-1}] = \tau^{-1}\tilde{\sigma}^{-1}\tau\tilde{\sigma} = \tau^{-1}c. \tag{49}$$

If σ is uniform on S_k, then $\tilde{\sigma}$ is uniform on the set H from (45), and therefore c is uniform among all $(k+1)$-cycles by Lemma B.6. It follows that

$$\frac{1}{|S_k|}\sum_{\sigma \in S_k} n^{\mathcal{C}([\tau^{-1}, \tilde{\sigma}^{-1}])} = \mathbb{E}\big[n^{\mathcal{C}(\tau^{-1}c)}\big]. \tag{50}$$

Now apply Lemma B.5 with $M = k$ and $t = n$:

$$\mathbb{E}\big[n^{\mathcal{C}(\tau^{-1}c)}\big] = \frac{1}{(k+2)!}\big(F^+_{k+2}(n) - F^-_{k+2}(n)\big). \tag{51}$$

Multiplying by $|S_k| = k!$ gives the claim. □

We are now ready to complete the trace computation.

Proposition B.8. *For every $k \geq 0$,*

$$\mathbb{E}\left[\frac{1}{n}\operatorname{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big)\right] = \frac{F^+_{k+2}(n) - F^-_{k+2}(n)}{n^{k+1}(k+1)(k+2)}. \tag{52}$$

Proof. By Lemma B.3,

$$\mathbb{E}\left[\frac{1}{n}\operatorname{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big)\right] = \sum_{\sigma \in S_k} n^{-r(\sigma)}. \tag{53}$$

Using the definition of the defect and Lemma B.4, this becomes

$$\mathbb{E}\left[\frac{1}{n}\operatorname{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big)\right] = n^{-k-1}\sum_{\sigma \in S_k} n^{F(\sigma)} = n^{-k-1}\sum_{\sigma \in S_k} n^{\mathcal{C}([\tau^{-1}, \tilde{\sigma}^{-1}])}. \tag{54}$$

Applying Lemma B.7 yields

$$\mathbb{E}\left[\frac{1}{n}\operatorname{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big)\right] = n^{-k-1} \cdot \frac{k!}{(k+2)!}\big(F^+_{k+2}(n) - F^-_{k+2}(n)\big), \tag{55}$$

which simplifies to

$$\mathbb{E}\left[\frac{1}{n}\operatorname{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big)\right] = \frac{F^+_{k+2}(n) - F^-_{k+2}(n)}{n^{k+1}(k+1)(k+2)}. \tag{56}$$

□

In the next subsection, we show how Proposition B.8 yields the formulas for $Q_{n,k}$ and $S_{n,t}$ stated in Theorem 2.1.

B.3 Reduction of $Q_{n,k}$ and $S_{n,t}$ to a trace moment

We now return to the signal propagation quantities $Q_{n,k}$ and $S_{n,t}$ and express them in terms of the trace moments computed in the previous subsections, which was evaluated explicitly in Proposition B.8.

Lemma B.9. *Consider a linear RNN (as defined in (2)) with weight matrix $\mathbf{W} \in \mathbb{C}^{n\times n}$, initialized with i.i.d. proper complex Gaussian entries of variance $1/n$ (as in (3)). Then the RNN signal energy $Q_{n,k}$ (defined in (4)) satisfies*

$$Q_{n,k} = \frac{1}{n}\operatorname{tr}(\Sigma^{(0)}) \cdot \frac{1}{n}\,\mathbb{E}\Big[\operatorname{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big)\Big], \tag{57}$$

where $\Sigma^{(0)} := \mathbb{E}\big[\mathbf{x}^{(0)}(\mathbf{x}^{(0)})^\top\big]$ is the input covariance.

Proof. We begin by rewriting the squared norm of the hidden state as a trace:

$$\|\mathbf{h}_{\text{RNN}}^{(k)}\|_2^2 = \|\mathbf{W}^k\mathbf{x}^{(0)}\|_2^2 = \operatorname{tr}\Big(\mathbf{x}^{(0)}(\mathbf{x}^{(0)})^\top(\mathbf{W}^k)^*\mathbf{W}^k\Big). \tag{58}$$

Taking expectations and using the independence between $\mathbf{x}^{(0)}$ and $\mathbf{W}$, we obtain

$$Q_{n,k} = \frac{1}{n}\mathbb{E}\big[\|\mathbf{h}_{\text{RNN}}^{(k)}\|_2^2\big] = \frac{1}{n}\operatorname{tr}\Big(\mathbb{E}\big[\mathbf{x}^{(0)}(\mathbf{x}^{(0)})^\top\big]\,\mathbb{E}\big[(\mathbf{W}^k)^*\mathbf{W}^k\big]\Big). \tag{59}$$

Thus it remains to understand the matrix

$$\mathbb{E}\big[(\mathbf{W}^k)^*\mathbf{W}^k\big].$$

We notice that this matrix is proportional to the identity, i.e.

$$\mathbb{E}\big[(\mathbf{W}^k)^*\mathbf{W}^k\big] = \alpha\mathbb{I}_n \tag{60}$$

for some scalar $\alpha \in \mathbb{C}$. Indeed, by unitary invariance of the complex Gaussian law, for every fixed unitary matrix U we have

$$U\mathbf{W}U^* \overset{d}{=} \mathbf{W}. \tag{61}$$

Raising to the k-th power and taking adjoints gives

$$(\mathbf{W}^k)^*\mathbf{W}^k \overset{d}{=} U(\mathbf{W}^k)^*\mathbf{W}^kU^*. \tag{62}$$

Taking expectations, we conclude that for every unitary U,

$$\mathbb{E}\big[(\mathbf{W}^k)^*\mathbf{W}^k\big] = U\mathbb{E}\big[(\mathbf{W}^k)^*\mathbf{W}^k\big]U^*. \tag{63}$$

Equivalently,

$$\mathbb{E}\big[(\mathbf{W}^k)^*\mathbf{W}^k\big]\,U = U\,\mathbb{E}\big[(\mathbf{W}^k)^*\mathbf{W}^k\big], \tag{64}$$

so $\mathbb{E}\big[(\mathbf{W}^k)^*\mathbf{W}^k\big]$ commutes with every unitary matrix. This is only possible if it is proportional to the identity.

Substituting this back into the trace formula yields

$$Q_{n,k} = \frac{1}{n}\operatorname{tr}\big(\Sigma^{(0)}\,\alpha I_n\big) = \frac{1}{n}\operatorname{tr}(\Sigma^{(0)})\,\alpha. \tag{65}$$

On the other hand,

$$\operatorname{tr}\Big(\mathbb{E}\big[(\mathbf{W}^k)^*\mathbf{W}^k\big]\Big) = n\alpha, \tag{66}$$

so

$$\alpha = \frac{1}{n}\operatorname{tr}\Big(\mathbb{E}\big[(\mathbf{W}^k)^*\mathbf{W}^k\big]\Big) = \frac{1}{n}\,\mathbb{E}\operatorname{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big). \tag{67}$$

Combining the two identities gives

$$Q_{n,k} = \frac{1}{n}\operatorname{tr}(\Sigma^{(0)}) \cdot \frac{1}{n}\,\mathbb{E}\Big[\operatorname{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big)\Big], \tag{68}$$

as claimed. □

The next lemma gives the analogous reduction for the LRU signal energy. The only new point is that, after expanding the squared norm, one must show that the mixed terms with different powers of $\mathbf{W}$ vanish.

Lemma B.10. *Consider an LRU (as defined in (1)) with weight matrix $\mathbf{W} \in \mathbb{C}^{n\times n}$, initialized with i.i.d. proper complex Gaussian entries of variance $1/n$ (as in (3)). Then the LRU signal energy $S_{n,t}$ (defined in (4)) satisfies*

$$S_{n,t} = \sum_{k=0}^{t} \frac{1}{n} \operatorname{tr}\big(\Sigma^{(t-k)}\big)\, \tilde{Q}_{n,k}, \tag{69}$$

where $\Sigma^{(k)} := \mathbb{E}\big[\mathbf{x}^{(k)}(\mathbf{x}^{(k)})^\top\big]$ are the input covariances for all time steps $k = 0, \dots, t$, and $\tilde{Q}_{n,k}$ is the RNN signal energy under whitened data, i.e. when $\Sigma^{(0)} = \mathbb{I}_n$.

Proof. We start from the explicit representation of the LRU hidden state:

$$\mathbf{h}_{\mathrm{LRU}}^{(t)} = \sum_{k=0}^{t} \mathbf{W}^k \mathbf{x}^{(t-k)}. \tag{70}$$

Expanding its squared norm gives

$$\|\mathbf{h}_{\mathrm{LRU}}^{(t)}\|_2^2 = \sum_{k,r=0}^{t} (\mathbf{x}^{(t-r)})^\top (\mathbf{W}^r)^* \mathbf{W}^k \mathbf{x}^{(t-k)} \tag{71}$$

$$= \sum_{k,r=0}^{t} \operatorname{tr}\Big(\mathbf{x}^{(t-k)}(\mathbf{x}^{(t-r)})^\top (\mathbf{W}^r)^* \mathbf{W}^k\Big). \tag{72}$$

Taking expectations and using the independence between the inputs and the weight matrix, we obtain

$$\mathbb{E}\big[\|\mathbf{h}_{\mathrm{LRU}}^{(t)}\|_2^2\big] = \sum_{k,r=0}^{t} \operatorname{tr}\Big(\mathbb{E}\big[\mathbf{x}^{(t-k)}(\mathbf{x}^{(t-r)})^\top\big]\, \mathbb{E}\big[(\mathbf{W}^r)^* \mathbf{W}^k\big]\Big). \tag{73}$$

We now show that the mixed terms with $k \neq r$ vanish. As in Lemma B.9, unitary invariance implies that $\mathbb{E}\big[(\mathbf{W}^r)^*\mathbf{W}^k\big]$ is proportional to identity. It therefore suffices to check that all diagonal entries are zero whenever $k \neq r$.

Fix $i \in \{1, \dots, n\}$. Then

$$\big[(\mathbf{W}^r)^*\mathbf{W}^k\big]_{ii} = \sum_{a_0=1}^{n} \overline{(\mathbf{W}^r)_{a_0 i}}\, (\mathbf{W}^k)_{a_0 i} \tag{74}$$

$$= \sum_{\substack{a_0,\dots,a_{r-1}=1 \\ b_1,\dots,b_{k-1}=1}}^{n} \overline{\mathbf{W}_{a_0 a_1}\mathbf{W}_{a_1 a_2}} \cdots \overline{\mathbf{W}_{a_{r-1} i}} \mathbf{W}_{a_0 b_1} \mathbf{W}_{b_1 b_2} \cdots \mathbf{W}_{b_{k-1} i}. \tag{75}$$

The family of Gaussian variables appearing here is proper complex Gaussian, so Lemma B.1 applies. Since there are r conjugated factors and k non-conjugated factors, the expectation vanishes whenever $k \neq r$. Therefore,

$$\mathbb{E}\big[(\mathbf{W}^r)^*\mathbf{W}^k\big] = 0 \qquad \text{for all } k \neq r. \tag{76}$$

Thus only the terms with $k = r$ remain, and the signal energy simplifies to

$$S_{n,t} = \frac{1}{n}\mathbb{E}\big[\|\mathbf{h}_{\mathrm{LRU}}^{(t)}\|_2^2\big] = \frac{1}{n}\sum_{k=0}^{t} \operatorname{tr}\Big(\mathbb{E}\big[\mathbf{x}^{(t-k)}(\mathbf{x}^{(t-k)})^\top\big]\, \mathbb{E}\big[(\mathbf{W}^k)^*\mathbf{W}^k\big]\Big). \tag{77}$$

At this point, we can apply exactly the same argument as in Lemma B.9 to each term of the sum, obtaining

$$S_{n,t} = \sum_{k=0}^{t} \frac{1}{n} \operatorname{tr}(\Sigma^{(t-k)}) \cdot \frac{1}{n}\, \mathbb{E}\Big[\operatorname{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big)\Big]. \tag{78}$$

By definition, the quantity

$$\tilde{Q}_{n,k} = \frac{1}{n}\,\mathbb{E}\Big[\mathrm{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big)\Big] \tag{79}$$

is precisely the RNN signal energy corresponding to whitened inputs. Substituting this into the previous expression gives

$$S_{n,t} = \sum_{k=0}^{t} \frac{1}{n}\,\mathrm{tr}\big(\Sigma^{(t-k)}\big)\,\tilde{Q}_{n,k}, \tag{80}$$

as claimed. □

Finally, Proposition B.8 can now be substituted into Lemmas B.9 and B.10, which yields the explicit formulas for $Q_{n,k}$ and $S_{n,t}$ stated in Theorem 2.1.

B.4 Real case: Proof of Remark 3

Proof. We first treat the RNN quantity $Q_{n,k}$. Let $\mathbf{W}^{\mathbb{R}}$ denote a real Gaussian matrix with i.i.d. entries of variance $1/n$, and let

$$Q^{\mathbb{R}}_{n,k} = \frac{1}{n}\,\mathbb{E}\big[\|\left(\mathbf{W}^{\mathbb{R}}\right)^k \mathbf{x}^{(0)}\|_2^2\big].$$

As in Lemma B.9, orthogonal invariance implies that

$$\mathbb{E}\Big[\big((\mathbf{W}^{\mathbb{R}})^k\big)^\top(\mathbf{W}^{\mathbb{R}})^k\Big] = \alpha^{\mathbb{R}}_k I_n$$

for some scalar $\alpha^{\mathbb{R}}_k \geq 0$, and therefore

$$Q^{\mathbb{R}}_{n,k} = \frac{1}{n}\,\mathrm{tr}(\Sigma^{(0)})\,\frac{1}{n}\mathbb{E}\big[\mathrm{tr}\big(((\mathbf{W}^{\mathbb{R}})^k)^\top(\mathbf{W}^{\mathbb{R}})^k\big)\big]\,.$$

So it suffices to compare the corresponding trace moments.

Expanding the trace exactly as in Appendix B.1, we obtain a sum of monomials in the entries of $\mathbf{W}$. In the complex case, by Lemma B.1, only pairings between conjugated and non-conjugated factors contribute. In the real Gaussian case, Wick's theorem sums instead over *all* pairings of the $2k$ factors. Since

$$\mathbb{E}[\,\mathbf{W}^{\mathbb{R}}_{ij}\mathbf{W}^{\mathbb{R}}_{pq}\,] = \frac{1}{n}\delta_{i,p}\delta_{j,q} \geq 0,$$

every contributing pairing in the real case gives a nonnegative contribution. Moreover, the pairings that survive in the complex case form a subset of the pairings appearing in the real case. Hence

$$\mathbb{E}\left[\frac{1}{n}\,\mathrm{tr}\big(((\mathbf{W}^{\mathbb{R}})^k)^\top(\mathbf{W}^{\mathbb{R}})^k\big)\right] \geq \mathbb{E}\left[\frac{1}{n}\,\mathrm{tr}\big((\mathbf{W}^k)^*\mathbf{W}^k\big)\right],$$

and therefore

$$Q^{\mathbb{R}}_{n,k} \geq Q_{n,k}.$$

We now turn to the LRU quantity $S^{\mathbb{R}}_{n,t}$. Expanding the squared norm as in Lemma B.10, we obtain

$$S^{\mathbb{R}}_{n,t} = \frac{1}{n}\sum_{k,r=0}^{t}\mathrm{tr}\Big(\mathbb{E}[\mathbf{x}^{(t-k)}(\mathbf{x}^{(t-r)})^\top]\,\mathbb{E}[((\mathbf{W}^{\mathbb{R}})^r)^\top(\mathbf{W}^{\mathbb{R}})^k]\Big).$$

Again by orthogonal invariance,

$$\mathbb{E}[((\mathbf{W}^{\mathbb{R}})^r)^\top(\mathbf{W}^{\mathbb{R}})^k] = \alpha^{\mathbb{R}}_{k,r} I_n$$

for some scalar $\alpha^{\mathbb{R}}_{k,r} \geq 0$. Indeed,

$$\alpha^{\mathbb{R}}_{k,r} = \frac{1}{n}\mathbb{E}\big[\mathrm{tr}\big(((\mathbf{W}^{\mathbb{R}})^r)^\top(\mathbf{W}^{\mathbb{R}})^k\big)\big]\,,$$

and this trace moment is again a sum of nonnegative Wick-pairing contributions. Therefore

$$S^{\mathbb{R}}_{n,t} = \sum_{k,r=0}^{t}\frac{1}{n}\,\mathrm{tr}\big(\mathbb{E}[\mathbf{x}^{(t-k)}(\mathbf{x}^{(t-r)})^\top]\big)\,\alpha^{\mathbb{R}}_{k,r}.$$

Under the assumption

$$\operatorname{tr}\big(\mathbb{E}[\mathbf{x}^{(s)}(\mathbf{x}^{(r)})^\top]\big) \geq 0 \qquad \text{for all } 0 \leq s, r \leq t,$$

each summand is nonnegative. In particular, restricting to the diagonal terms $k = r$ gives

$$S^{\mathbb{R}}_{n,t} \geq \sum_{k=0}^{t} \frac{1}{n} \operatorname{tr}\big(\mathbb{E}[\mathbf{x}^{(t-k)}(\mathbf{x}^{(t-k)})^\top]\big)\, \alpha^{\mathbb{R}}_{k,k}.$$

By the already established inequality for the diagonal trace moments,

$$\alpha^{\mathbb{R}}_{k,k} \geq \tilde{Q}_{n,k},$$

where $\tilde{Q}_{n,k}$ is the corresponding complex RNN signal energy under whitened inputs. Hence

$$S^{\mathbb{R}}_{n,t} \geq \sum_{k=0}^{t} \frac{1}{n} \operatorname{tr}\big(\mathbb{E}[\mathbf{x}^{(t-k)}(\mathbf{x}^{(t-k)})^\top]\big)\, \tilde{Q}_{n,k} = S_{n,t}.$$

This proves

$$Q^{\mathbb{R}}_{n,k} \geq Q_{n,k}, \qquad S^{\mathbb{R}}_{n,t} \geq S_{n,t}.$$

The final claim follows immediately: if the complex model exhibits exploding signal energy in a given depth–width scaling regime, then so does the real model. □

C Proofs for Scaling Regimes of $Q_{n,k}$

This section contains the proofs of the asymptotic statements for the signal energy $Q_{n,k}$ in the three joint (k, n) scaling regimes identified in the main text: subcritical, critical, and supercritical. Each proof is obtained by a direct asymptotic analysis of the exact finite-width formula for $Q_{n,k}$ given in Theorem 2.1.

Theorem C.1 (Subcritical regime)**.** *If $k = o(\sqrt{n})$, we have*

$$Q_{n,k} \to 1. \tag{81}$$

Proof. Let us define $P_\pm = \prod_{j=0}^{k+1}(1 \pm j/n)$, then we have:

$$Q_{n,k} = \frac{n}{(k+1)(k+2)}[P_+ - P_-]. \tag{82}$$

As $k \ll n$, we can get the following by Taylor expansion of the logarithm:

$$\log P_\pm = \sum_{j=0}^{k+1} \log\Big(1 \pm \frac{j}{n}\Big) = \sum_{j=0}^{k+1} \Big(\pm\frac{j}{n} + O\Big(\frac{j^2}{n^2}\Big)\Big). \tag{83}$$

Under $k = o(\sqrt{n})$, we have the following for the terms above:

$$a_{n,k} := \sum_{j=0}^{k+1} \frac{j}{n} = \frac{(k+1)(k+2)}{2n} = O\Big(\frac{k^2}{n}\Big), \quad b_{n,k} := \sum_{j=0}^{k+1} O\Big(\frac{j^2}{n^2}\Big) = O\Big(\frac{k^3}{n^2}\Big) = o(a_{n,k}). \tag{84}$$

From here, we can write $P_\pm = e^{\pm a_{n,k}}(1 + o(a_{n,k}))$. Further taking into account that $a_{n,k} = o(1)$ and thus $e^{\pm a_{n,k}} = 1 + o(1)$, we can write:

$$P_+ - P_- = (e^{a_{n,k}} - e^{-a_{n,k}}) + o(a_{n,k}) = 2a_{n,k} + o(a_{n,k}) = \frac{(k+1)(k+2)}{n}(1 + o(1)). \tag{85}$$

Plugging this result into $Q_{n,k}$, we get:

$$Q_{n,k} = 1 + o(1) \to 1 \tag{86}$$

for large k, n whenever $k = o(\sqrt{n})$. □

Theorem C.2 (Critical regime)**.** *If $k/\sqrt{n} \to c$ for some constant $c > 0$, we have*

$$Q_{n,k} \to \frac{2}{c^2} \sinh\Big(\frac{c^2}{2}\Big). \tag{87}$$

Proof. Analogous to the proof of Theorem C.1, we can write:

$$P_+ - P_- = (e^{a_{n,k}} - e^{-a_{n,k}}) + o(1). \tag{88}$$

When $k/\sqrt{n} \to c$, we then have:

$$a_{n,k} \to \frac{c^2}{2}, \quad P_+ - P_- = e^{c^2/2} - e^{-c^2/2} = 2\sinh\Big(\frac{c^2}{2}\Big). \tag{89}$$

Then for $Q_{n,k}$ we get the following:

$$Q_{n,k} = \frac{n}{(k+1)(k+2)}(P_+ - P_-) \to \frac{2}{c^2} \sinh\Big(\frac{c^2}{2}\Big). \tag{90}$$

□

Theorem C.3 (Supercritical sublinear regime)**.** *If $\sqrt{n} \ll k \ll n$ or, in other words, $k/\sqrt{n} \to \infty$ and $k/n \to 0$, then*

$$Q_{n,k} \sim \frac{n}{(k+1)(k+2)} e^{n\Psi(k/n)}, \tag{91}$$

where $\Psi(x) := (1+x)\log(1+x) - x$. In particular, this regime holds when $k \sim n^\alpha$ for $\alpha \in (1/2, 1)$.

Proof. First, we notice that in this regime P_- is negligible in comparison with P_+. Indeed, when $k \ll n$ we can write:

$$\log \frac{P_-}{P_+} = \sum_{j=0}^{k+1} \log \frac{1 - j/n}{1 + j/n} = -\frac{2}{n} \sum_{j=0}^{k+1} j + O\Big(\frac{1}{n^3} \sum_{j=0}^{k+1} j^3\Big) = -\frac{(k+1)(k+2)}{n} + O\Big(\frac{k^4}{n^3}\Big) \tag{92}$$

Further, we have $O(k^4/n^3) = o(k^2/n)$ when $k = o(n)$, and therefore under $k^2/n \to \infty$ we get:

$$\log \frac{P_-}{P_+} = -\frac{k^2}{n}(1 + o(1)) \to -\infty, \quad P_+ - P_- = P_+(1 + o(1)). \tag{93}$$

It remains to estimate P_+, which we will do via approximating the sum with an integral:

$$\log P_+ = \sum_{j=0}^{k+1} \log\Big(1 + \frac{j}{n}\Big) = n \int_0^{(k+1)/n} \log(1+x)dx + O(\log(1 + (k+1)/n)). \tag{94}$$

Taking into account that $O(\log(1 + (k+1)/n)) = O(k/n) = o(1)$ and computing the integral, we get:

$$\log P_+ = n\Psi\Big(\frac{k+1}{n}\Big) + o(1) = n\Psi\Big(\frac{k}{n}\Big) + o(1). \tag{95}$$

Combining everything, we get:

$$P_+ - P_- = \exp(n\Psi(k/n) + o(1)), \tag{96}$$

which results in the final expression for $Q_{n,k}$:

$$Q_{n,k} = \frac{n}{(k+1)(k+2)} \exp(n\Psi(k/n) + o(1)). \tag{97}$$

□

D Proofs for Scaling Regimes of $S_{n,t}$

This section proves the asymptotic behavior of the signal energy $S_{n,t}$ in the three joint (t, n) scaling regimes identified in the main text: subcritical, critical, and supercritical. The arguments combine the exact finite-width representation of $S_{n,t}$ from Theorem 2.1 with the asymptotic behavior of the corresponding RNN signal energies $Q_{n,k}$ established in Appendix C.

Theorem D.1 (Subcritical regime)**.** *If $t = o(\sqrt{n})$, we have:*

$$S_{n,t} \sim t + 1. \tag{98}$$

Proof. Under $t = o(\sqrt{n})$, we have

$$Q_{n,k} = 1 + o(1) \tag{99}$$

uniformly over $k \le t$ by Theorem C.1. This directly implies the claim:

$$S_{n,t} = \sum_{k=0}^{t} Q_{n,k} = (t+1)(1 + o(1)). \tag{100}$$

□

Theorem D.2 (Critical regime)**.** *If $t/\sqrt{n} \to c$ for some constant $c > 0$, we have*

$$S_{n,t} \sim \sqrt{n} \int_0^c q(x)\, dx, \tag{101}$$

where

$$q(x) := \begin{cases} 1, & x = 0, \\ \dfrac{2}{x^2} \sinh\Big(\dfrac{x^2}{2}\Big), & x > 0. \end{cases} \tag{102}$$

Proof. We first show that, in the critical window, $Q_{n,k}$ admits a uniform approximation by the profile $q(k/\sqrt{n})$ for all $k \le t$. As before, we rewrite $Q_{n,k}$ via $P_\pm$ as follows:

$$P_\pm(n,k) := \prod_{j=0}^{k+1} \Big(1 \pm \frac{j}{n}\Big), \qquad Q_{n,k} = \frac{n}{(k+1)(k+2)} \big(P_+(n,k) - P_-(n,k)\big). \tag{103}$$

Similarly to the proof of Theorem C.2, we define

$$a_{n,k} := \frac{1}{n} \sum_{j=0}^{k+1} j = \frac{(k+1)(k+2)}{2n}, \qquad b_{n,k} := \frac{1}{n^2} \sum_{j=0}^{k+1} j^2 = \frac{(k+1)(k+2)(2k+3)}{6n^2}. \tag{104}$$

Under $k = O(\sqrt{n})$, we have uniformly in $k \le t$:

$$a_{n,k} = O(1), \qquad b_{n,k} = O(n^{-1/2}). \tag{105}$$

Therefore, by the same logarithmic expansion as before,

$$\log P_\pm(n,k) = \pm a_{n,k} + O(b_{n,k}) \tag{106}$$

uniformly over $k \le t$, and exponentiating yields

$$P_+(n,k) - P_-(n,k) = 2 \sinh(a_{n,k}) + O(b_{n,k}). \tag{107}$$

Plugging this into the formula for $Q_{n,k}$, we obtain

$$Q_{n,k} = \frac{2n}{(k+1)(k+2)} \sinh(a_{n,k}) + O\Big(\frac{n\, b_{n,k}}{(k+1)(k+2)}\Big). \tag{108}$$

Since

$$\frac{n\, b_{n,k}}{(k+1)(k+2)} = \frac{2k+3}{6n} = O(n^{-1/2}) \tag{109}$$

uniformly for $k \le t$, it follows that

$$Q_{n,k} = \frac{2n}{(k+1)(k+2)} \sinh(a_{n,k}) + O(n^{-1/2}). \tag{110}$$

Next, define

$$x_{n,k} := \frac{k}{\sqrt{n}}. \tag{111}$$

Then

$$a_{n,k} = \frac{(k+1)(k+2)}{2n} = \frac{k^2}{2n} + \frac{3k+2}{2n} = \frac{x_{n,k}^2}{2} + O(n^{-1/2}) \tag{112}$$

uniformly for $k \le t$. Since the function

$$g(u) := \begin{cases} 1, & u = 0, \\ \dfrac{\sinh(u)}{u}, & u > 0 \end{cases} \tag{113}$$

is continuous on bounded intervals, we obtain

$$\frac{2n}{(k+1)(k+2)} \sinh(a_{n,k}) = g(a_{n,k}) = g\Big(\frac{x_{n,k}^2}{2}\Big) + o(1) \tag{114}$$

uniformly over $k \le t$. In other words,

$$\sup_{0 \le k \le t} \left| Q_{n,k} - q\Big(\frac{k}{\sqrt{n}}\Big) \right| \to 0, \tag{115}$$

where

$$q(x) = g\Big(\frac{x^2}{2}\Big) = \begin{cases} 1, & x = 0, \\ \dfrac{2}{x^2} \sinh\Big(\dfrac{x^2}{2}\Big), & x > 0. \end{cases} \tag{116}$$

We now sum over $k \le t$. Since $t/\sqrt{n} \to c$, we have $t = O(\sqrt{n})$, hence

$$S_{n,t} = \sum_{k=0}^{t} q\Big(\frac{k}{\sqrt{n}}\Big) + o(\sqrt{n}). \tag{117}$$

Dividing by $\sqrt{n}$ gives

$$\frac{S_{n,t}}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_{k=0}^{t} q\Big(\frac{k}{\sqrt{n}}\Big) + o(1). \tag{118}$$

Since q is continuous on $[0, c+\varepsilon]$ for any fixed $\varepsilon > 0$, the sum is a Riemann sum, and therefore

$$\frac{1}{\sqrt{n}} \sum_{k=0}^{t} q\Big(\frac{k}{\sqrt{n}}\Big) \to \int_0^c q(x)\, dx. \tag{119}$$

This proves that

$$\frac{S_{n,t}}{\sqrt{n}} \to \int_0^c q(x)\, dx. \tag{120}$$

□

Theorem D.3 (Supercritical sublinear regime)**.** *If $t/\sqrt{n} \to \infty$ and $t/n \to 0$, then*

$$S_{n,t} \sim \frac{Q_{n,t}}{\log(1 + t/n)}. \tag{121}$$

Equivalently,

$$S_{n,t} \sim \frac{n^2}{t^3} \exp\Big(n\Psi(t/n)\Big), \tag{122}$$

where $\Psi(x) := (1+x)\log(1+x) - x$.

Proof. Recall from Theorem 2.1 that

$$S_{n,t} = \sum_{k=0}^{t} Q_{n,k}. \tag{123}$$

We will show that this sum is asymptotically dominated by an endpoint window near $k = t$. Specifically, we choose an integer sequence L_n satisfying

$$\frac{n}{t} \ll L_n \ll \sqrt{n}. \tag{124}$$

Such a sequence exists because $t/\sqrt{n} \to \infty$. For $\delta_n := \log(1 + t/n) \sim t/n$, this choice implies

$$L_n \delta_n \to \infty, \qquad L_n = o(t). \tag{125}$$

We then split the sum into the endpoint window and its complement as follows:

$$S_{n,t} = \sum_{k=0}^{t-L_n-1} Q_{n,k} + \sum_{k=t-L_n}^{t} Q_{n,k}. \tag{126}$$

We first analyze the endpoint window. For $k = t - j$ with $0 \le j \le L_n$, we have $j = o(t)$, hence $k \sim t$. Therefore, uniformly over $0 \le j \le L_n$,

$$\frac{k}{\sqrt{n}} \to \infty, \qquad \frac{k}{n} \to 0. \tag{127}$$

The proof of Theorem C.3, applied uniformly on this window, gives

$$Q_{n,t-j} = f_n(t-j)(1 + o(1)), \qquad 0 \le j \le L_n, \tag{128}$$

where we defined

$$f_n(k) := \frac{n}{(k+1)(k+2)} \exp\Big(n\Psi(k/n)\Big). \tag{129}$$

We briefly justify this local uniformity. Writing

$$P_{\pm}(n,k) := \prod_{j=0}^{k+1} \left(1 \pm \frac{j}{n}\right), \tag{130}$$

the exact formula for $Q_{n,k}$ gives

$$Q_{n,k} = \frac{n}{(k+1)(k+2)} \big(P_+(n,k) - P_-(n,k)\big). \tag{131}$$

Uniformly for $k = t - j$, $0 \le j \le L_n$, we have

$$\log \frac{P_-(n,k)}{P_+(n,k)} = -\frac{k^2}{n}(1 + o(1)) \to -\infty, \tag{132}$$

because $k \sim t$ and $t^2/n \to \infty$. Thus

$$P_-(n,k) = o(P_+(n,k)) \tag{133}$$

uniformly on the endpoint window. Moreover, by the same Riemann-sum estimate as in Theorem C.3,

$$\log P_+(n,k) = n\Psi(k/n) + o(1) \tag{134}$$

uniformly for $k = t - j$, $0 \le j \le L_n$, since all such k satisfy $k \sim t$ and $t/n \to 0$. This proves (128).

We now evaluate the endpoint contribution. For $0 \le j \le L_n$,

$$\frac{f_n(t-j)}{f_n(t)} = \frac{(t+1)(t+2)}{(t-j+1)(t-j+2)} \exp\Big(-n\big[\Psi(t/n) - \Psi((t-j)/n)\big]\Big). \tag{135}$$

Since $j \le L_n = o(t)$, the prefactor satisfies

$$\frac{(t+1)(t+2)}{(t-j+1)(t-j+2)} = 1 + o(1) \tag{136}$$

uniformly for $0 \le j \le L_n$. Using the Taylor expansion of $\Psi((t-j)/n)$ around t/n, we get

$$\Psi((t-j)/n) = \Psi(t/n) - \frac{j}{n}\Psi'(t/n) + O\left(\frac{j^2}{n^2}\right), \tag{137}$$

where the error term is uniform over $0 \le j \le L_n$, since $\Psi''(x) = 1/(1+x)$ is bounded on $[0, t/n]$ for large n. Therefore,

$$n\big[\Psi(t/n) - \Psi((t-j)/n)\big] = j\Psi'(t/n) + O\left(\frac{j^2}{n}\right). \tag{138}$$

Since $\Psi'(x) = \log(1+x)$ and $\delta_n := \log(1+t/n)$, this gives

$$n\big[\Psi(t/n) - \Psi((t-j)/n)\big] = j\delta_n + O\left(\frac{j^2}{n}\right). \tag{139}$$

By the choice of the endpoint window, $L_n^2/n \to 0$, we then have:

$$n\big[\Psi(t/n) - \Psi((t-j)/n)\big] = j\delta_n + o(1) \tag{140}$$

uniformly over $0 \le j \le L_n$. Therefore,

$$\frac{f_n(t-j)}{f_n(t)} = e^{-j\delta_n}(1+o(1)) \tag{141}$$

uniformly in the endpoint window. Consequently,

$$\sum_{j=0}^{L_n} f_n(t-j) = f_n(t)(1+o(1))\sum_{j=0}^{L_n} e^{-j\delta_n} \sim f_n(t)\sum_{j=0}^{L_n} e^{-j\delta_n}. \tag{142}$$

Because $L_n\delta_n \to \infty$, the truncated geometric sum captures the full endpoint mass:

$$\sum_{j=0}^{L_n} e^{-j\delta_n} \sim \frac{1}{1-e^{-\delta_n}} \sim \frac{1}{\delta_n}. \tag{143}$$

Combining this with (128), we obtain

$$\sum_{k=t-L_n}^{t} Q_{n,k} \sim \frac{f_n(t)}{\delta_n}. \tag{144}$$

It remains to show that the contribution outside the endpoint window is negligible. We use the exact formula and the trivial bound

$$Q_{n,k} \le \frac{n}{(k+1)(k+2)} P_+(n,k). \tag{145}$$

The same integral estimate gives the crude bound

$$Q_{n,k} \le C\frac{n}{(k+1)(k+2)} \exp\Big(n\Psi(k/n)\Big) \tag{146}$$

for all $k \le t$, where C is independent of k and n.

First fix $\eta \in (0,1)$. For $k \le \eta t$, the exponent gap satisfies

$$n\big[\Psi(t/n) - \Psi(k/n)\big] \ge n\big[\Psi(t/n) - \Psi(\eta t/n)\big] = \frac{1-\eta^2}{2}\frac{t^2}{n}(1+o(1)). \tag{147}$$

Since $t^2/n \to \infty$, the contribution of $k \le \eta t$ is exponentially smaller than the endpoint contribution:

$$\sum_{k=0}^{\lfloor \eta t \rfloor} Q_{n,k} = o\left(\frac{f_n(t)}{\delta_n}\right). \tag{148}$$

It remains to control the intermediate range $\eta t < k < t - L_n$. Write $k = t - j$, so now $j \geq L_n$. By (146) and the mean value theorem, uniformly for $\eta t \leq k \leq t$,

$$n\big[\Psi(t/n) - \Psi(k/n)\big] \geq c_\eta (t-k)\delta_n \tag{149}$$

for some constant $c_\eta > 0$, because $\Psi'(x) = \log(1+x)$ and k is comparable to t throughout this range. Hence

$$Q_{n,t-j} \leq C_\eta f_n(t) e^{-c_\eta j \delta_n} \tag{150}$$

for $L_n \leq j \leq (1-\eta)t$. Therefore

$$\sum_{\eta t < k < t - L_n} Q_{n,k} \leq C_\eta f_n(t) \sum_{j \geq L_n} e^{-c_\eta j \delta_n} \leq C_\eta f_n(t) \frac{e^{-c_\eta L_n \delta_n}}{\delta_n}. \tag{151}$$

Since $L_n \delta_n \to \infty$, this gives

$$\sum_{\eta t < k < t - L_n} Q_{n,k} = o\left(\frac{f_n(t)}{\delta_n}\right). \tag{152}$$

Combining (144), (148), and (152), we conclude that

$$S_{n,t} \sim \frac{f_n(t)}{\delta_n}. \tag{153}$$

Finally, by Theorem C.3,

$$Q_{n,t} \sim f_n(t), \tag{154}$$

and therefore

$$S_{n,t} \sim \frac{Q_{n,t}}{\log(1 + t/n)}. \tag{155}$$

Since $\log(1 + t/n) \sim t/n$ and $(t+1)(t+2) \sim t^2$, we also get

$$S_{n,t} \sim \frac{n^2}{t^3} \exp\Big(n\Psi(t/n)\Big). \tag{156}$$

□

Corollary D.4. *If $t/\sqrt{n} \to \infty$ and $t/n^{2/3} \to 0$, then*

$$S_{n,t} \sim \frac{n^2}{t^3} \exp\Big(\frac{t^2}{2n}\Big). \tag{157}$$

Proof. By Theorem D.3,

$$S_{n,t} \sim \frac{n^2}{t^3} \exp\Big(n\Psi(t/n)\Big). \tag{158}$$

Since $t/n \to 0$, we may expand

$$\Psi(x) = \frac{x^2}{2} + O(x^3), \tag{159}$$

and therefore

$$n\Psi(t/n) = \frac{t^2}{2n} + O\Big(\frac{t^3}{n^2}\Big). \tag{160}$$

Under $t = o(n^{2/3})$, we have

$$\frac{t^3}{n^2} = o(1), \tag{161}$$

and hence

$$n\Psi(t/n) = \frac{t^2}{2n} + o(1). \tag{162}$$

Substituting this into the previous expression gives

$$S_{n,t} \sim \frac{n^2}{t^3} \exp\Big(\frac{t^2}{2n}\Big). \tag{163}$$

□

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The introduction includes an accurate summary of the contributions in Section 1.2, while the precise setting studied in the paper is specified in Section 1.1. The abstract provides a further summary of both.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The limitations are discussed explicitly in Section 5.1.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The proofs of all stated results are given in Appendices B (main theorem), C (scaling regimes of the RNN signal energy), and D (scaling regimes of the LRU signal energy). The main text also includes a summary of the proof strategy in Section 2.1.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The numerical results reported in Section 4 consist of straightforward Monte Carlo estimates of the signal energies and are included primarily to validate the theoretical predictions. Thus, experimental results do not form a significant contribution of the paper. For reproducibility, we specify the numerical setup and Monte Carlo sample sizes used in the experiments.

Guidelines:

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

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No]

   Justification: The code used to generate the numerical results in Section 4 consists of straightforward Monte Carlo estimation of the signal energies and is not a substantial contribution of the paper. For this reason, we do not include it with the submission.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [N/A]

   Justification: The experiments do not involve model training and are therefore independent of any hyperparameters related to training or data selection.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: The figures in Section 4 include error bars representing the Monte Carlo standard error of the empirical mean. Since the distributions of the studied observables are heavy-tailed, particularly at large depth, these bars should be interpreted with caution, as they may underestimate the true variability. We make this point explicit in the text.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: The experiments reported in Section 4 are lightweight and require less than 10 hours on a single NVIDIA A100 GPU. We therefore do not discuss computational cost in further detail.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Our work does not raise any particular ethical concerns and adheres to the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [N/A]

Justification: This work is theoretical in nature and does not have any immediate societal impact.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: The work does not release any data or models with a risk on misuse.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [N/A]

Justification: The results of this work do not rely on any licensed datasets, pretrained models, or proprietary code.

Guidelines:

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

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release any datasets, models, or code.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs were used during the preparation of this work for text editing and basic code generation, but they did not affect the core methodology and were not used in any non-standard way.

    Guidelines:

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