ACTIVATION-ADJACENT NORMALIZATION: APPROXIMATION POWER OF RMSNORM AND LAYERNORM

Anonymous authors
Paper under double-blind review

ABSTRACT

Layer normalization (LN) and root mean square normalization (RMSNorm) remain input-dependent and nonlinear at inference, so inserting either next to an activation need not preserve represented functions. We study when activation-normalized (AN) and preactivation-normalized (PN) shallow networks can approximate the function represented by a prescribed width-N one-hidden-layer MLP. Viewing RMSNorm as input-dependent scaling, we use width augmentation to locally linearize normalization on compact feature sets. For AN networks, an unbounded activation range allows width $N + 1$, while every continuous nonconstant activation admits width $N + r$ with error $O_f(r^{-1})$. For PN networks, affine parameters or, without them, degree-one positive homogeneity recover width $N + 1$; general continuous activations admit width $N + r$ with error tending to zero. The corresponding LN constructions use centered coordinate pairs, yielding widths $N + 2$ and $N + 2r$, respectively. We prove that $N + 1$ for RMSNorm and $N + 2$ for LN are also the worst-case bound for approximations, and show that applicable universal-approximation theorems for ordinary MLPs transfer to these normalized architectures. Numerical experiments recover the predicted approximation rates for both normalizations across all five settings and illustrate the shifted-ReLU finite-width obstruction.

1 INTRODUCTION

Classical universal approximation results (Cybenko, 1989; Hornik, 1991; Leshno et al., 1993) provide a basic foundation for neural networks: sufficiently expressive classes can approximate large families of target functions. Modern architectures augment these standard networks with components such as residual connections (He et al., 2016) and normalization (Ioffe & Szegedy, 2015; Ba et al., 2016), often to improve optimization. The representational effects of residual architectures have also been studied (Lu et al., 2017; Lin & Jegelka, 2018). Far less is known about how normalization changes the function class represented by a network.

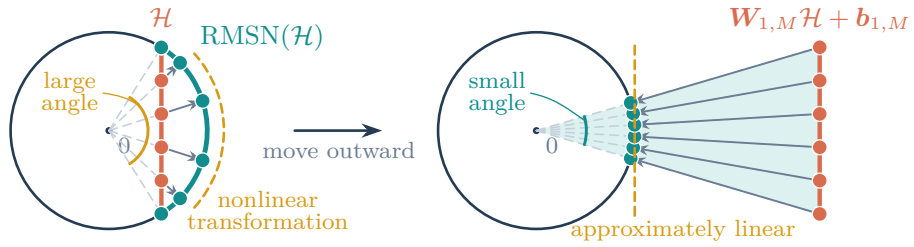
Among normalization mechanisms, this representational question is especially subtle for layer normalization (LN) (Ba et al., 2016) and its centering-free variant, root mean square normalization (RMSNorm) (Zhang & Sennrich, 2019). Both recompute their statistics from each representation and therefore remain nonlinear at inference. Recent theory has primarily treated this nonlinearity as a source of expressivity. Ni et al. (2024) describe exact no-affine LN geometrically as the composition of centering and scaling: the former projects onto the zero-mean hyperplane, whereas the latter projects onto a sphere. Because centering is linear, the nonlinearity lies in the input-dependent scaling (or, equivalently, the RMSNorm operation). Exploiting this operation, Ni et al. (2024) show that width-3 LN networks can fit arbitrary labels on finite samples and derive a depth-dependent VC-dimension lower bound. Building on this view, Ni et al. (2025) arrange multiple LN blocks with linear maps in parallel layer normalization (PLN), treating LN itself as an approximation primitive to prove universal approximation. Whether adding normalization preserves a function already represented by an MLP with its own coordinate-wise activation has received little attention.

To study this preservation problem, we insert a single LN or RMSNorm after or before the activation function of a prescribed one-hidden-layer MLP, obtaining the activation-normalized (AN) architectures in Eqn. 5 and Eqn. 11, and the preactivation-normalized (PN) architectures in Eqn. 8 and Eqn. 12. Directly analyzing whether these architectures preserve the approximation power of

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

the underlying MLP is challenging because LN and RMSNorm couple the hidden neurons (Ni et al., 2024). Therefore, we adopt a relative-approximation perspective and ask when such a normalized architecture can approximate a prescribed target MLP.

For RMSNorm, the only distortion is input-dependent scaling. Inspired by the geometry in Ni et al. (2024), we show that width augmentation moves a bounded feature set away from the origin, thereby shrinking the angle it subtends there and making the scaling factor nearly constant after an input-independent rescaling (Figure 1). Section 4 uses this geometric mechanism to construct AN approximations.



(a) Finite distance from the origin (b) Distance from the origin $\to \infty$

Figure 1: RMSNorm scaling at finite and diverging distances from the origin. In each panel, distances and angles are measured from the origin of the corresponding pre-normalization space. In (a), input-dependent radial scaling maps equally spaced collinear points to nonuniformly spaced points on the sphere. In (b), the lift $\boldsymbol{h} \mapsto [\boldsymbol{h}^\top, M]^\top$ moves the bounded signal set outward as $M \to \infty$, shrinking its angular span. After an input-independent rescaling and a readout that ignores the appended coordinate, the resulting map converges uniformly on $\mathcal{H}$ to a linear map.

For LN, centering introduces a second distortion. We absorb it on the linear side of the activation: by imposing zero-mean preactivations in PN networks and an equivalent zero-sum condition on the readout in AN networks. Width augmentation then makes the remaining input-dependent scaling nearly constant. Section 6 derives the corresponding LN counterparts of the RMSNorm constructions.

Section 7 establishes that the width-$N+1$ RMSNorm and width-$N+2$ LN bounds are worst-case sharp and shows that applicable universal-approximation theorems for ordinary MLPs transfer to the normalized architectures. The experiments in Section 8 recover the predicted approximation rates.

We provide an overview of our theoretical results in Table 1. Here ✓, ×, and ✓/× indicate that affine normalization parameters are present, absent, or arbitrary, respectively. The two AN forms coincide, and affine RMSNorm contains only a coordinate-wise scale. In the lower-bound position, – indicates that no matching lower bound is claimed.

Table 1: Width and error guarantees for a width-N target. Entries are constructive upper bounds/worst-case lower bounds (UB/LB); matching entries are sharp, and – denotes that no matching lower bound is claimed. All rows assume continuous ϕ.

Mode	Parameters	Additional condition on ϕ	RMSNorm (UB/LB)	LN (UB/LB)	Error
AN	✓/×	unbounded range	$N+1\,/\,N+1$	$N+2\,/\,N+2$	$<\varepsilon$
AN	✓/×	nonconstant	$N+r\,/\,-$	$N+2r\,/\,-$	$\le C_f/r$
PN	✓	none	$N+1\,/\,N+1$	$N+2\,/\,N+2$	$<\varepsilon$
PN	×	degree-one homogeneous	$N+1\,/\,N+1$	$N+2\,/\,N+2$	$<\varepsilon$
PN	×	none	$N+r\,/\,-$	$N+2r\,/\,-$	$\to 0$

2 Related Work

Classical universal approximation theory studies when a neural-network class is dense in a target function space. Foundational shallow-network results (Cybenko, 1989; Hornik, 1991; Leshno et al.,

1993) cover sigmoidal activations and, more generally, broad classes of nonpolynomial activations. Work on approximation rates and minimum-width conditions (Barron, 1993; Park et al., 2021) addresses related but distinct questions. Width-limited fully connected ReLU networks exhibit a non-universality threshold (Lu et al., 2017), whereas deep ResNets with one-neuron residual blocks can be universal in L^1 (Lin & Jegelka, 2018). Together, these results show that additional architectural structure can materially change the function class represented by a network.

Batch normalization (Ioffe & Szegedy, 2015) and LN (Ba et al., 2016) were originally introduced to improve neural-network training. With fixed population statistics, inference-time BN is a channel-wise affine transformation and can be folded into a neighboring affine layer. LN is fundamentally different: its mean and variance depend on the current representation, so it remains nonlinear at inference. Ni et al. (2024) decompose LN into centering and scaling. Geometrically, centering projects onto the zero-mean hyperplane, whereas scaling projects onto a sphere. Since centering is linear, their constructions exploit the nonlinear scaling to show that width-3 LN networks can fit arbitrary labels on finite samples and to derive a depth-dependent VC-dimension lower bound. Building on LN as a nonlinear approximation primitive, parallel layer normalization networks (Ni et al., 2025) combine multiple LN blocks with linear maps and establish universal approximation directly for that architecture.

Unlike these studies, which use normalization itself as a source of expressivity, we study a preservation problem. The target MLP already contains a fixed coordinate-wise activation; after inserting a single RMSNorm or LN, we determine how much additional width suffices to recover its specified function. Our constructions use the centering–scaling decomposition in the opposite direction: width augmentation makes scaling nearly input independent, while preactivation or readout reparameterization absorbs LN's centering. The lower bounds show that this extra width can be unavoidable, and the universal approximation results follow by transferring classical density through the relative approximation rather than by constructing targets from LN blocks.

3 Notation and preliminaries

In this paper, ordinary italic letters such as x denote scalars, bold lowercase letters such as $\boldsymbol{x}$ denote vectors, and bold uppercase letters such as $\boldsymbol{W}$ denote matrices. We write $\mathbf{1}_d \in \mathbb{R}^d$ for the all-ones column vector. Unless stated otherwise, network inputs and outputs lie in $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_y}$.

3.1 Target one-hidden-layer MLPs

In this paper, we take a prescribed width-N one-hidden-layer MLP without normalization as the approximation target, namely

$$f(\boldsymbol{x}) = \boldsymbol{W}_2\phi(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2, \tag{1}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ acts coordinate-wise, $\boldsymbol{W}_1 \in \mathbb{R}^{N\times d_x}$, $\boldsymbol{b}_1 \in \mathbb{R}^N$, $\boldsymbol{W}_2 \in \mathbb{R}^{d_y\times N}$, and $\boldsymbol{b}_2 \in \mathbb{R}^{d_y}$. We write $\mathcal{M}_N(\phi)$ for the class of functions represented by all such networks. In this paper, we take $\mathcal{M}_N(\phi)$ to be the target function class. Approximation is measured on a compact set $K \subset \mathbb{R}^{d_x}$ using $\|g\|_{\infty,K} = \sup_{\boldsymbol{x}\in K} \|g(\boldsymbol{x})\|$. For each $f \in \mathcal{M}_N(\phi)$, we fix one width-$N$ realization as in Eqn. 1; constants indexed by f may depend on this realization and on K.

Accordingly, we compare a normalized network directly with a target $f \in \mathcal{M}_N(\phi)$. For $\varepsilon > 0$, a width-d normalized network $\widehat{f}$ is an ε-approximation of f on K if

$$\left\|f - \widehat{f}\right\|_{\infty,K} < \varepsilon. \tag{2}$$

3.2 RMS normalization

We begin with normalization by scale alone. For $\boldsymbol{h} \in \mathbb{R}^d$, RMSN is defined as $\sqrt{d}\,\boldsymbol{h}/\|\boldsymbol{h}\|$ when $\boldsymbol{h} \neq \mathbf{0}$ and as $\mathbf{0}$ when $\boldsymbol{h} = \mathbf{0}$. The map removes dependence on the input magnitude and sends every nonzero input to the sphere $\{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|^2 = d\}$. For $\diamond \in \{0, \mathrm{aff}\}$, let $\mathrm{RMSN}^{(0)}(\boldsymbol{h}) = \mathrm{RMSN}(\boldsymbol{h})$ and $\mathrm{RMSN}^{(\mathrm{aff})}(\boldsymbol{h}) = \boldsymbol{\alpha} \odot \mathrm{RMSN}(\boldsymbol{h})$, where $\boldsymbol{\alpha} \in \mathbb{R}^d$. Thus, the affine mode adds only a learnable coordinate-wise scale, with no additive bias.

With these conventions in place, we now specify the target network class and the relative approximation problem studied throughout the paper.

Theorem 3.1 (Approximation of linear maps by RMSN)**.** *Let $\mathcal{H} \subset \mathbb{R}^N$ be compact and let $\boldsymbol{W}^\star \in \mathbb{R}^{d_y \times N}$. For every $M > 0$, there exists a width-$(N+1)$ map*

$$f_M(\boldsymbol{h}) = \boldsymbol{W}_2 \,\mathrm{RMSN}(\boldsymbol{W}_1 \boldsymbol{h} + \boldsymbol{b}_1) + \boldsymbol{b}_2 \tag{3}$$

such that

$$\sup_{\boldsymbol{h} \in \mathcal{H}} \|f_M(\boldsymbol{h}) - \boldsymbol{W}^\star \boldsymbol{h}\| = O(M^{-2}). \tag{4}$$

The parameters may depend on M. They can be chosen with $\boldsymbol{W}_1$ independent of M, $\|\boldsymbol{b}_1\|_\infty = M$, and $\|\boldsymbol{W}_2\|_{\max} = O(M)$, where $\|\boldsymbol{A}\|_{\max} := \max_{i,j} |A_{ij}|$.
The exact identity and proof are given in Appendix A.1. The construction augments the width by one, lifting each $\boldsymbol{h} \in \mathcal{H}$ to $[\boldsymbol{h}^\top, M]^\top$. Because $\mathcal{H}$ is bounded, the augmented family moves away from the origin and the angle it subtends there tends to zero as $M \to \infty$. Figure 1 illustrates this far-field geometry. After an input-independent rescaling of order M, the resulting RMSN map approaches a linear map uniformly on $\mathcal{H}$. The appended coordinate receives zero readout weight, so the decoder recovers $\boldsymbol{W}^\star \boldsymbol{h}$. Figure 2 summarizes the construction. This lift-linearize-decode mechanism provides the basic template for the constructions that follow, with additional reparameterizations used to accommodate the activation and, in the LN case, centering.

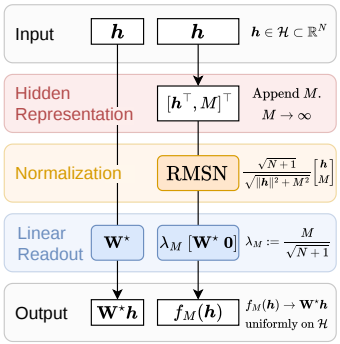


Figure 2: Target map and its one-coordinate RMSNorm realization (Theorem 3.1).

We next study whether inserting RMSNorm immediately before or after the activation of a target one-hidden-layer MLP preserves its represented function.

4 Activation-normalized RMS networks

The width-N activation-normalized RMS class is

$$\mathcal{R}_N^{\mathrm{AN},\diamond}(\phi) := \left\{ \boldsymbol{x} \mapsto \boldsymbol{W}_2 \,\mathrm{RMSN}^{(\diamond)}(\phi(\boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{b}_1)) + \boldsymbol{b}_2 \right\}. \tag{5}$$

Here and below all displayed weights, biases, and available normalization parameters vary freely. The affine scale is absorbed into $\boldsymbol{W}_2$, so the affine and no-affine AN classes coincide; we denote their common value by $\mathcal{R}_N^{\mathrm{AN}}(\phi)$.

The common proof strategy is to augment the width until the original feature map is recovered, up to a uniformly vanishing scaling distortion, from the first N normalized coordinates. The added coordinates control the normalization statistics and are ignored by the target readout, as illustrated in Figure 2.

4.1 Unbounded activations: magnitude controls the normalization scale

Theorem 4.1 (Approximation by RMS-AN networks with unbounded activations)**.** *Let ϕ be continuous with unbounded range. For every $f \in \mathcal{M}_N(\phi)$, there are constants $C_f, R_f > 0$ such that every attainable value $M \in \phi(\mathbb{R})$ with $|M| \geq R_f$ admits*

$$\widehat{f}_M \in \mathcal{R}_{N+1}^{\mathrm{AN}}(\phi), \qquad \left\| \widehat{f}_M - f \right\|_{L^\infty(K)} \leq \frac{C_f}{M^2}. \tag{6}$$

The construction appends the activated value M and assigns its coordinate zero readout weight. After a fixed rescaling, the signal is multiplied by $(1 + \|\boldsymbol{z}(\boldsymbol{x})\|^2 / M^2)^{-1/2}$, which converges uniformly to one. This is exactly the mechanism in Figure 2.

Driving the error in Eqn. 6 to zero requires attainable values $|M| \to \infty$, which is why this one-coordinate construction needs an activation with unbounded range. For bounded activations this regime is unavailable; the next subsection replaces magnitude by multiplicity.

4.2 General activations: multiplicity replaces magnitude

Theorem 4.2 (Approximation by RMS-AN networks with general activations)**.** *Let ϕ be continuous and nonconstant. For every $f \in \mathcal{M}_N(\phi)$, there are constants $C_f > 0$, $r_f \in \mathbb{N}$, and a sequence $\widehat{f}_r \in \mathcal{R}^{\mathrm{AN}}_{N+r}(\phi)$ such that*

$$\left\|\widehat{f}_r - f\right\|_{L^\infty(K)} \le \frac{C_f}{r} \qquad (r \ge r_f). \tag{7}$$

Choose any nonzero attainable value a and append r copies. Their squared norm grows like ra^2, while the target feature vector remains bounded, so the remaining scaling factor is $(1 + \|\boldsymbol{z}(\boldsymbol{x})\|^2/(ra^2))^{-1/2}$. Appendix B proves both theorems.

Summary. RMS-AN networks require only scale stabilization. One large attainable value gives width $N + 1$, while r repeated bounded values give width $N + r$. The next section moves normalization before the activation, where its scale can no longer generally be absorbed into the readout.

5 Preactivation-normalized RMS networks

Unlike in AN networks, the affine RMSNorm scale in PN networks acts before the nonlinear activation and therefore cannot generally be absorbed into the output linear map. The affine and no-affine PN classes can therefore differ, so we treat them separately. The corresponding width-N PN classes are

$$\mathcal{R}^{\mathrm{PN},\diamond}_N(\phi) := \left\{\boldsymbol{x} \mapsto \boldsymbol{W}_2\phi\Big(\mathrm{RMSN}^{(\diamond)}(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\Big) + \boldsymbol{b}_2\right\}. \tag{8}$$

5.1 PN networks with affine parameters

Theorem 5.1 (Approximation by affine PN networks)**.** *Let ϕ be continuous. For every $f \in \mathcal{M}_N(\phi)$ and every $\varepsilon > 0$, there exists $\widehat{f} \in \mathcal{R}^{\mathrm{PN,aff}}_{N+1}(\phi)$ such that $\left\|\widehat{f} - f\right\|_{L^\infty(K)} < \varepsilon$. If ϕ is Lipschitz on compact sets, the approximants have error $O_f(M^{-2})$ as $M \to \infty$.*

The proof follows the construction of Theorem 3.1, with the affine RMSNorm scale performing the required linear rescaling before ϕ. For each target preactivation $\boldsymbol{t}$, append the constant M as one extra coordinate and choose the affine scale so that the first N normalized coordinates are $\left(1 + \|\boldsymbol{t}\|^2/M^2\right)^{-1/2}\boldsymbol{t}$. Because $\boldsymbol{t}$ ranges over a compact set, the scalar factor converges uniformly to one as $|M| \to \infty$. The continuity of ϕ therefore transfers this identity limit through the activation. Appendix C gives the proof.

5.2 A finite-width obstruction without affine parameters

Proposition 5.2 (A finite-width obstruction)**.** *For every finite normalized width d, let $\psi_d(t) = \max\{0, t - \sqrt{d}\}$. Every width-$d$ no-affine RMS-PN network with activation ψ_d is constant, whereas ordinary ψ_d-networks can represent nonconstant functions.*

Every coordinate of an RMSNorm output lies in $[-\sqrt{d}, \sqrt{d}]$, where ψ_d vanishes. Thus no fixed d gives a guarantee uniform over all continuous activations: at finite width, normalization can erase the activation's nonconstant regime and hence its expressive power. The next two subsections give two distinct ways around this obstruction. Positive homogeneity makes an inverse rescaling of the second-layer weights equivalent to applying the required scale before ϕ, retaining width $N + 1$; for a general continuous activation, increasing width recovers approximation asymptotically.

5.3 No-affine PN networks with positively homogeneous activations

Theorem 5.3 (Approximation by no-affine PN networks with homogeneous activations)**.** *Suppose that ϕ is continuous and positively homogeneous of degree one, namely, $\phi(\lambda t) = \lambda\phi(t)$ for all*

$\lambda \geq 0$ *and* $t \in \mathbb{R}$. *For every* $f \in \mathcal{M}_N(\phi)$ *and every* $\varepsilon > 0$, *there exists* $\widehat{f} \in \mathcal{R}^{\mathrm{PN},0}_{N+1}(\phi)$ *such that* $\left\|\widehat{f} - f\right\|_{L^\infty(K)} < \varepsilon$. *The approximants have error* $O_f(M^{-2})$ *as* $M \to \infty$.

Without affine parameters, normalization leaves a positive common factor before ϕ, so no homogeneity condition for negative multipliers is needed. Positive homogeneity moves this factor through the activation, and an inverse rescaling of the second-layer weights cancels it. Equivalently, the output-layer rescaling supplies the scale that affine RMSNorm would have applied before ϕ.

5.4 No-affine PN networks with general continuous activations

Theorem 5.4 (Approximation by no-affine PN networks with general activations)**.** *Let* ϕ *be continuous. For every* $f \in \mathcal{M}_N(\phi)$, *there is a sequence* $\widehat{f}_r \in \mathcal{R}^{\mathrm{PN},0}_{N+r}(\phi)$ *such that* $\left\|\widehat{f}_r - f\right\|_{L^\infty(K)} \to 0$. *If* ϕ *is Lipschitz on compact sets, then for some* $C_f > 0$ *and* $r_f \in \mathbb{N}$,

$$\left\|\widehat{f}_r - f\right\|_{L^\infty(K)} \leq \frac{C_f}{r} \qquad (r \geq r_f). \tag{9}$$

For the target preactivation $\boldsymbol{t}(\boldsymbol{x})$, append r constant coordinates equal to one, forming $[\boldsymbol{t}(\boldsymbol{x})^\top, \mathbf{1}_r^\top]^\top \in \mathbb{R}^{N+r}$. By the definition of RMSNorm, its first N normalized coordinates are $\Big((N+r)/(\|\boldsymbol{t}(\boldsymbol{x})\|^2 + r)\Big)^{1/2} \boldsymbol{t}(\boldsymbol{x})$. On the compact input set, the scalar factor converges uniformly to one, so the normalized preactivation converges to the original one. The networks exist for every r, but the displayed error bound is asserted only for $r \geq r_f$. Appendix C gives the proofs of all results in this section.

Summary. Affine parameters give an $N+1$ RMS-PN approximation for every continuous activation, and positive homogeneity retains width $N+1$ without affine parameters. General continuous activations instead require replicated-coordinate width growth. The next section adds centering and derives the corresponding LayerNorm guarantees.

6 Adding centering: From RMSNorm to LayerNorm

LayerNorm (Ba et al., 2016) differs from RMSNorm by an additional centering operation: before normalization by scale, it subtracts the coordinate mean. Let $\boldsymbol{P}_d = \boldsymbol{I}_d - d^{-1}\mathbf{1}_d\mathbf{1}_d^\top$. With this notation, LN is

$$\mathrm{LN}(\boldsymbol{h}) = \mathrm{RMSN}(\boldsymbol{P}_d\boldsymbol{h}) = \frac{\boldsymbol{h} - \mu(\boldsymbol{h})\mathbf{1}_d}{\sigma(\boldsymbol{h})}, \tag{10}$$

where $\mu(\boldsymbol{h}) = d^{-1}\mathbf{1}_d^\top\boldsymbol{h}$ and $\sigma(\boldsymbol{h}) = (d^{-1}\,\|\boldsymbol{h} - \mu(\boldsymbol{h})\mathbf{1}_d\|^2)^{1/2}$. We set $\mathrm{LN}(\boldsymbol{h}) = \mathbf{0}$ when $\boldsymbol{h}$ has zero variance. Its affine mode is $\mathrm{LN}_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{h}) = \boldsymbol{\alpha} \odot \mathrm{LN}(\boldsymbol{h}) + \boldsymbol{\beta}$ for $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^d$. For $\diamond \in \{0, \mathrm{aff}\}$, let $\mathrm{LN}^{(0)} = \mathrm{LN}$ and $\mathrm{LN}^{(\mathrm{aff})} = \mathrm{LN}_{\boldsymbol{\alpha},\boldsymbol{\beta}}$.

For $\diamond \in \{0, \mathrm{aff}\}$, define the LN classes

$$\mathcal{P}^{\mathrm{AN},\diamond}_N(\phi) := \left\{\boldsymbol{x} \mapsto \boldsymbol{W}_2\,\mathrm{LN}^{(\diamond)}(\phi(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)) + \boldsymbol{b}_2\right\}, \tag{11}$$

$$\mathcal{P}^{\mathrm{PN},\diamond}_N(\phi) := \left\{\boldsymbol{x} \mapsto \boldsymbol{W}_2\phi\Big(\mathrm{LN}^{(\diamond)}(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)\Big) + \boldsymbol{b}_2\right\}, \tag{12}$$

where $\mathrm{LN}^{(0)} = \mathrm{LN}$ and $\mathrm{LN}^{(\mathrm{aff})} = \mathrm{LN}_{\boldsymbol{\alpha},\boldsymbol{\beta}}$. For AN networks the affine parameters are absorbed into the readout, so $\mathcal{P}^{\mathrm{AN},\mathrm{aff}}_N(\phi) = \mathcal{P}^{\mathrm{AN},0}_N(\phi)$; we denote the common class by $\mathcal{P}^{\mathrm{AN}}_N(\phi)$.

The width-augmentation principle is the same as for RMSNorm: the added coordinates make the normalization scale nearly constant on the bounded signal family. The new issue is LN's centering projection $\boldsymbol{P}_d$, which we absorb through reparameterization rather than by approximating the subtracted mean. Theorem 3.1 extends to LN by replacing its one-coordinate augmentation with the following mean-zero two-coordinate construction. Write $s(\boldsymbol{h}) = \mathbf{1}_N^\top\boldsymbol{h}$ and replace the one-coordinate augmentation by $\boldsymbol{y}_M(\boldsymbol{h}) = [\boldsymbol{h}^\top, -s(\boldsymbol{h})/2 + M, -s(\boldsymbol{h})/2 - M]^\top$. Since $\mathbf{1}_{N+2}^\top\boldsymbol{y}_M(\boldsymbol{h}) = 0$, LN equals RMSNorm on the augmented vector; the same decoder gives width $N+2$ and error $O(M^{-2})$. Appendix D.1 gives the exact calculation.

6.1 CENTERING THROUGH REPARAMETERIZATION

The centering operation can be absorbed on the side of the activation where the relevant map remains linear. For PN networks, the mean-zero two-coordinate construction above absorbs centering before the activation and retains the $O(M^{-2})$ rate; see Appendix D.4.

For AN networks, this input-side argument is unavailable because generally $\phi(\boldsymbol{P}_d\boldsymbol{h}) \neq \boldsymbol{P}_d\phi(\boldsymbol{h})$. Moreover, an activation such as ReLU may be unbounded only on one side, so the mean-zero pair $M, -M$ used above need not be attainable. The following result shows how one attainable large value still suffices.

Theorem 6.1 (Approximation by LN-AN networks with unbounded activations)**.** *Let ϕ be continuous with unbounded range. For every $f \in \mathcal{M}_N(\phi)$, there are constants $C_f, R_f > 0$ such that every attainable value $M \in \phi(\mathbb{R})$ with $|M| \geq R_f$ admits*

$$\widehat{f}_M \in \mathcal{P}^{\mathrm{AN}}_{N+2}(\phi), \qquad \left\|\widehat{f}_M - f\right\|_{L^\infty(K)} \leq \frac{C_f}{|M|}. \tag{13}$$

Proof analysis. Let the target be the width-n MLP $g(\boldsymbol{x}) = \boldsymbol{A}\phi(\boldsymbol{B}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{a} = \boldsymbol{A}\boldsymbol{z}(\boldsymbol{x}) + \boldsymbol{a}$. We first append a single attainable value M with zero readout weight. For clarity take $M > 0$; the negative case has the same calculation with the corresponding signs. Put $d = n + 1$, $s = \mathbf{1}_n^\top \boldsymbol{z}$, $\mu_M = (s + M)/d$, $\sigma_M = \sigma([\boldsymbol{z}^\top, M]^\top)$, $\tau_M = M\sqrt{n}/d$, and $\lambda_M = \tau_M/\sigma_M$. The candidate network is

$$\widehat{g}_M(\boldsymbol{x}) = \tau_M \begin{bmatrix}\boldsymbol{A} & \mathbf{0}\end{bmatrix} \mathrm{LN}\left(\begin{bmatrix}\boldsymbol{z}(\boldsymbol{x}) \\ M\end{bmatrix}\right) + \boldsymbol{a} + \frac{M}{d}\boldsymbol{A}\mathbf{1}_n. \tag{14}$$

A direct calculation gives the approximation error

$$\widehat{g}_M - g = (\lambda_M - 1)\boldsymbol{A}\boldsymbol{z} - \left(\lambda_M\mu_M - \frac{M}{d}\right)\boldsymbol{A}\mathbf{1}_n. \tag{15}$$

The last term in Eqn. 14 cancels the leading input-independent centering error. Uniformly on the compact input set, $\lambda_M = 1 + s/(nM) + O(M^{-2})$ and $\lambda_M\mu_M - M/d = s/n + O(M^{-1})$. Thus the remaining error generally tends to $-(s(\boldsymbol{x})/n)\boldsymbol{A}\mathbf{1}_n$. The one-coordinate construction succeeds precisely under the zero-sum condition $\boldsymbol{A}\mathbf{1}_n = \mathbf{0}$.

Lemma 6.1 (Conditional one-coordinate approximation)**.** *Let ϕ be continuous with unbounded range, and let $g(\boldsymbol{x}) = \boldsymbol{A}\phi(\boldsymbol{B}\boldsymbol{x} + \boldsymbol{b}) + \boldsymbol{a}$ be a width-n MLP with $\boldsymbol{A}\mathbf{1}_n = \mathbf{0}$. There are constants $C_g, R_g > 0$ such that every attainable $M \in \phi(\mathbb{R})$ with $|M| \geq R_g$ admits $\widehat{g}_M \in \mathcal{P}^{\mathrm{AN}}_{n+1}(\phi)$ with $\sup_{\boldsymbol{x}\in K} \|\widehat{g}_M(\boldsymbol{x}) - g(\boldsymbol{x})\| \leq C_g/|M|$.*

The complete proof is given in Appendix D.2. Lemma 6.1 thus approximates a zero-sum target after appending only one coordinate. It remains to put an arbitrary target MLP in this favorable form; one additional constant activated coordinate suffices.

Lemma 6.2 (Zero-sum readout reparameterization)**.** *Every $f \in \mathcal{M}_N(\phi)$ has an exactly equivalent width-$(N + 1)$ one-hidden-layer MLP $\bar{f}(\boldsymbol{x}) = \bar{\boldsymbol{W}}_2\phi(\bar{\boldsymbol{W}}_1\boldsymbol{x} + \bar{\boldsymbol{b}}_1) + \bar{\boldsymbol{b}}_2$ with $\bar{\boldsymbol{W}}_2\mathbf{1}_{N+1} = \mathbf{0}$.*

The proof is given in Appendix D.2. Combining Lemmas 6.1 and 6.2, first reparameterizing the target at width $N + 1$ and then appending M, gives width $N + 2$ and proves Theorem 6.1.

Remark 6.1. Viewed in isolation, Lemma 6.2 may look inapplicable: its zero-sum readout acts after scaling, whereas LN centers before scaling. In our construction, the large coordinate M makes the RMSNorm-type scale nearly input independent and hence approximately linear on the lifted compact family. Because the residual scaling is a common scalar, the readout centering passes through it and aligns with LN's inner centering. Thus Lemma 6.2 removes centering exactly, while Lemma 6.1 bounds the residual scaling error.

6.2 REMAINING LAYERNORM APPROXIMATION GUARANTEES

Proposition 6.2 (Remaining LayerNorm counterparts)**.** *The affine-PN and homogeneous-PN conclusions of Theorems 5.1 and 5.3 remain valid for the corresponding LN classes with width $N + 2$. The general-activation conclusions of Theorems 4.2 and 5.4 both remain valid with width $N + 2r$. All stated rates are unchanged. Moreover, the finite-width obstruction in Proposition 5.2 remains valid when RMSNorm is replaced by LN.*

The contrast with RMSNorm is explicit in the widths that grow with r. Its general AN construction repeats one attainable value r times and has width $N+r$. After LN centering, copies of a single value cannot provide a nondegenerate limiting variance, so we repeat each of two distinct attainable values r times, giving width $N + 2r$. The constant part of the induced mean is absorbed into the output bias; the remaining input-dependent centering and scaling errors are both $O(r^{-1})$. Thus, unlike the one-sided unbounded construction, this argument does not require Lemma 6.2. For PN networks, the analogous replacement of r identical coordinates by r symmetric pairs also gives width $N+2r$. The two-coordinate linearization in Appendix D.1 supplies the affine and homogeneous constructions, while replicated symmetric coordinates handle a general continuous activation. Appendix D gives the detailed arguments.

Summary. This section extends the five RMSNorm approximation results to LayerNorm by reparameterizing its centering. The next section establishes their joint worst-case width lower bounds and transfers them to universal approximation.

7 Width lower bounds and universal approximation

7.1 Joint worst-case width lower bounds

The $N + 1$ RMSNorm and $N + 2$ LayerNorm widths are jointly sharp. In each case the identity target on the unit ball exposes the obstruction: when the normalized layer is too narrow, it collapses a nontrivial input segment.

Proposition 7.1 (Worst-case necessity of one additional coordinate)**.** *Let $N \geq 1$ and $K = \{\boldsymbol{x} \in \mathbb{R}^N : \|\boldsymbol{x}\| \leq 1\}$. For every $d \leq N$ and every width-d map*

$$g(\boldsymbol{x}) = \boldsymbol{W}_2 \,\mathrm{RMSN}(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) + \boldsymbol{b}_2 \in \mathbb{R}^N, \tag{16}$$

we have $\sup_{\boldsymbol{x}\in K} \|g(\boldsymbol{x}) - \boldsymbol{x}\| \geq \frac{1}{2}$.

The proof uses rank deficiency when $d < N$ and the ray invariance of RMSNorm when $d = N$. The same argument applies to LN after replacing the affine map by its centered version $\boldsymbol{P}_d(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1)$, where $\boldsymbol{P}_d = \boldsymbol{I}_d - d^{-1}\mathbf{1}_d\mathbf{1}_d^\top$ is the centering projection. Since $\mathrm{rank}(\boldsymbol{P}_d\boldsymbol{W}_1) \leq d - 1$, every width $d \leq N + 1$ LN map likewise has error at least $1/2$, so the first admissible width is $N + 2$.

With $\phi(t) = t$, Proposition 7.1 and its LN counterpart show that the $N + 1$ RMSNorm and $N + 2$ LayerNorm width bounds are worst-case sharp. The $N + r$ and $N + 2r$ widths remain constructive upper bounds. Appendix E.1 gives the shared argument.

7.2 Universal approximation transfer

The preceding relative-approximation guarantees transfer any applicable universal-approximation theorem for one-hidden-layer MLPs to the corresponding normalized classes.

Corollary 7.2 (Universal approximation transfer)**.** *Suppose that $\bigcup_{N\geq 1} \mathcal{M}_N(\phi)$ is dense in $C(K; \mathbb{R}^{d_y})$ under the uniform norm. If ϕ is continuous and nonconstant, then the unions of both RMS-AN and LN-AN classes are dense. If ϕ is continuous, then the corresponding affine and no-affine PN unions are dense as well.*

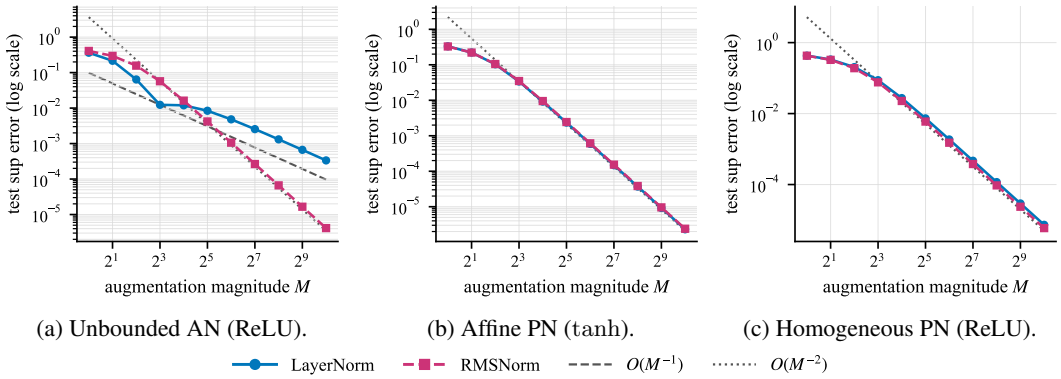
Indeed, approximate the target first by a one-hidden-layer MLP and then approximate that MLP by the corresponding normalized network, the triangle inequality gives the claim. Appendix E.2 gives the details.

8 Numerical illustration

We conduct three experiments to test the fixed-overhead approximation rates by varying the augmentation magnitude M without training. In each experiment, we use six width-16 target MLPs generated from independent random seeds and report the mean uniform error in Figure 3. Figure 3(a) experimentally verifies the $O(M^{-2})$ RMSNorm rate in Theorem 4.1 and the $O(M^{-1})$

LayerNorm rate in Theorem 6.1. Figures 3(b) and 3(c) verify the $O(M^{-2})$ rates in Theorems 5.1 and 5.3, together with their LayerNorm counterparts in Proposition 6.2.

We also test the remaining width-growing constructions, other activations, and target widths $N \in \{4, 8, 16, 32\}$. These experiments exhibit the same predicted behavior. Appendix F gives the complete results, including interquartile ranges and the shifted-ReLU width obstruction.



(a) Unbounded AN (ReLU). (b) Affine PN ($\tanh$). (c) Homogeneous PN (ReLU).

Figure 3: Representative verification of the three M-controlled constructions at $N = 16$. Each curve reports the mean uniform error over six target MLPs generated from independent random seeds. ReLU is used for unbounded AN and homogeneous PN, whereas $\tanh$ is used for affine PN. Both axes use logarithmic scales. Markers distinguish LayerNorm and RMSNorm, and gray lines show the predicted orders. The complete four-width activation sweep and the width-growing experiments appear in Appendix F.

9 Conclusion

Starting from the nonlinearity of input-dependent scaling, we show that width augmentation can make normalization approximately linear on compact feature sets. Under the corresponding assumptions, inserting RMSNorm after or before the activation of a one-hidden-layer MLP can therefore preserve its represented function arbitrarily well by increasing width. LayerNorm admits analogous guarantees after its centering operation is absorbed by reparameterization. Applicable universality results transfer, the $N + 1$ RMSNorm and $N + 2$ LayerNorm bounds are worst-case sharp, and experiments recover the predicted approximation rates.

Limitations. The results concern LN and RMSNorm on compact sets in one shallow normalization–activation composition. Fixed numerical stabilizers, coupled normalization layers, residual architectures, and finite-precision effects lie outside the analysis. Whether the constructive widths $N + r$ and $N + 2r$ are necessary for standard fixed activations remains open. The scalar synthetic experiments do not establish behavior in high-dimensional or deep networks.

AI use statement

We used generative AI tools to assist with manuscript editing, proof checking, literature search and citation checking, code implementation and debugging, and the preparation and interpretation of synthetic experiments and figures. The research question, mathematical claims and constructions, and experimental objectives were formulated by the authors. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content.

References

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

Luca Baroni, Galvin Khara, Joachim Schaeffer, Marat Subkhankulov, and Stefan Heimersheim. Transformers Don't Need LayerNorm at Inference Time: Scaling LayerNorm Removal to GPT-2 XL and the Implications for Mechanistic Interpretability. *arXiv preprint arXiv:2507.02559*, 2025.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Shaked Brody, Uri Alon, and Eran Yahav. On the expressivity role of LayerNorm in transformers' attention. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 14211–14221, 2023.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

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*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Jialong Guo, Xinghao Chen, Yehui Tang, and Yunhe Wang. SLAB: Efficient Transformers with Simplified Linear Attention and Progressive Re-parameterized Batch Normalization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16802–16812, 2024. URL https://proceedings.mlr.press/v235/guo24a.html.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4 (2):251–257, 1991.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Zixuan Jiang, Jiaqi Gu, Hanqing Zhu, and David Z. Pan. Pre-RMSNorm and Pre-CRMSNorm Transformers: Equivalent and Efficient Pre-LN Transformers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 45777–45793, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/8f1bacee31caf990a4f08d84f0ccb322-Abstract-Conference.html.

Andrei Kanavalau, Carmen Amo Alonso, and Sanjay Lall. Gated Normalization Removal and Scale Anchoring in Pre-Norm Transformers, 2026. URL https://arxiv.org/abs/2602.10408v2. arXiv version v2.

Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

Hongzhou Lin and Stefanie Jegelka. Resnet with one-neuron hidden layers is a universal approximator. *Advances in neural information processing systems*, 31, 2018.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. *Advances in neural information processing systems*, 30, 2017.

Yunhao Ni, Yuxin Guo, Junlong Jia, and Lei Huang. On the Nonlinearity of Layer Normalization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 37957–37998, 2024. URL https://proceedings.mlr.press/v235/ni24b.html.

Yunhao Ni, Yuxin Guo, Yuhe Liu, Wenxin Sun, Jie Luo, Wenjun Wu, and Lei Huang. Parallel Layer Normalization for Universal Approximation. *arXiv preprint arXiv:2505.13142*, 2025. URL https://arxiv.org/abs/2505.13142v2.

Sejun Park, Chulhee Yun, Jaeho Lee, and Jinwoo Shin. Minimum Width for Universal Approximation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=O-XJwyoIF-k.

Paul M Riechers. Geometry and dynamics of LayerNorm. *arXiv preprint arXiv:2405.04134*, 2024.

Felix Stollenwerk. On the mathematical relationship between layer normalization and dynamic activation functions. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 674–681, 2026.

Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32, 2019.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Jiachen Zhu, Xinlei Chen, Kaiming He, Yann LeCun, and Zhuang Liu. Transformers without normalization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14901–14911. IEEE, 2025.

A Proofs for Section 3

A.1 Proof of Theorem 3.1

Proof of Theorem 3.1. Put $d = N + 1$ and define

$$\boldsymbol{W}_{1,M} = \begin{bmatrix} \boldsymbol{I}_N \\ \boldsymbol{0}^\top \end{bmatrix}, \qquad \boldsymbol{b}_{1,M} = \begin{bmatrix} \boldsymbol{0}_N \\ M \end{bmatrix}, \qquad \boldsymbol{W}_{2,M} = \frac{|M|}{\sqrt{d}} \begin{bmatrix} \boldsymbol{W}^\star & \boldsymbol{0} \end{bmatrix}, \qquad \boldsymbol{b}_{2,M} = \boldsymbol{0}. \tag{17}$$

Then

$$f_M(\boldsymbol{h}) = \left(1 + \frac{\|\boldsymbol{h}\|^2}{M^2}\right)^{-1/2} \boldsymbol{W}^\star \boldsymbol{h}. \tag{18}$$

Since $\mathcal{H}$ is compact, the elementary inequality $0 \le 1 - (1+u)^{-1/2} \le u/2$ for $u \ge 0$ gives Eqn. 4. □

B Proofs for Section 4

B.1 Proof of Theorem 4.1

Proof of Theorem 4.1. Write $f(\boldsymbol{x}) = \boldsymbol{W}_2 \boldsymbol{z}(\boldsymbol{x}) + \boldsymbol{b}_2$, where $\boldsymbol{z}(\boldsymbol{x}) = \phi(\boldsymbol{W}_1 \boldsymbol{x} + \boldsymbol{b}_1)$. For an attainable value M, append one constant activated coordinate and set $d = N + 1$:

$$\boldsymbol{y}_M(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{z}(\boldsymbol{x}) \\ M \end{bmatrix}, \qquad \widehat{f}_M(\boldsymbol{x}) = \frac{|M|}{\sqrt{d}} \begin{bmatrix} \boldsymbol{W}_2 & \boldsymbol{0} \end{bmatrix} \mathrm{RMSN}(\boldsymbol{y}_M(\boldsymbol{x})) + \boldsymbol{b}_2. \tag{19}$$

Then

$$\widehat{f}_M(\boldsymbol{x}) = \left(1 + \frac{\|\boldsymbol{z}(\boldsymbol{x})\|^2}{M^2}\right)^{-1/2} \boldsymbol{W}_2 \boldsymbol{z}(\boldsymbol{x}) + \boldsymbol{b}_2. \tag{20}$$

Since $\boldsymbol{z}(K)$ is bounded, the inequality used in Appendix A.1 gives Eqn. 6. □

B.2 Proof of Theorem 4.2

Proof of Theorem 4.2. Choose $a \in \phi(\mathbb{R}) \setminus \{0\}$ and append r copies:

$$\boldsymbol{y}_r(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{z}(\boldsymbol{x}) \\ a\boldsymbol{1}_r \end{bmatrix} \in \mathbb{R}^{N+r}. \tag{21}$$

With $d = N + r$, define

$$\widehat{f}_r(\boldsymbol{x}) = \frac{\sqrt{r}|a|}{\sqrt{d}} \begin{bmatrix} \boldsymbol{W}_2 & \boldsymbol{0} \end{bmatrix} \mathrm{RMSN}(\boldsymbol{y}_r(\boldsymbol{x})) + \boldsymbol{b}_2. \tag{22}$$

The signal factor is

$$\left(1 + \frac{\|\boldsymbol{z}(\boldsymbol{x})\|^2}{ra^2}\right)^{-1/2} = 1 + O_f(r^{-1}) \tag{23}$$

uniformly on K, proving the claim. □

C Proofs for Section 5

C.1 Proof of Theorem 5.1

Proof of Theorem 5.1. Write $f(\boldsymbol{x}) = \boldsymbol{W}_2 \phi(\boldsymbol{t}(\boldsymbol{x})) + \boldsymbol{b}_2$ and put $d = N + 1$. Apply RMSNorm to $[\boldsymbol{t}(\boldsymbol{x})^\top, M]^\top$ and choose the affine scale

$$\boldsymbol{\alpha}_M = \frac{|M|}{\sqrt{d}} \begin{bmatrix} \boldsymbol{1}_N \\ 0 \end{bmatrix}. \tag{24}$$

The first N coordinates after affine scaling are $\lambda_M(\boldsymbol{t}(\boldsymbol{x}))\boldsymbol{t}(\boldsymbol{x})$, where $\lambda_M(\boldsymbol{t}) = (1 + \|\boldsymbol{t}\|^2 / M^2)^{-1/2}$, and the added coordinate has zero readout weight. Uniform continuity of ϕ proves convergence; local Lipschitz continuity and $\lambda_M - 1 = O(M^{-2})$ give the stated rate. □

C.2 Proof of Proposition 5.2

Proof of Proposition 5.2. Every RMSNorm output $\boldsymbol{u} \in \mathbb{R}^d$ satisfies $\|\boldsymbol{u}\|^2 = d$ when nonzero, hence $|u_i| \le \sqrt{d}$; the same coordinate bound holds at zero. Thus $\psi_d(u_i) = 0$ for every coordinate, and every no-affine RMS-PN network has only its final bias left. By contrast, $\psi_d(\boldsymbol{w}^\top \boldsymbol{x} + c) = \mathrm{ReLU}(\boldsymbol{w}^\top \boldsymbol{x} + c - \sqrt{d})$ is nonconstant for suitable parameters in a one-hidden-layer MLP. □

C.3 Proof of Theorem 5.3

Proof of Theorem 5.3. With the notation above, the first N no-affine RMSNorm coordinates are $c_M \lambda_M(\boldsymbol{t}(\boldsymbol{x}))\boldsymbol{t}(\boldsymbol{x})$, where $c_M = \sqrt{N+1}/|M|$. Positive homogeneity moves c_M through ϕ, and the readout rescales by c_M^{-1}. The remaining factor $\lambda_M = 1 + O(M^{-2})$ proves the result exactly as above. □

C.4 Proof of Theorem 5.4

Proof of Theorem 5.4. Append r unit coordinates to the preactivation:

$$\boldsymbol{y}_r(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{t}(\boldsymbol{x}) \\ \mathbf{1}_r \end{bmatrix}. \tag{25}$$

The first N RMSNorm coordinates are

$$\boldsymbol{u}_r(\boldsymbol{x}) = \left(\frac{N+r}{\|\boldsymbol{t}(\boldsymbol{x})\|^2 + r} \right)^{1/2} \boldsymbol{t}(\boldsymbol{x}) = \boldsymbol{t}(\boldsymbol{x}) + O_f(r^{-1}) \tag{26}$$

uniformly on K. Define $\widehat{f}_r = [\boldsymbol{W}_2 \ \mathbf{0}]\phi(\mathrm{RMSN}(\boldsymbol{y}_r)) + \boldsymbol{b}_2$. Uniform continuity of ϕ proves convergence, and local Lipschitz continuity gives Eqn. 9 for all sufficiently large r. □

D Proofs for Section 6

D.1 LayerNorm counterpart of Theorem 3.1

Derivation of the two-coordinate LN linearization. Put $d = N + 2$, let $s(\boldsymbol{h}) = \mathbf{1}_N^\top \boldsymbol{h}$, and define

$$\boldsymbol{W}_{1,M} = \begin{bmatrix} \boldsymbol{I}_N \\ -\frac{1}{2}\mathbf{1}_N^\top \\ -\frac{1}{2}\mathbf{1}_N^\top \end{bmatrix}, \qquad \boldsymbol{b}_{1,M} = \begin{bmatrix} \mathbf{0}_N \\ M \\ -M \end{bmatrix}, \qquad \boldsymbol{W}_{2,M} = \frac{\sqrt{2}M}{\sqrt{d}} \begin{bmatrix} \boldsymbol{W}^\star & \mathbf{0} & \mathbf{0} \end{bmatrix}, \qquad \boldsymbol{b}_{2,M} = \mathbf{0}. \tag{27}$$

The augmented vector is already centered:

$$\mathbf{1}_d^\top (\boldsymbol{W}_{1,M}\boldsymbol{h} + \boldsymbol{b}_{1,M}) = 0. \tag{28}$$

Moreover, the signal-dependent part is orthogonal to the direction defined by the two augmentation coordinates, and hence

$$\|\boldsymbol{W}_{1,M}\boldsymbol{h} + \boldsymbol{b}_{1,M}\|^2 = 2M^2 + \|\boldsymbol{h}\|^2 + \frac{1}{2}s(\boldsymbol{h})^2. \tag{29}$$

Using Eqn. 10, we therefore obtain the exact identity

$$f_M(\boldsymbol{h}) = \left(1 + \frac{\|\boldsymbol{h}\|^2 + \frac{1}{2}s(\boldsymbol{h})^2}{2M^2} \right)^{-1/2} \boldsymbol{W}^\star \boldsymbol{h}. \tag{30}$$

Since $\mathcal{H}$ is compact, both $\|\boldsymbol{h}\|$ and $\|\boldsymbol{h}\|^2 + \frac{1}{2}s(\boldsymbol{h})^2$ are uniformly bounded on $\mathcal{H}$. The elementary inequality

$$0 \le 1 - (1+u)^{-1/2} \le \frac{u}{2}, \qquad u \ge 0, \tag{31}$$

then gives the uniform $O(M^{-2})$ error asserted in Section 6.1. □

D.2 Proof of Theorem 6.1

AN affine equivalence for LayerNorm. For an affine AN network,

$$\boldsymbol{W}_2 \operatorname{LN}_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\boldsymbol{y}) + \boldsymbol{b}_2 = \boldsymbol{W}_2 \operatorname{diag}(\boldsymbol{\alpha}) \operatorname{LN}(\boldsymbol{y}) + (\boldsymbol{W}_2\boldsymbol{\beta} + \boldsymbol{b}_2). \tag{32}$$

Absorbing $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ into the linear readout and output bias gives $\mathcal{P}_N^{\mathrm{AN,aff}}(\phi) \subseteq \mathcal{P}_N^{\mathrm{AN},0}(\phi)$. Conversely, every no-affine AN network is obtained from the affine class by choosing $\boldsymbol{\alpha} = \mathbf{1}_N$ and $\boldsymbol{\beta} = \mathbf{0}$. This proves the reverse inclusion and hence the equality. □

Proof of Lemma 6.2. Write $f(\boldsymbol{x}) = \boldsymbol{W}_2\boldsymbol{z}(\boldsymbol{x}) + \boldsymbol{b}_2$, where $\boldsymbol{z}(\boldsymbol{x}) = \phi(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) \in \mathbb{R}^N$. Choose any $c \in \mathbb{R}$ and append the constant activated feature $\phi(c)$:

$$\bar{\boldsymbol{W}}_1 = \begin{bmatrix} \boldsymbol{W}_1 \\ \mathbf{0}^\top \end{bmatrix}, \qquad \bar{\boldsymbol{b}}_1 = \begin{bmatrix} \boldsymbol{b}_1 \\ c \end{bmatrix}, \qquad \bar{\boldsymbol{z}}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{z}(\boldsymbol{x}) \\ \phi(c) \end{bmatrix}. \tag{33}$$

Set

$$\bar{\boldsymbol{W}}_2 = [\boldsymbol{W}_2 \quad -\boldsymbol{W}_2\mathbf{1}_N], \qquad \bar{\boldsymbol{b}}_2 = \boldsymbol{b}_2 + (\boldsymbol{W}_2\mathbf{1}_N)\phi(c). \tag{34}$$

Then

$$\bar{\boldsymbol{W}}_2\bar{\boldsymbol{z}}(\boldsymbol{x}) + \bar{\boldsymbol{b}}_2 = \boldsymbol{W}_2\boldsymbol{z}(\boldsymbol{x}) - (\boldsymbol{W}_2\mathbf{1}_N)\phi(c) + \boldsymbol{b}_2 + (\boldsymbol{W}_2\mathbf{1}_N)\phi(c) = f(\boldsymbol{x}), \tag{35}$$

while

$$\bar{\boldsymbol{W}}_2\mathbf{1}_{N+1} = \boldsymbol{W}_2\mathbf{1}_N - \boldsymbol{W}_2\mathbf{1}_N = \mathbf{0}. \tag{36}$$

Thus both the representation and the feature-centering condition hold exactly. □

Proof of Lemma 6.1. Compactness of K and continuity of ϕ imply that $\boldsymbol{z}(K)$ is bounded. For an attainable $M \in \phi(\mathbb{R})$, choose $c_M \in \mathbb{R}$ with $\phi(c_M) = M$ and append one constant activated coordinate:

$$\boldsymbol{y}_M(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{z}(\boldsymbol{x}) \\ M \end{bmatrix} \in \mathbb{R}^d, \qquad d = n + 1. \tag{37}$$

This is a valid AN feature vector: append a zero row to $\boldsymbol{B}$ and the bias c_M to $\boldsymbol{b}$. Let $s(\boldsymbol{x}) = \mathbf{1}_n^\top \boldsymbol{z}(\boldsymbol{x})$, $v(\boldsymbol{x}) = \|\boldsymbol{z}(\boldsymbol{x})\|^2$, and $\sigma_M(\boldsymbol{x}) = \sigma(\boldsymbol{y}_M(\boldsymbol{x}))$. With

$$\tau_M = \frac{|M|\sqrt{n}}{d}, \tag{38}$$

a direct calculation gives

$$\sigma_M(\boldsymbol{x})^2 = \frac{v(\boldsymbol{x}) + M^2}{d} - \frac{(s(\boldsymbol{x}) + M)^2}{d^2}, \tag{39}$$

$$\frac{\sigma_M(\boldsymbol{x})^2}{\tau_M^2} = 1 - \frac{2s(\boldsymbol{x})}{nM} + \frac{dv(\boldsymbol{x}) - s(\boldsymbol{x})^2}{nM^2}. \tag{40}$$

The functions s and v are uniformly bounded on K. Hence there are constants $C_0, R_0 > 0$ such that every attainable $|M| \geq R_0$ satisfies

$$\sup_{\boldsymbol{x}\in K} \left| \frac{\sigma_M(\boldsymbol{x})^2}{\tau_M^2} - 1 \right| \leq \frac{C_0}{|M|}. \tag{41}$$

Increasing R_0 if necessary makes the ratio lie in $[1/2, 3/2]$. Since $u \mapsto u^{-1/2}$ is Lipschitz on this interval, there is $C_1 > 0$ such that

$$\sup_{\boldsymbol{x}\in K} \left| \frac{\tau_M}{\sigma_M(\boldsymbol{x})} - 1 \right| \leq \frac{C_1}{|M|}. \tag{42}$$

Define the width-d AN network

$$\widehat{g}_M(\boldsymbol{x}) = \tau_M \, [\boldsymbol{A} \quad \mathbf{0}] \operatorname{LN}(\boldsymbol{y}_M(\boldsymbol{x})) + \boldsymbol{a}. \tag{43}$$

Because $[\boldsymbol{A}\ \mathbf{0}]\mathbf{1}_d = \boldsymbol{A}\mathbf{1}_n = \mathbf{0}$, the readout cancels the mean subtracted by LN; the augmentation coordinate itself has zero direct readout weight. Therefore

$$\widehat{g}_M(\boldsymbol{x}) = \frac{\tau_M}{\sigma_M(\boldsymbol{x})}\boldsymbol{A}\boldsymbol{z}(\boldsymbol{x}) + \boldsymbol{a}. \tag{44}$$

Combining Eqn. 42 with boundedness of $\boldsymbol{A}\boldsymbol{z}(K)$ yields

$$\sup_{\boldsymbol{x}\in K} \|\widehat{g}_M(\boldsymbol{x}) - g(\boldsymbol{x})\| \le \sup_{\boldsymbol{x}\in K} \left|\frac{\tau_M}{\sigma_M(\boldsymbol{x})} - 1\right| \sup_{\boldsymbol{x}\in K} \|\boldsymbol{A}\boldsymbol{z}(\boldsymbol{x})\| \le \frac{C_g}{|M|}. \tag{45}$$

The constructed width is $d = n + 1$. □

Proof of Theorem 6.1. Lemma 6.2 represents the target f exactly as a width- $(N + 1)$ MLP with a zero-sum readout. Apply Lemma 6.1 with $n = N + 1$. The resulting LN-AN network has width $n + 1 = N + 2$ and satisfies Eqn. 13. Since the unbounded range of ϕ contains attainable values of arbitrarily large magnitude, the error can be made smaller than any prescribed $\varepsilon > 0$. □

D.3 LayerNorm counterpart of Theorem 4.2

General AN case of Proposition 6.2. Write the target as

$$f(\boldsymbol{x}) = \boldsymbol{W}_2\boldsymbol{z}(\boldsymbol{x}) + \boldsymbol{b}_2, \qquad \boldsymbol{z}(\boldsymbol{x}) = \phi(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) \in \mathbb{R}^N. \tag{46}$$

Choose $\xi_a, \xi_b \in \mathbb{R}$ such that $a = \phi(\xi_a) \neq b = \phi(\xi_b)$. For $r \ge 1$, append r copies of each value:

$$\boldsymbol{y}_r(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{z}(\boldsymbol{x}) \\ a\mathbf{1}_r \\ b\mathbf{1}_r \end{bmatrix} \in \mathbb{R}^D, \qquad D = N + 2r. \tag{47}$$

These constant activated coordinates are realized by constant preactivations ξ_a and ξ_b. Give all of them zero readout weight and put

$$\widetilde{\boldsymbol{W}}_{2,r} = [\boldsymbol{W}_2 \quad \mathbf{0}]. \tag{48}$$

Then

$$\widetilde{\boldsymbol{W}}_{2,r}\boldsymbol{y}_r(\boldsymbol{x}) = \boldsymbol{W}_2\boldsymbol{z}(\boldsymbol{x}), \qquad \widetilde{\boldsymbol{W}}_{2,r}\mathbf{1}_D = \boldsymbol{W}_2\mathbf{1}_N. \tag{49}$$

Write

$$s(\boldsymbol{x}) = \mathbf{1}_N^\top \boldsymbol{z}(\boldsymbol{x}), \quad v(\boldsymbol{x}) = \|\boldsymbol{z}(\boldsymbol{x})\|^2, \quad A_r = r(a^2 + b^2), \quad B_r = r(a + b), \tag{50}$$

and define the augmentation-only baseline mean and variance by

$$m_r = \frac{B_r}{D}, \qquad \tau_r^2 = \frac{A_r}{D} - \frac{B_r^2}{D^2}. \tag{51}$$

The full mean is $\mu_r(\boldsymbol{x}) = m_r + s(\boldsymbol{x})/D$, and the full variance is

$$\sigma_r(\boldsymbol{x})^2 = \frac{v(\boldsymbol{x}) + A_r}{D} - \frac{(s(\boldsymbol{x}) + B_r)^2}{D^2} \tag{52}$$

$$= \tau_r^2 + \Delta_r(\boldsymbol{x}), \tag{53}$$

where

$$\Delta_r(\boldsymbol{x}) = \frac{v(\boldsymbol{x})}{D} - \frac{2B_r s(\boldsymbol{x}) + s(\boldsymbol{x})^2}{D^2}. \tag{54}$$

Because $\boldsymbol{z}(K)$ is bounded, $\sup_{\boldsymbol{x}\in K} |\Delta_r(\boldsymbol{x})| = O(r^{-1})$, while

$$\tau_r^2 \longrightarrow \frac{(a-b)^2}{4} > 0. \tag{55}$$

Hence, uniformly on K,

$$\frac{\tau_r}{\sigma_r(\boldsymbol{x})} = 1 + O(r^{-1}). \tag{56}$$

Finally, define

$$\widehat{f}_r(\boldsymbol{x}) = \tau_r\widetilde{\boldsymbol{W}}_{2,r}\,\mathrm{LN}(\boldsymbol{y}_r(\boldsymbol{x})) + \boldsymbol{b}_2 + m_r\boldsymbol{W}_2\mathbf{1}_N. \tag{57}$$

Put $\rho_r(\boldsymbol{x}) = \tau_r/\sigma_r(\boldsymbol{x})$. Using Eqn. 49 and $\mu_r = m_r + s/D$, the approximation error is

$$\widehat{f}_r(\boldsymbol{x}) - f(\boldsymbol{x}) = \big(\rho_r(\boldsymbol{x}) - 1\big)\boldsymbol{W}_2\boldsymbol{z}(\boldsymbol{x}) + \left[\big(1 - \rho_r(\boldsymbol{x})\big)m_r - \rho_r(\boldsymbol{x})\frac{s(\boldsymbol{x})}{D}\right]\boldsymbol{W}_2\mathbf{1}_N. \tag{58}$$

Now $m_r = O(1)$, $D^{-1} = O(r^{-1})$, and s is uniformly bounded on K. Together with Eqn. 56, these facts give $\left\|\widehat{f}_r - f\right\|_{L^\infty(K)} = O_f(r^{-1})$, completing the proof with width $D = N + 2r$. □

D.4 LayerNorm counterpart of Theorem 5.1

Affine-PN case of Proposition 6.2. Write $f(\boldsymbol{x}) = \boldsymbol{W}_2\phi(\boldsymbol{t}(\boldsymbol{x})) + \boldsymbol{b}_2$, where $\boldsymbol{t}(\boldsymbol{x}) = \boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1 \in \mathbb{R}^N$, and apply the first layer parameters $\boldsymbol{W}_{1,M}, \boldsymbol{b}_{1,M}$ from Eqn. 27 to $\boldsymbol{t}(\boldsymbol{x})$. Put $d = N + 2$, $c_M = \sqrt{2}M/\sqrt{d}$, and choose affine LN parameters

$$\boldsymbol{\alpha}_M = c_M \begin{bmatrix} \mathbf{1}_N \\ 0 \\ 0 \end{bmatrix}, \qquad \boldsymbol{\beta}_M = \mathbf{0}. \tag{59}$$

By Eqn. 30 with $\boldsymbol{W}^\star = \boldsymbol{I}_N$, the first N coordinates after affine LN are $\lambda_M(\boldsymbol{t}(\boldsymbol{x}))\boldsymbol{t}(\boldsymbol{x})$, where $\lambda_M(\boldsymbol{t}) \to 1$ uniformly on $\boldsymbol{t}(K)$ with error $O(M^{-2})$. The last two activated coordinates are assigned zero readout weight, so

$$\widehat{f}_M(\boldsymbol{x}) = [\boldsymbol{W}_2 \quad \mathbf{0} \quad \mathbf{0}]\, \phi(\mathrm{LN}_{\boldsymbol{\alpha}_M,\mathbf{0}}(\boldsymbol{W}_{1,M}\boldsymbol{t}(\boldsymbol{x}) + \boldsymbol{b}_{1,M})) + \boldsymbol{b}_2 \longrightarrow f(\boldsymbol{x}) \tag{60}$$

uniformly by continuity of ϕ. Local Lipschitz continuity gives the stated rate. □

D.5 LayerNorm counterpart of Proposition 5.2

Finite-width obstruction case of Proposition 6.2. Every nondegenerate LN output $\boldsymbol{u} \in \mathbb{R}^d$ satisfies $\|\boldsymbol{u}\|^2 = d$, so $u_i \le \sqrt{d}$ for each coordinate. Thus $\psi_d(u_i) = 0$ for all i. At a degenerate input, our definition $\mathrm{LN}(\boldsymbol{u}) = \mathbf{0}$ gives the same conclusion. Hence the network output is its final bias.

On the other hand,

$$\psi_d(\boldsymbol{w}^\top\boldsymbol{x} + c) = \mathrm{ReLU}(\boldsymbol{w}^\top\boldsymbol{x} + c - \sqrt{d}), \tag{61}$$

so the threshold is absorbed into the first-layer bias of an ordinary ψ_d-network, which can therefore represent nonconstant functions. □

D.6 LayerNorm counterpart of Theorem 5.3

Homogeneous PN case of Proposition 6.2. Write $f(\boldsymbol{x}) = \boldsymbol{W}_2\phi(\boldsymbol{t}(\boldsymbol{x})) + \boldsymbol{b}_2$, where $\boldsymbol{t}(\boldsymbol{x}) = \boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1$, and use the same augmented preactivation $\boldsymbol{W}_{1,M}\boldsymbol{t}(\boldsymbol{x}) + \boldsymbol{b}_{1,M}$ as in Eqn. 27. Put $d = N + 2$ and $c_M = \sqrt{2}M/\sqrt{d}$. If $\lambda_M(\boldsymbol{t})$ denotes the scalar in Eqn. 30 for $\boldsymbol{W}^\star = \boldsymbol{I}_N$, then the first N coordinates of $\mathrm{LN}(\boldsymbol{W}_{1,M}\boldsymbol{t} + \boldsymbol{b}_{1,M})$ equal $c_M^{-1}\lambda_M(\boldsymbol{t})\boldsymbol{t}$. Positive homogeneity therefore gives

$$c_M\,[\boldsymbol{W}_2 \quad \mathbf{0} \quad \mathbf{0}]\, \phi(\mathrm{LN}(\boldsymbol{W}_{1,M}\boldsymbol{t}(\boldsymbol{x}) + \boldsymbol{b}_{1,M})) + \boldsymbol{b}_2 = \lambda_M(\boldsymbol{t}(\boldsymbol{x}))\boldsymbol{W}_2\phi(\boldsymbol{t}(\boldsymbol{x})) + \boldsymbol{b}_2. \tag{62}$$

The factor λ_M converges uniformly to one at rate $O(M^{-2})$ on the compact set $\boldsymbol{t}(K)$, proving the claim. □

D.7 LayerNorm counterpart of Theorem 5.4

Lemma D.1 (Replicated-coordinate identity limit)**.** *Let $\boldsymbol{t}(\boldsymbol{x}) = \boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1 \in \mathbb{R}^N$. For $r \ge 1$, define*

$$\boldsymbol{y}_r(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{t}(\boldsymbol{x}) \\ -\mathbf{1}_r \\ \mathbf{1}_r \end{bmatrix} \in \mathbb{R}^{N+2r}. \tag{63}$$

Let $\boldsymbol{u}_r(\boldsymbol{x})$ be the first N coordinates of $\mathrm{LN}(\boldsymbol{y}_r(\boldsymbol{x}))$*. Then*

$$\sup_{\boldsymbol{x}\in K} \|\boldsymbol{u}_r(\boldsymbol{x}) - \boldsymbol{t}(\boldsymbol{x})\| = O(r^{-1}). \tag{64}$$

Proof. Write $s(\boldsymbol{x}) = \mathbf{1}_N^\top\boldsymbol{t}(\boldsymbol{x})$ and $D = N + 2r$. The replicated values -1 and 1 cancel in the sum, so

$$\mu_r(\boldsymbol{x}) = \frac{s(\boldsymbol{x})}{D}. \tag{65}$$

They contribute $2r$ to the squared norm, and hence

$$\rho_r(\boldsymbol{x})^2 = \frac{\|\boldsymbol{t}(\boldsymbol{x})\|^2 + 2r}{D} - \frac{s(\boldsymbol{x})^2}{D^2} \tag{66}$$

$$= 1 + \frac{\|\boldsymbol{t}(\boldsymbol{x})\|^2 - N}{D} - \frac{s(\boldsymbol{x})^2}{D^2}. \tag{67}$$

Compactness of $\boldsymbol{t}(K)$ gives $\sup_{\boldsymbol{x}\in K}|\rho_r(\boldsymbol{x})^2-1| = O(r^{-1})$. Thus, for all sufficiently large r, ρ_r is uniformly bounded away from zero and $\rho_r^{-1}-1 = O(r^{-1})$ uniformly on K.

The first N coordinates of $\mathrm{LN}(\boldsymbol{y}_r(\boldsymbol{x}))$ are

$$\boldsymbol{u}_r(\boldsymbol{x}) = \frac{\boldsymbol{t}(\boldsymbol{x}) - D^{-1}s(\boldsymbol{x})\mathbf{1}_N}{\rho_r(\boldsymbol{x})}. \tag{68}$$

Consequently,

$$\boldsymbol{u}_r(\boldsymbol{x}) - \boldsymbol{t}(\boldsymbol{x}) = \big(\rho_r(\boldsymbol{x})^{-1}-1\big)\boldsymbol{t}(\boldsymbol{x}) - \frac{s(\boldsymbol{x})}{D\rho_r(\boldsymbol{x})}\mathbf{1}_N. \tag{69}$$

Both terms are uniformly $O(r^{-1})$, which proves Eqn. 64. □

General PN case of Proposition 6.2. Write $f(\boldsymbol{x}) = \boldsymbol{W}_2\phi(\boldsymbol{t}(\boldsymbol{x})) + \boldsymbol{b}_2$ and use $\boldsymbol{y}_r(\boldsymbol{x})$ from Lemma D.1. Define

$$\widehat{f}_r(\boldsymbol{x}) = [\boldsymbol{W}_2 \quad \mathbf{0}]\,\phi(\mathrm{LN}(\boldsymbol{y}_r(\boldsymbol{x}))) + \boldsymbol{b}_2, \tag{70}$$

where the zero block ignores the $2r$ augmentation coordinates after activation. By Lemma D.1, the first N normalized coordinates converge uniformly to $\boldsymbol{t}(\boldsymbol{x})$. They and $\boldsymbol{t}(K)$ lie in a common compact set for all sufficiently large r, so uniform continuity of ϕ proves uniform convergence to f. If ϕ is Lipschitz on compact sets, Eqn. 64 yields constants $C_f > 0$ and r_f for which the $O_f(r^{-1})$ bound holds whenever $r \geq r_f$. □

E Proofs for Section 7

E.1 Proof of Proposition 7.1

Proof of Proposition 7.1. If $\operatorname{rank}\boldsymbol{W}_1 < N$, choose a unit $\boldsymbol{v} \in \ker \boldsymbol{W}_1$. The inputs $\boldsymbol{v}$ and $-\boldsymbol{v}$ have the same normalized representation, so the error is at least 1. Otherwise $d = N$ and $\boldsymbol{W}_1$ is invertible. Let $\boldsymbol{x}_0$ be the unique solution of $\boldsymbol{W}_1\boldsymbol{x}_0+\boldsymbol{b}_1 = \mathbf{0}$ and write $\boldsymbol{x}_0 = t\boldsymbol{v}$ with $t \geq 0$ and $\|\boldsymbol{v}\| = 1$ (choose any unit $\boldsymbol{v}$ if $\boldsymbol{x}_0 = \mathbf{0}$). For $-1 \leq s < \min\{t, 1\}$, the representations of $s\boldsymbol{v}$ lie on the same nonzero ray, so their RMSNorm outputs coincide. This segment has diameter at least one, giving error at least $1/2$; an endpoint of zero variance is handled by taking a limit.

LayerNorm extension. Set $\boldsymbol{C} = \boldsymbol{P}_d\boldsymbol{W}_1$ and $\boldsymbol{c} = \boldsymbol{P}_d\boldsymbol{b}_1$. Any two inputs whose centered affine images lie on the same nonzero ray have the same LN representation and hence the same network output. Since $\operatorname{rank}\boldsymbol{C} \leq d-1$, if $\operatorname{rank}\boldsymbol{C} < N$ we may choose a unit $\boldsymbol{v} \in \ker \boldsymbol{C}$; the inputs $\boldsymbol{v}$ and $-\boldsymbol{v}$ then have the same centered affine image $\boldsymbol{c}$, so the approximation error is at least 1. Otherwise $\operatorname{rank}\boldsymbol{C} = N$, which forces $d = N+1$, and $\boldsymbol{C}$ is a bijection from $\mathbb{R}^N$ onto $\mathbf{1}_d^{\perp}$. Because $\boldsymbol{c} \in \mathbf{1}_d^{\perp}$, there is a unique $\boldsymbol{x}_0$ satisfying $\boldsymbol{C}\boldsymbol{x}_0 + \boldsymbol{c} = \mathbf{0}$. Write $\boldsymbol{x}_0 = t\boldsymbol{v}$ with $t \geq 0$ and $\|\boldsymbol{v}\| = 1$, choosing any unit $\boldsymbol{v}$ when $\boldsymbol{x}_0 = \mathbf{0}$. For $-1 \leq s < \min\{t, 1\}$,

$$\boldsymbol{C}(s\boldsymbol{v}) + \boldsymbol{c} = (s-t)\boldsymbol{C}\boldsymbol{v} \tag{71}$$

lies on one nonzero ray. The corresponding inputs form a segment of diameter $1 + \min\{t, 1\} \geq 1$, so their common network output has uniform error at least $1/2$. If the limiting endpoint has zero variance, taking points approaching it gives the same bound. Thus the LN obstruction holds for every $d \leq N+1$. □

E.2 Proof of Corollary 7.2

Proof of Corollary 7.2. Given a target $f_\star \in C(K;\mathbb{R}^{d_y})$ and $\varepsilon > 0$, density of the one-hidden-layer MLPs gives some $f \in \mathcal{M}_N(\phi)$ with $\|f_\star - f\|_{\infty,K} < \varepsilon/2$. Apply Theorem 4.2, 5.1, or 5.4 for the three RMSNorm claims, and then Proposition 6.2 for their LayerNorm counterparts. In each case choose the approximation error below $\varepsilon/2$; the triangle inequality gives $\left\|f_\star - \widehat{f}\right\|_{\infty,K} < \varepsilon$. □

F EXPERIMENTAL DETAILS

All experiments use $d_x = d_y = 1$ and target networks with parameters

$$\boldsymbol{W}_1 = 1.15\boldsymbol{G}_1, \qquad \boldsymbol{b}_1 = 0.45\boldsymbol{G}_2, \qquad \boldsymbol{W}_2 = \frac{0.7\boldsymbol{G}_3 + 0.25\mathbf{1}_N^\top}{\sqrt{N}}, \qquad \boldsymbol{b}_2 = 0.15\boldsymbol{G}_4, \tag{72}$$

where all entries of $\boldsymbol{G}_1, \boldsymbol{G}_2, \boldsymbol{G}_3, \boldsymbol{G}_4$ are standard normal draws. Figure 3 fixes $N = 16$ and reports the mean over six independently generated targets. The full five-regime sweep in Figure 4 uses $N \in \{4, 8, 16, 32\}$, six targets at each width, and $M, r \in \{2^0, \ldots, 2^{10}\}$. We evaluate the explicit constructions for exact LN and exact RMSNorm in float64 arithmetic. The reported metric is the maximum error on an 8193-point test grid. The full sweep contains 1320 evaluations per normalization, with no nonfinite metrics. The shifted-ReLU experiment in Figures 5 and 6 sets $N = 1$, $K = [-1, 1]$, $\phi_a(t) = \mathrm{ReLU}(t - a)$ for $a \in \{10, 30, 100\}$, and $f(x) = \phi_a(a + 2x) = \mathrm{ReLU}(2x)$. For each threshold, the coordinate bound $|u_i| \le \sqrt{d}$ forces every hidden activation to vanish when $d \le a^2$, so every no-affine PN network is constant and its uniform error is at least 1. The plotted RMSNorm and LayerNorm constructions use replicated unit coordinates and converge as d grows. Figure 4 uses the same targets, grid, and parameter values, changing only the activation. The shifted-ReLU experiments use the same grid and evaluate the corresponding explicit PN constructions directly, again without training.

Table 2 summarizes the rate test numerically. For each construction, we fit the slope over the final six log–log points and then report its median and interquartile range over the 24 target–width pairs. Thus a fitted slope near $-q$ supports an error of order q in the corresponding control parameter.

Table 2: Predicted and observed tail slopes of the five explicit constructions. Observed values are medians (interquartile ranges) over 24 target–width pairs; agreement with the predicted slope is the quantitative rate check used below.

Construction	Predicted RMS/LN	RMS observed	LN observed
AN, unbounded (ReLU)	$-2/-1$	-1.998 $(-1.999, -1.997)$	-0.953 $(-0.977, -0.886)$
AN, general ($\tanh$)	$-1/-1$	-0.956 $(-0.975, -0.931)$	-0.968 $(-0.982, -0.940)$
PN, affine ($\tanh$)	$-2/-2$	-1.999 $(-1.999, -1.997)$	-1.999 $(-1.999, -1.998)$
PN, homogeneous (ReLU)	$-2/-2$	-1.997 $(-1.999, -1.994)$	-1.998 $(-1.999, -1.995)$
PN, general ($\tanh$)	$-1/-1$	-0.956 $(-0.970, -0.913)$	-0.966 $(-0.979, -0.937)$

F.1 EXTENDED ACTIVATION-FAMILY VERIFICATION

Figure 4 checks whether the five rates depend on the representative activation used in the main text. Its rows correspond, in order, to unbounded AN, general AN, affine PN, homogeneous PN, and general PN; the three columns replace the activation within each regime. The first row separates the predicted $O(M^{-1})$ LayerNorm and $O(M^{-2})$ RMSNorm rates. Rows two and five test the width-growing $O(r^{-1})$ constructions, whereas rows three and four test the fixed-overhead $O(M^{-2})$ constructions. All fifteen plots have tails parallel to the appropriate reference line, so the rate agreement is not specific to one activation.

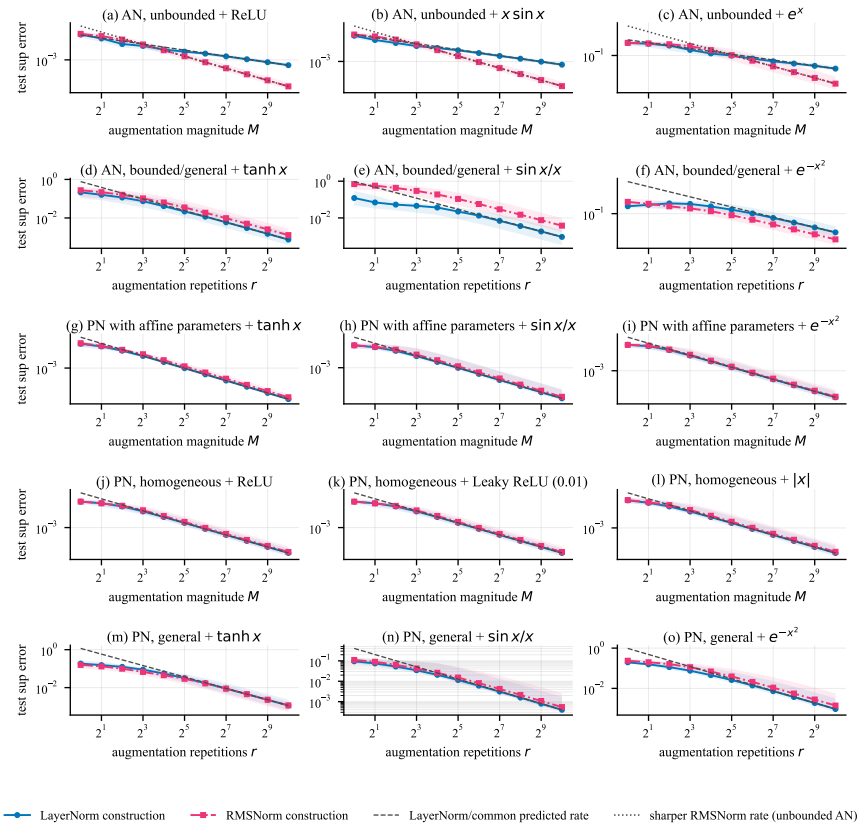


Figure 4: Activation-family verification of the five constructions. Blue and pink curves show LayerNorm and RMSNorm, respectively; shaded regions are interquartile ranges over 24 target–width pairs. Dashed lines show the LayerNorm or common predicted order, and dotted lines in the first row show the sharper RMSNorm order.

F.2 SHIFTED-RELU FINITE-WIDTH BARRIER

The activation $\phi(t) = \mathrm{ReLU}(t - 10)$ makes the finite-width mechanism behind Proposition 5.2 visible for a fixed activation. Figure 5 uses the explicit target $f(x) = \mathrm{ReLU}(2x)$. Figures 5(a) and 5(b) show the target and the no-affine RMS-PN and LN-PN outputs as the normalized width d increases. They remain constant until normalization can reach the shifted ReLU's nonzero region, then approach the target. Figure 5(c) summarizes this transition: the vertical line $d = 10^2$ is the coordinate-bound barrier, after which the explicit errors eventually follow $O(d^{-1})$. In contrast, Figure 5(d) shows that affine parameters retain widths 2 and 3 and yield the predicted $O(M^{-2})$ error. Thus increasing width overcomes the barrier for this fixed shifted ReLU; it does not contradict Proposition 5.2, whose activation depends on d.

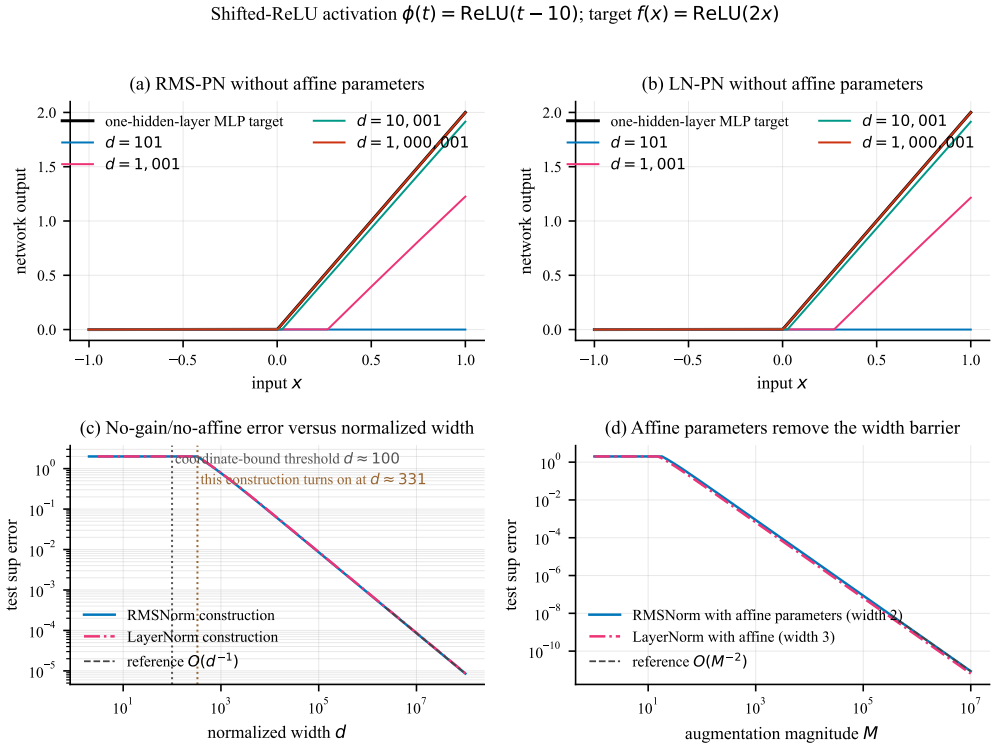


Figure 5: Shifted-ReLU PN constructions for $f(x) = \mathrm{ReLU}(2x)$. (a,b) Function-level recovery as width grows. (c) The finite-width barrier and subsequent $O(d^{-1})$ decay without affine parameters. (d) Fixed-width affine constructions with $O(M^{-2})$ error.

Figure 6 repeats the comparison for 24 random one-hidden-layer MLP targets. Figures 6(a) and 6(b) visualize one target and its approximations; Figure 6(c) aggregates the no-affine constructions over all targets and recovers the width-controlled $O(d^{-1})$ tail; and Figure 6(d) recovers the affine $O(M^{-2})$ tail at fixed width. The interquartile bands show target-to-target variation without changing either conclusion.

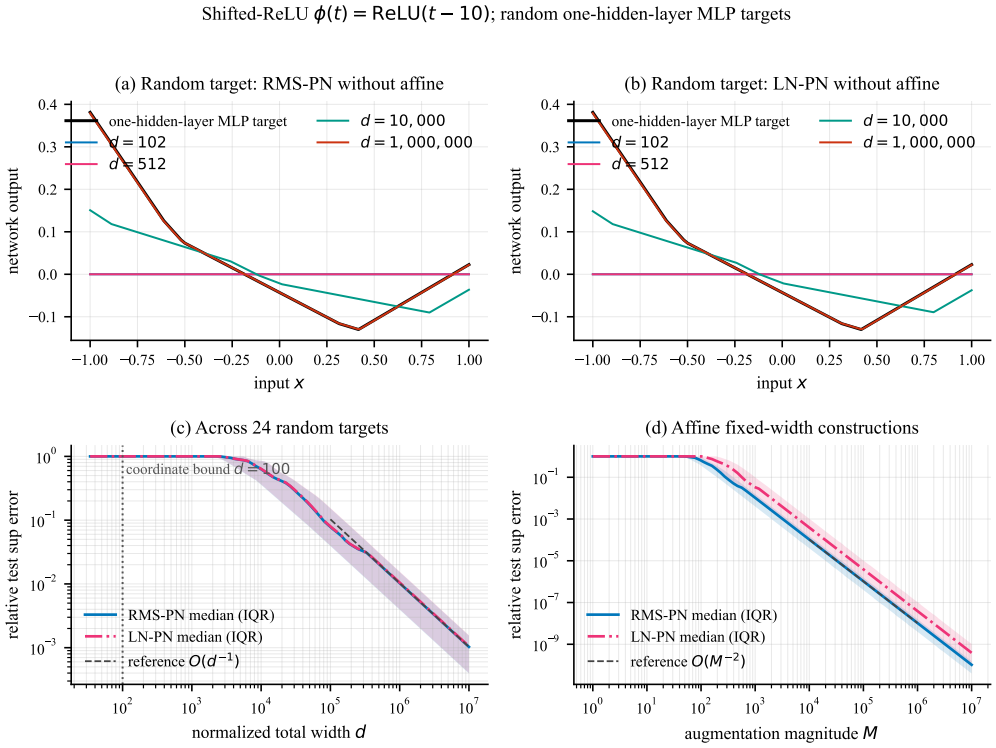


Figure 6: Shifted-ReLU PN verification over 24 random one-hidden-layer MLP targets. (a,b) One representative target. (c,d) Median relative uniform errors and interquartile ranges for the width-growing no-affine and fixed-width affine constructions, respectively.