WIDTH-INDEPENDENT COMPRESSIBILITY OF DEEP NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

It has long been known that well-trained neural networks can be compressed very strongly without affecting their performance, an important phenomenon that remains poorly understood. We prove a uniform compressibility theorem for deep multilayer perceptrons with analytic activations. For a deep, wide fixed teacher network, there exists a narrow (same depth) network that approximately represents the same function as the original. The reachable compressed width is strikingly independent of the original width, but is $O((\log(1/\varepsilon))^{d_{\text{in}}})$, where ε is the error budget and d_{in} is the effective input dimension. Our construction involves a novel derivative-matching technique which is aware of the low-dimensional input, and a layer-wise reweighting that preserves the input-output mapping.

1 INTRODUCTION

Modern neural networks often behave as though only a fraction of their parameters are essential. At language-model scale, weights can be quantized to four bits, matrices can be narrowed by removing rows and columns, and entire layers can be dropped while largely preserving model capabilities (Frantar et al., 2023; Ashkboos et al., 2024; Men et al., 2024; Gromov et al., 2025). That such different interventions succeed points to a common phenomenon: the function learned by a wide network may admit a much smaller representation.

This observation raises a basic theoretical question. Given a particular trained deep network, can we construct a narrow network that approximates the same function *uniformly over a bounded input domain*? If so, how does the size of the compressed network scale if the error budget is limited?

Existing theories do not yet provide this combination. Uniform compression theorems for fixed networks are available only in more restricted settings, such as two-layer permutation-symmetric networks (Wang et al., 2026). What is missing is a constructive, sup-norm guarantee for a fixed deep MLP that composes across layers with controlled error.

We close this gap for analytic-activation deep MLPs. Fix a layer ℓ and write f for the input-to-layer-$(\ell-1)$ map. The contribution of layer ℓ to the next pre-activation is

$$y(x) = \sum_{i=1}^{d_\ell} v_i \, \sigma(w_i^\top f(x)), \tag{1}$$

a finite sum over the d_ℓ neurons of layer ℓ. If f and σ extend holomorphically to suitable complex neighborhoods, every summand is a holomorphic function of the input x on a complex polydisk, and Cauchy's inequality converts finite-order *input-derivative matching at the origin* into a uniform sup-norm bound on a bounded real domain $\Omega \subset \mathbb{R}^d_{\text{in}}$ of radius R: matching all derivatives $\partial_x^\alpha y(0)$ for $|\alpha| \le k$ controls the Taylor remainder by a geometric factor $(R/\rho_0)^{k+1}$, where ρ_0 is determined by the activation's holomorphy radius. The matching condition is finite-dimensional: it imposes only $D_k = \binom{d_{\text{in}}+k}{d_{\text{in}}}$ scalar constraints per output coordinate. Therefore, a deterministic rank-reduction procedure on the derivative-feature matrix retains at most D_k of the original neurons and recomputes the outgoing weights.

Solving for the smallest k that meets a target accuracy ε_ℓ gives a compressed width

$$d'_\ell = O\left(\frac{1}{d_{\text{in}}!}\left(\frac{\log(2C_0/\varepsilon_\ell)}{\lambda}\right)^{d_{\text{in}}}\right), \tag{2}$$

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.

where λ is determined by the ratio between the input radius and the holomorphy radius, and C_0 is determined by the teacher network—mostly the range of weights. Eq. (2) is width-independent in the sense that: after these teacher-specific constants are evaluated, the cap $D_{k_\ell^\star}$ has no additional explicit d_ℓ factor. To compress all hidden layers of a deep MLP, we show that by composing layer-wise compressions in backward order (i.e., from the last hidden layer to the first), the error remains bounded in a similar form as Eq. (2) with inflated constants.

Conceptually, our theory establishes the following rather surprising idea:

> *For a broad class of models (specified in the assumptions), the effective complexity of neural networks solely depends on the data dimension.*

Our technical contributions that lead to this result are:

- A fixed-teacher deep-MLP compression theorem (Theorem 8): backward layer-wise compression of all $L-1$ hidden layers yields a uniformly accurate, same-depth network. For a fixed teacher network, its width caps are polylogarithmic in $1/\varepsilon$ with no additional explicit original-width factor.
- The per-layer derivative-matching error bound that powers it (Theorem 5): matching all input derivatives at the origin up to order k controls the Taylor remainder by a geometric factor $(R/\rho_0)^{k+1}$ in terms of holomorphic radii of f and σ.
- An explicit rank-reduction procedure (Algorithm 1) that selects at most $\binom{d_{\text{in}}+k}{d_{\text{in}}}$ original neurons and recomputes outgoing weights while controlling the error from the network's input to output.

2 Problem setup

Consider an L-layer neural network

$$h^{(0)} = x, \qquad h^{(\ell)} = \sigma\big(W^{(\ell)} h^{(\ell-1)}\big) \quad (\ell = 1, \dots, L-1), \qquad y_{\text{out}} = W^{(L)} h^{(L-1)} \in \mathbb{R}^{d_{\text{out}}}, \tag{3}$$

where $x \in \Omega$ and Ω is a bounded subset of $\mathbb{R}^{d_{\text{in}}}$, and $R := \sup_{x\in\Omega} \|x\|$. Without loss of generality, we assume $0 \in \Omega$. The dimension of the ℓth layer (i.e., width) is d_ℓ, and in particular we denote $d_0 = d_{\text{in}}$ and $d_L = d_{\text{out}}$. Usually, there is a bias term $b^{(\ell)}$ for each layer in Eq. (3). Here, we note that $b^{(\ell)}$ can be absorbed into $W^{(\ell)}$ (for notation simplicity) by appending a constant coordinate equal to 1 to each hidden layer, converting affine maps into linear maps.

We work in the regime $d_{\text{in}} = O(1)$ and $d_1, \dots, d_{L-1} \gg 1$. We aim to compress the hidden layers while approximately preserving the input-to-output mapping. We write $F : \Omega \to \mathbb{R}^{d_{\text{out}}}$ for the input-to-output map of the original network. The goal is to construct a same-depth approximation $\widetilde{F}$ satisfying $\sup_{x\in\Omega} \|\widetilde{F}(x) - F(x)\|_\infty \le \varepsilon$ for a given error budget $\varepsilon > 0$, with retained widths controlled by accuracy-driven dimension bounds.

Throughout, F is one fixed, finite teacher network. Unless uniform control of those quantities is imposed separately, our theorems do not assert a common retained-width bound for a family of teachers whose original widths vary.

In what follows, we will first show the compression of a single intermediate layer, say the ℓ-th layer, given an error budget ε_ℓ, and then compose the compressions of all hidden layers and discuss the error budget assignment across layers.

Effective input dimension. For the layer under consideration, write $f = h^{(\ell-1)}$ for its upstream map. We call m an *admissible effective input dimension* if, after translating coordinates if necessary, there are a bounded set $U \subset \mathbb{R}^m$ with $0 \in U$, a map $\pi : \Omega \to U$, and a map $\bar{f} : U \to \mathbb{R}^{d_{\ell-1}}$ such that $f = \bar{f} \circ \pi$, and the transformed problem obtained by replacing $(\Omega, x, d_{\text{in}}, f)$ with $(U, u, m, \bar{f})$ satisfies Assumptions 1–4. We denote the smallest admissible value by $m_{\text{eff},\ell}$. The ambient choice $m = d_{\text{in}}$, $U = \Omega$, and $\pi(x) = x$ is always the baseline.

We will use the following assumptions, mostly for bounding the Taylor expansion remainder. $\ell \in [L-1]$ is any hidden layer, and $f : \Omega \to \mathbb{R}^{d_{\ell-1}}$ is the input-to-layer-$(\ell-1)$ map, i.e., $f(x) = h^{(\ell-1)}$.

Assumption 1 (Holomorphy of f and σ)**.**

- *f extends holomorphically to a complex ball $B_{\mathbb{C}}(0, \rho_f) \subset \mathbb{C}^d_{\text{in}}$. For $0 < r < \rho_f$, define*

$$L_f(r) \coloneqq \sup_{\|z\| \le r} \|Jf(z)\|_{\text{op}} < \infty. \tag{4}$$

- *$\sigma : \mathbb{R} \to \mathbb{R}$ is real analytic. Let $\mathcal{D}_\sigma \subset \mathbb{C}$ be the domain of its maximal holomorphic extension and, with $W^{(\ell)}_{\max}$ as in Assumption 2 below, define*

$$\rho_\sigma \coloneqq \inf_{\|w\| \le W^{(\ell)}_{\max}} \operatorname{dist}\big(w^\top f(0), \mathbb{C} \setminus \mathcal{D}_\sigma\big) > 0. \tag{5}$$

Real analytic activation functions are common, e.g., sigmoid, tanh, and GELU. In particular, GELU is holomorphic on the entire $\mathbb{C}$, so we expect our error bounds to work for any size of Ω. This assumption does not hold for ReLU. Extending the construction to ReLU would require a separate approximation argument and is left for future work.

Assumption 2 (Per-layer weight bound)**.** *The incoming weights at layer ℓ satisfy*

$$W^{(\ell)}_{\max} \coloneqq \max_{i \in [d_\ell]} \|w^{(\ell)}_i\| < \infty. \tag{6}$$

Assumption 3 (Outgoing ℓ_1-norm bound)**.** *The outgoing weights at layer ℓ satisfy*

$$V_{\max} \coloneqq \max_{j \in [d_{\ell+1}]} \|W^{(\ell+1)}_{j,:}\|_1 < \infty. \tag{7}$$

The effective holomorphic radius of $g(\cdot, w)$ is defined as

$$\rho \coloneqq \sup\Big\{ r \in (0, \rho_f) : r\, W^{(\ell)}_{\max}\, L_f(r) < \rho_\sigma \Big\}. \tag{8}$$

Assumption 4 (Error-budget compatibility)**.** $R < \rho$.

Assumption 4 ensures that the MLP function from the input ($\forall x \in \Omega$) to $h^{(\ell+1)}$ is analytic.

3 Related works

Approximation theory. Classical universal approximation theorems show that shallow networks of sufficient width approximate continuous functions on compact domains (Cybenko, 1989; Hornik et al., 1989; Hornik, 1991; Pinkus, 1999). DeVore et al. (2021) survey the approximation theory literature, including matching parameter-count bounds for Sobolev/Besov targets and the role of unstable parameterization. These works approximate an external target class. We instead approximate a fixed teacher network uniformly on a bounded domain, with an accuracy-driven retained-width bound. Kolmogorov-width and metric-entropy asymptotics for analytic function balls strongly indicate that the resulting $[\log(1/\varepsilon)]^m$ dependence, with m the effective input dimension, is not an artifact of our algorithm but has information-theoretic grounds; Appendix E makes this comparison precise.

Depth, width, and smoothness. Depth improves approximation in Sobolev and related smoothness classes (Yarotsky, 2017; 2018; Lu et al., 2021; Schmidt-Hieber, 2020; Liu et al., 2026), while *depth–width separations* show that certain functions require exponential width when depth is fixed (Telgarsky, 2016; Eldan & Shamir, 2016), and that bounded-width networks recover universality only by growing depth (Lu et al., 2017; Kidger & Lyons, 2020; Hanin, 2019; Vardi et al., 2022). Our theory preserves the original depth and gives each fixed covered teacher a same-depth approximation. The depth–width separation claims crucially rely on a family of increasingly hard functions, whereas our theory treats properties of a teacher network as fixed.

Empirical compression at scale. Practical compression is dominated by four primitives: pruning, quantization, low-rank factorization, and distillation (Cheng et al., 2018; Hoefler et al., 2021), with foundational milestones in deep compression (Han et al., 2016), lottery tickets (Frankle & Carbin, 2019), and the methodological review of Blalock et al. (2020). At LLM scale, post-training

compression also spans structured row/column deletion (Ashkboos et al., 2024), depth pruning (Men et al., 2024; Gromov et al., 2025), and weight-quantization variants (Frantar et al., 2023). These primitives are data-dependent and architecture-specific. They overlap with our work on application to post-training compression of a deep network without fine-tuning and thus provide motivation for this work. Our result is different in kind: it proves a uniform function-error certificate and a corresponding retained-width cap, not a claim of empirical superiority over these methods.

Theoretical compression bounds and distillation. Existing constructive guarantees occupy a narrow strip: data-dependent coresets (Baykal et al., 2019; Liebenwein et al., 2020), spectral/activation-covariance pruning (Suzuki et al., 2020), sparse-linear-approximation viewpoints (Yang et al., 2022), NTK-preserving pruning of two-layer ReLU networks (Yang et al., 2023), low-rank recovery under approximately-low-rank activations (Zhang & Saab, 2026), and randomized greedy schemes (El Cheairi et al., 2025). Each is data-dependent, primitive-specific, or restricted to two-layer networks. Distillation (Hinton et al., 2015; Romero et al., 2015), where a student is trained to mimic the teacher's function, is the closest practical cousin of derivative matching; theoretical accounts (Phuong & Lampert, 2019; Allen-Zhu & Li, 2023) establish formal reductions in restricted settings (linear classifiers; multi-view data) but do not give explicit width bounds for deep nets, and the matching is on data rather than on the local Taylor expansion.

Tensor moment methods. Tensor-moment–like methods have been used to understand neural networks in several ways, see e.g., (Anandkumar et al., 2014; Janzamin et al., 2015). The methodological predecessor of this work is Wang et al. (2026). It proves a polylogarithmic compression theorem for permutation-symmetric objects via tensor moment matching of the weight cloud. The present work advances that direction in three structural ways. First, the finite basis is the input-derivative basis at the origin, not a tensor-moment basis over weights. This resolves the problem of (Wang et al., 2026) being unable to compress a hidden layer whose previous or next layer is also wide. Second, for one fixed teacher our retained-width formula has no additional explicit d_ℓ factor, whereas the bound of Wang et al. (2026) contains a polylogarithmic factor in the original width. Third, here we establish compressibility of all layers of a deep network, whereas Wang et al. (2026) treats a two-layer MLP with narrow input and output layers.

4 Proof idea for compressing one layer

In this section, we focus on the following function defined by the ℓ-th layer:

$$\begin{aligned} x \mapsto y &= W^{(\ell+1)}\,\sigma\big(W^{(\ell)}h^{(\ell-1)}\big) \\ &= \sum_{i=1}^{d_\ell} v_i\,\sigma\big(w_i^\top f(x)\big), \end{aligned} \tag{9}$$

where $f:\Omega\to\mathbb{R}^{d_{\ell-1}}$ is the input-to-layer-($\ell-1$) map, v_i's are the column vectors of $W^{(\ell+1)}$, and $w_i^\top$'s are the row vectors of $W^{(\ell)}$. We also write $g(x,w) \coloneqq \sigma(w^\top f(x))$. We aim to replace

$$\begin{aligned} W^{(\ell)} \in \mathbb{R}^{d_\ell\times d_{\ell-1}} \;\text{ by }\; \widetilde{W}^{(\ell)} \in \mathbb{R}^{d'_\ell\times d_{\ell-1}}, \\ W^{(\ell+1)} \in \mathbb{R}^{d_{\ell+1}\times d_\ell} \;\text{ by }\; \widetilde{W}^{(\ell+1)} \in \mathbb{R}^{d_{\ell+1}\times d'_\ell}, \end{aligned} \tag{10}$$

where d'_ℓ is the compressed width.

Measure compression. We start by reformulating the ℓ-th layer as an integration over the empirical measure of its neurons, the mean-field view of a network (Mei et al., 2018). Define the counting measure $\mu=\sum_{i=1}^{d_\ell}\delta_{(v_i,w_i)}$ on $\mathbb{R}^{d'_\ell}\times\mathbb{R}^{d_{\ell-1}}$ over pairs of outgoing and incoming weights. For any function $a(v,w)$ and measure ν,

$$\langle a(v,w)\rangle_\nu \coloneqq \int a(v,w)\,d\nu(v,w). \tag{11}$$

With this notation, the ℓ-th layer can be written as $y(x)=\langle v g(x,w)\rangle_\mu$. The compression problem is then to find a new measure $\tilde\mu$ supported on d'_ℓ points, such that

$$\sup_{x\in\Omega}\big\|\langle v\,g(x,w)\rangle_{\tilde\mu}-\langle v\,g(x,w)\rangle_\mu\big\|_\infty \;\le\; \varepsilon_\ell, \tag{12}$$

where $\varepsilon_\ell>0$ is the error budget allocated to this layer.

Derivative matching. Eq. (12) has to be satisfied for a large class of functions $y(x)$. One approach (Wang et al., 2026) is to match the expectation of low-degree polynomials of v, w, so that the error occurs only in the Taylor expansion tail (assuming any relevant function of $W^{(\ell)}$ is, in a sense, analytic, so that Taylor expansions converge to the actual function). Wang et al. (2026) matches a monomial basis with $\binom{d_{\text{in}}+d_{\text{out}}+k}{k}$ compressed points and proves that this is optimal if y's dependence on v and w is a *generic* permutation-symmetric function. However, here we surpass the bound by exploiting the specific expression of MLPs: it suffices to *match the expectation of the derivatives of g, which has a much smaller basis* than the monomial basis.

Note that we need to match the expectations $\langle v\, g(x,w)\rangle_\mu$—nothing more and nothing less, and that g as a function of w is parametrized by $O(1)$ parameters, i.e., by $x \in \Omega$. Assume g is analytic in a sufficiently large range (see Assumption 4) so that a Taylor expansion exists and converges to the actual function. From

$$g(x,w) \approx \sum_{|\alpha|\le k} \frac{\partial_x^\alpha g(0,w)}{\alpha!} x^\alpha, \tag{13}$$

(where α is a multi-index defined in Appendix A), it follows that

$$\langle v\, g(x,w)\rangle_\mu \approx \sum_{|\alpha|\le k} \frac{x^\alpha}{\alpha!} \langle v\, \partial_x^\alpha g(0,w)\rangle_\mu. \tag{14}$$

Therefore, if we change a measure but preserve the expectations $\langle v\, \partial_x^\alpha g(0,w)\rangle$ for $|\alpha| \le k$, they represent almost the same input-output map. The compression error starts from the degree-$(k+1)$ Taylor remainder, which is thus of order $(R/\rho)^{k+1}$. This is summarized in the following theorem:

Theorem 5 (Derivative-matching error)**.** *Let $k \ge 0$, let $\tilde\mu = \sum_{i\in\mathcal{K}} \delta_{(\tilde v_i, w_i)}$ with $\mathcal{K} \subseteq [d_\ell]$, and write $\tilde y(x) \coloneqq \langle v\, g(x,w)\rangle_{\tilde\mu} = \sum_{i\in\mathcal{K}} \tilde v_i\, g(x,w_i)$. Suppose*

$$\big\langle v\, \partial_x^\alpha g(x,w)\big|_{x=0}\big\rangle_{\tilde\mu} \;=\; \big\langle v\, \partial_x^\alpha g(x,w)\big|_{x=0}\big\rangle_{\mu} \qquad \textit{for every } \alpha \in \mathbb{N}_{\text{in}}^d \textit{ with } |\alpha| \le k, \tag{15}$$

where both sides are $\mathbb{R}^{d_{\ell+1}}$-valued. Then for every $j \in [d_{\ell+1}]$ and every $x \in \Omega$,

$$\big|y_j(x) - \tilde y_j(x)\big| \;\le\; \Big(\|v_{:,j}\|_1 + \|\tilde v_{:,j}\|_1\Big)\, M_g\, \frac{\rho_0}{\rho_0 - R} \Big(\frac{R}{\rho_0}\Big)^{k+1}, \tag{16}$$

where $v_{:,j} \coloneqq (v_{ij})_{i\in[d_\ell]} \in \mathbb{R}^{d_\ell}$ and $\tilde v_{:,j} \coloneqq (\tilde v_{ij})_{i\in\mathcal{K}} \in \mathbb{R}^{|\mathcal{K}|}$. M_g is a constant determined by g: pick any $\rho_0 \in (R,\rho)$,

$$M_g \;\coloneqq\; \sup_{\substack{\|w\|\le W_{\max}^{(\ell)} \\ \|z\|\le\rho_0}} \big|g(z,w)\big|. \tag{17}$$

From here on, we define the *derivative feature map*:

$$\Phi(w) \;\coloneqq\; \big(\partial_x^\alpha g(x,w)\big|_{x=0}\big)_{|\alpha|\le k} \in \mathbb{R}^{D_k}, \qquad D_k \;\coloneqq\; \binom{d_{\text{in}}+k}{d_{\text{in}}}. \tag{18}$$

Stacking over the d_ℓ incoming rows of $W^{(\ell)}$,

$$\Phi(W^{(\ell)}) \;\coloneqq\; \big(\Phi(w_1) \quad \cdots \quad \Phi(w_{d_\ell})\big) \;\in\; \mathbb{R}^{D_k \times d_\ell}. \tag{19}$$

Rank reduction and reweighting. The task of finding a compressed measure $\tilde\mu$ has now become a linear algebra problem: find a matrix $V \in \mathbb{R}^{d_{\ell+1}\times d_\ell}$ that satisfies $V\,\Phi(W^{(\ell)})^\top = W^{(\ell+1)}\,\Phi(W^{(\ell)})^\top$. The i-th column of V is $\tilde v_i$, and the new measure is $\tilde\mu = \sum_{\substack{i\in[d_\ell]\\ \tilde v_i \ne 0}} \delta_{(\tilde v_i, w_i)}$. It turns out there exists a V with at most D_k nonzero columns: we prove this in Lemma 10 and also propose Algorithm 1 to find such a V. Furthermore, $V_{\max}$ can be guaranteed to be inflated by at most a factor of D_k, which could affect the next layer's error bound but by a controlled amount.

Main theorem and width bound. Fix any $\rho_0 \in (R,\rho)$, recall M_g from Eq. (17), and define the k-independent constants

$$C_0 \;\coloneqq\; V_{\max}\, M_g\, \frac{\rho_0}{\rho_0 - R} \quad \text{and} \quad \lambda \;\coloneqq\; \log(\rho_0/R) \;>\; 0. \tag{20}$$

If $C_0 = 0$, the layer contribution is degenerate and the zero-output replacement has zero error. We therefore state the nontrivial case $C_0 > 0$ below.

The whole approach above culminates in the following theorem for compressing a single hidden layer. In plain language, it answers the question: if there is an error budget ε_ℓ, what matching order k suffices? By solving error $\sim (R/\rho_0)^{k+1}$, it suffices to take $k^* = O(\log(1/\varepsilon_\ell))$, which gives the compressed width $D_{k^\star(\varepsilon_\ell)} = O((\log(1/\varepsilon_\ell))^{d_{\text{in}}})$.

Theorem 6 (Intermediate-layer compression)**.** *Fix the teacher network and a hidden layer satisfying Assumptions 1–4. For any $\rho_0 \in (R, \rho)$ with $C_0 > 0$ and λ as in Eq.* (20)*, any error budget $\varepsilon_\ell > 0$, and any integer $k \geq 0$ satisfying*

$$(k+1)\,\lambda \;\geq\; \log(2\,C_0/\varepsilon_\ell) \;+\; d_{\text{in}}\,\log(k+d_{\text{in}}) \;-\; \log(d_{\text{in}}!), \tag{21}$$

there exist an index set $\mathcal{K} = \{\kappa_1, \ldots, \kappa_n\} \subseteq [d_\ell]$ with $n = |\mathcal{K}| \leq D_k$, where $D_k := \binom{d_{\text{in}}+k}{d_{\text{in}}}$, and weight matrices $\widetilde{W}^{(\ell)} \in \mathbb{R}^{n \times d_{\ell-1}}$, $\widetilde{W}^{(\ell+1)} \in \mathbb{R}^{d_{\ell+1} \times n}$ (constructed in Corollary 12) satisfying $\widetilde{W}^{(\ell)}_{r,:} = W^{(\ell)}_{\kappa_r,:}$ for every $r \in [n]$, such that the compressed network $\tilde{y}(x) = \widetilde{W}^{(\ell+1)}\sigma(\widetilde{W}^{(\ell)} f(x))$ satisfies

$$\sup_{x \in \Omega} \big\| y(x) - \tilde{y}(x) \big\|_\infty \;\leq\; \varepsilon_\ell. \tag{22}$$

Asymptotics. *Define*

$$k^\star(\varepsilon_\ell) \;:=\; \min\big\{k \in \mathbb{N}_0 : k \text{ satisfies Eq. } (21)\big\}. \tag{23}$$

Then, as $\varepsilon_\ell \to 0^+$ with this teacher and $d_{\text{in}}, \rho_0, R, V_{\max}, M_g$ fixed,

$$k^\star(\varepsilon_\ell) \;=\; \frac{\log(2C_0/\varepsilon_\ell)}{\lambda} \;+\; O\big(\log\log(1/\varepsilon_\ell)\big) \;=\; \Theta\big(\log(1/\varepsilon_\ell)\big), \tag{24}$$

and consequently

$$D_{k^\star(\varepsilon_\ell)} \;=\; O\left(\frac{1}{d_{\text{in}}!}\left(\frac{\log(2C_0/\varepsilon_\ell)}{\lambda}\right)^{d_{\text{in}}}\right). \tag{25}$$

Remark 7 (Width dependence and finite-scale compression)**.** With the fixed teacher constants and d_{in} held fixed, $k^\star(\varepsilon_\ell) = \Theta(\log(1/\varepsilon_\ell))$ and $D_{k^\star(\varepsilon_\ell)} = O((\log(1/\varepsilon_\ell))^{d_{\text{in}}})$. If M_g and the radius quantities are controlled uniformly across a family, dependence on $V_{\max}$ enters the cap only through $\log V_{\max}$; if that outgoing norm is also uniformly bounded, the cap is uniform in d_ℓ. For this bound to give a genuinely narrower layer, the original width must satisfy $d_\ell > D_{k^\star(\varepsilon_\ell)}$. Appendix D gives sharper finite-width bookkeeping, a closed-form sufficient threshold, and the realized-rank caveat.

Corollary (exact analytic parameterization). Suppose $\Omega = \iota(U)$ for a one-to-one real-analytic parameterization $\iota : U \subset \mathbb{R}^m \to \mathbb{R}^d_{\text{in}}$, and suppose the reparameterized upstream map $f \circ \iota$ satisfies Assumptions 1–4 in the u-coordinates (after translation if needed). Applying Theorem 6 to u gives the same uniform error guarantee on Ω with $D_k = \binom{m+k}{m}$ and hence retained width $O((\log(1/\varepsilon_\ell))^m)$, with constants determined in the reparameterized coordinates.

Corollary (architectural bottleneck). Fix $j < \ell$ and use $u = h^{(j)}(x)$ as the input to the suffix from layer $j+1$ through layer ℓ. If that suffix satisfies the transformed assumptions on $h^{(j)}(\Omega)$ after recentering, then d_j is admissible and Theorem 6 holds with $D_k = \binom{d_j+k}{d_j}$. Whenever the transformed assumptions hold for the candidate bottlenecks, one may therefore use the architectural bound

$$m_{\text{eff},\ell} \;\leq\; m^{\text{arch}}_\ell \;:=\; \min\{d_{\text{in}}, d_1, \ldots, d_{\ell-1}\}. \tag{26}$$

Numerics. We present small-scale numerical experiments to test the construction and its predicted geometric error decay over a finite range of derivative orders and input radii.

We instantiate Theorem 6 on a single hidden layer of a randomly initialized two-layer network with input dimension $d_{\text{in}} = 3$, hidden width $d_1 = 2048$, and output dimension $d_{\text{out}} = 2048$. We compress the hidden layer from 2048 neurons down to $D_k = \binom{3+k}{3}$ derivative-matched neurons. Two activations with qualitatively different properties are tested on identical random weights:

1. sigmoid, which is real-analytic with poles at $\pm i\pi$ in $\mathbb{C}$, therefore $\rho_\sigma = \pi$;

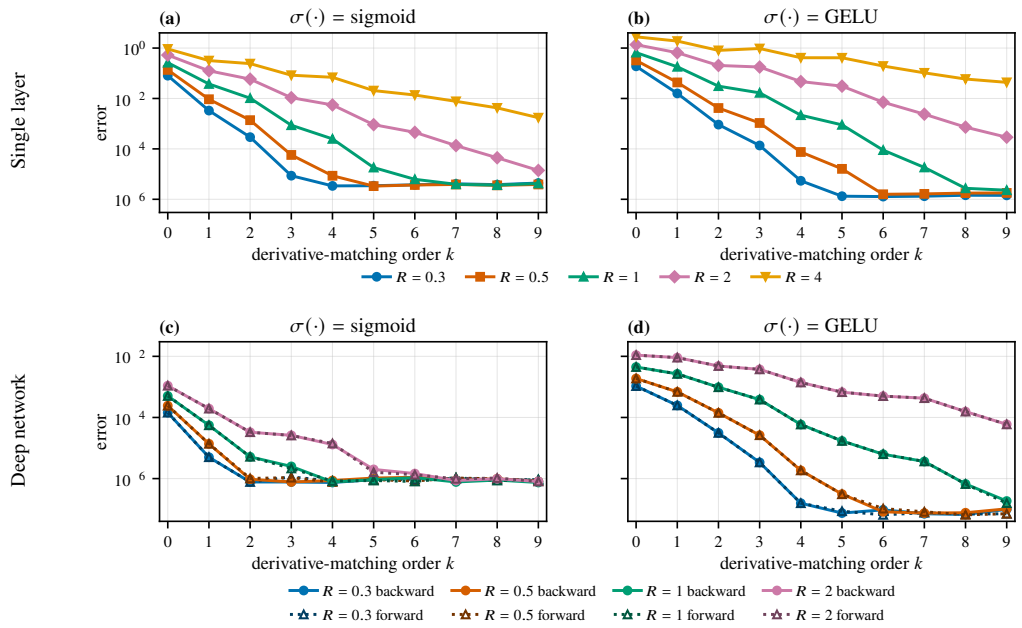

Figure 1: Compression error for randomly initialized single-layer (top, $[3, 2048, 2048]$) and four-hidden-layer (bottom, $[2, 1024, 1024, 1024, 1024, 2]$) networks, with sigmoid (a,c) and GELU (b,d), for $k = 0, \ldots, 9$ and varying input radius R. Here "error" is $\max_{x \in \mathcal{X}_R} \|F(x) - \widetilde{F}(x)\|_\infty$, where $\mathcal{X}_R$ is the evaluation sample from the radius-R ball.

2. GELU, which is entire on $\mathbb{C}$ but whose k-th derivative at the origin grows superexponentially with k.

Figure 1(a,b) is consistent with the qualitative finite-range prediction of Theorem 6 for both sigmoid and GELU. Both top-row panels show clean geometric decay in k at every radius, with the per-step rate slowing as R grows — consistent with the $(R/\rho_0)^{k+1}$ prediction. For $R \leq 1$ the error reaches the float32 noise floor ($\sim 10^{-6}$) by $k \approx 5$ for both activations; even at the largest tested radius $R = 4$, which already exceeds $\rho_\sigma = \pi$ for sigmoid, the error still decreases exponentially over $k = 0, \ldots, 9$. As a remark, we would have expected that in the GELU case the error decays visibly faster than in the sigmoid case since $\rho_\sigma = \infty$ for $\sigma = \mathrm{GELU}$. However, the error decreases with k for every tested R. A possible explanation is that the higher derivatives of GELU grow rapidly at the origin, which hinders the error decay with k.

5 Compressing all layers

In this section, we discuss compression of all layers of a deep MLP. Applying Theorem 6 to every hidden layer of one fixed teacher yields retained-width bounds that can be composed. We need to address two additional questions here: (1) What is the optimal order to compress the layers? (2) If given a total error budget ε, how to assign it to each layer's compression? We will briefly discuss these caveats in the rest of this section, and leave a formal treatment to Appendix C.

The short answers are: (1) Compressing the layers backward from the last hidden layer to the first keeps each step's analytic constants bounded by their original-network values (Appendix C, Proposition 17), which licenses the budget split; forward order inflates these constants recursively (Proposition 18) and yields no rigorous guarantee. Backward order is therefore preferred, though for moderately sized networks one is empirically free to choose any order without paying a measurable price. (2) When compressing in backward order, splitting the budget evenly as $\varepsilon_\ell = \varepsilon / \left((L-1)\,\overline{\Lambda}\right)$ — where $\overline{\Lambda} \geq 1$ is an a-priori upper bound on the downstream Lipschitz factor, computed from the *original* network (Appendix C, Lemma 19) — guarantees total error at most ε.

We provide intuitive reasoning for these facts. Per Sec. 4, after compressing the ℓ-th layer, $W^{(\ell)}$ becomes a subset of the original, whereas $W^{(\ell+1)}$ is reweighted, and its maximum row ℓ_1-norm increases at most by a factor of D_k. If we compress backward, the inflated $W^{(\ell+1)}$ is never compressed again in the subsequent compressions, so the error bounds of the layers essentially do not influence each other. However, if we compress forward, the inflated $W^{(\ell+1)}$ should be used to calculate the effective holomorphic radius at the next layer, which in the worst case shrinks ρ by D_k—and this effect stacks as we compress deeper layers. The practical reason why the error of forward-order compression does not explode (see Fig. 1(c,d)) is presumably that the function $f : x \to h^{(\ell+2)}$ is not changed by much due to the compression of the ℓ-th layer, even if $W^{(\ell+1)}$ is inflated. Nevertheless, this does not lead to a rigorous error guarantee for forward-order compression.

Our main theoretical error bound for deep MLP compression is as follows, rigorously stated as Theorem 22.

Theorem 8 (Deep MLP compression; informal)**.** *For any error budget $\varepsilon > 0$, compressing the hidden layers of an L-layer MLP F in backward order with a uniform per-layer budget yields a compressed network $\widetilde{F}$ satisfying*

$$\sup_{x\in\Omega} \left\| \widetilde{F}(x) - F(x) \right\|_\infty \;\le\; \varepsilon, \tag{27}$$

in which every hidden layer has compressed width

$$d'_\ell \;=\; O\Big(\big(\log(L/\varepsilon)\big)^{d_{\mathrm{in}}} \Big), \qquad \ell = 1, \dots, L-1, \tag{28}$$

where the suppressed constants depend only on the activation σ and on the weight matrices of the original network (through the per-layer holomorphic radii, the downstream Lipschitz factors, and the constants in Theorem 6). Consequently, for fixed $L, d_{\mathrm{in}}, d_{\mathrm{out}}$ and the fixed teacher's constants, the total parameter count of $\widetilde{F}$ is polylogarithmic in $1/\varepsilon$. Equation (69) *records the explicit dependence on L.*

Numerics. We instantiate the composition theorem on a deep network with $L = 5$, four hidden layers of width 1024, $d_{\mathrm{in}} = d_{\mathrm{out}} = 2$. Compression is applied to all four hidden layers sequentially with a uniform per-step order k, in backward or forward order. The two orders are applied to the same initialized weights for two activations (sigmoid and GELU).

The bottom row of Figure 1 reports the resulting test error. Both bottom-row panels exhibit clean geometric decay in k at every radius, with the per-step rate slowing as R grows — the same $(R/\rho_0)^{k+1}$-type behavior as in the single-layer experiment, now compounded across four hidden layers. Sigmoid reaches the float32 noise floor by $k \approx 4$ for $R \le 1$ and by $k \approx 7$ for $R = 2$; at $k = 9$, GELU reaches order 10^{-7} for $R \le 1$ and approximately 6×10^{-5} for $R = 2$. At the largest plotted order ($k = 9$) every hidden layer has been compressed from 1024 to 55 neurons — a $\sim 19\times$ width reduction per layer, two orders of magnitude in per-layer parameter count.

Post-training compression. In Fig. 2, we test the same construction after unconstrained training. On the unit disk in $\mathbb{R}^2$, five teacher networks with sigmoid activation are fit to

$$q(x_1, x_2) = \sin(2x_1) + \tfrac{1}{2}\cos(3x_2) + \tfrac{1}{4}x_1x_2. \tag{29}$$

We use 20,000 Sobol training points and evaluate on 150,000 held-out Sobol points followed by projected local searches starting from the largest observed discrepancies. For each seed, the frozen randomly initialized network is compressed by the identical procedure as a control. Matching orders $k = 0, \dots, 10$ give $D_k = \binom{k+2}{2}$.

We use a predeclared task-preservation criterion: the compressed network's target MSE be at most 1.05 times its teacher's MSE. Across the five seeds, the median first width meeting this criterion is 28, corresponding to a $304\times$ reduction in total parameter count. Assumption 4 is violated for all trials (especially at layer 2, R/ρ is at least 3), but the compression construction still yields a practical width reduction whose error decreases substantially over the tested range as k increases. This further supports that the applicability of the construction is broader than the theorem's formal guarantee. Appendix G gives the full protocol, per-seed results, and numerical checks.

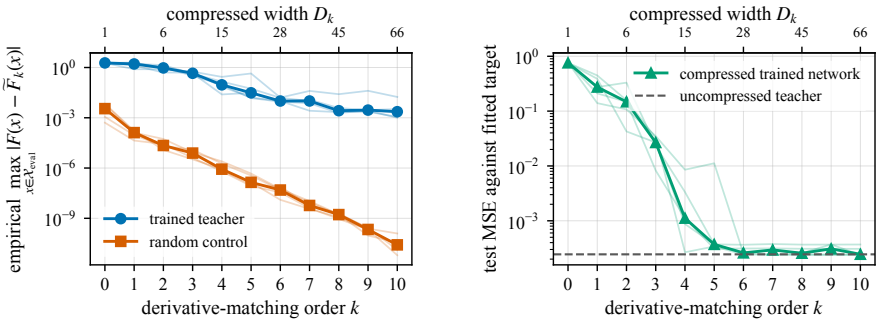

Figure 2: Compression of all three hidden layers of a trained $[2, 512, 512, 512, 1]$ sigmoid MLP, without fine-tuning. Thin curves are the five seeds and bold curves their medians. Left: empirical maximum teacher–compression discrepancy on the held-out set and projected local searches; frozen random controls are shown for comparison. Right: target MSE, with the median uncompressed-teacher MSE dashed.

6 Discussion and Outlook

We have established a uniform sup-norm approximation theorem for analytic deep MLPs based on input-derivative matching. At layer ℓ, the construction retains $d'_\ell \leq D_{k^\star_\ell}$ neurons. For the fixed teacher, $D_{k^\star_\ell}$ is polylogarithmic in the inverse per-layer error. The certificate therefore has no additional explicit original-width factor once these constants are evaluated, but it can depend indirectly on width through the teacher's norms and downstream factors. The construction is explicit: compute derivative features, select a rank basis, and recompute the outgoing weights. To our knowledge, this is the first uniform compression theorem that applies to arbitrary intermediate layers of a deep MLP and composes layer by layer with a controlled error budget. It supplies a function-level existence certificate complementary to practical post-training methods.

Our construction (and associated numerical results) serves as a proof of existence and is not necessarily practical as-is. In particular, it is compute- and memory-expensive to evaluate the high-order derivative features. We do expect future work to improve on the residual error (expectedly by a constant factor) and to reduce the computational cost of compression.

The ambient feature count $\binom{d_{\text{in}}+k}{d_{\text{in}}}$ can be replaced by $\binom{m+k}{m}$ for any admissible effective input dimension m. Extending this result to approximately low-dimensional or noisy near-manifold inputs is an important open problem that would bring the theory closer to settings relevant to practical post-training compression. Empirical intrinsic-dimension estimates for common image datasets lie roughly between 10 and 50 (Pope et al., 2021). These estimates do not by themselves provide an admissible effective dimension under our theorem, but they illustrate the potential scale: if we match up to the fifth derivatives, the retained width ranges from 10^3 to 10^6. Thus, a successful near-manifold extension could yield retained-width bounds relevant to large-model compression.

Finally, the construction has several conceptual implications, all of which are worth exploring in future work:

1. *Critical retained-width.* Whether a network class admits a universal critical width, or a matching information-theoretic lower bound, remains open.
2. *Token granularity.* There has been a tendency to train LLMs with low precision ε to gain speed. However, our theory (also Fig. 1 visually) suggests that wide networks do not store more patterns than a $\text{polylog}(1/\varepsilon)$-wide network. It is thus worth exploring the concrete tradeoff between model expressivity and numerical precision.
3. *Proof of lottery ticket hypothesis.* The existence of universal compression approaches could lead to a proof of the dynamical lottery ticket hypothesis (DLTH), as elucidated by Wang et al. (2026). Our compression method does not establish DLTH, but its derivative-matching construction may suggest tools for studying DLTH in deep neural networks.

AI USE STATEMENT

In this work, we used generative AI tools for writing codes for implementing the experiments (all designed by the authors), checking and improving derivation, and language refinement. We have not used generative AI tools for formulating ideas or proving mathematical claims. The rest of the required disclosure tasks (handling data sets and data analysis, propose or refine hypotheses, assist with translation etc.) are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *International Conference on Learning Representations*, 2023.

Anima Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15: 2773–2832, 2014.

Saleh Ashkboos, Maximilian L. Croci, Marcelo Gennari do Nascimento, Torsten Hoefler, and James Hensman. SliceGPT: Compress large language models by deleting rows and columns. In *International Conference on Learning Representations*, 2024.

Oscar F. Bandtlow and Stéphanie Nivoche. New solution of a problem of Kolmogorov on width asymptotics in holomorphic function spaces. *Journal of the European Mathematical Society*, 24 (7):2493–2532, 2022. doi: 10.4171/JEMS/1148.

Cenk Baykal, Lucas Liebenwein, Igor Gilitschenski, Dan Feldman, and Daniela Rus. Data-dependent coresets for compressing neural networks with applications to generalization bounds. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HJfwJ2A5KX`.

Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? In *Proceedings of Machine Learning and Systems*, volume 2, pp. 129–146, 2020.

Yu Cheng, Duo Wang, Pan Zhou, and Tao Zhang. Model compression and acceleration for deep neural networks: The principles, progress, and challenges. *IEEE Signal Processing Magazine*, 35 (1):126–136, 2018. doi: 10.1109/MSP.2017.2765695.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989. doi: 10.1007/BF02551274.

Ronald DeVore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, 2021. doi: 10.1017/S0962492921000052.

Ronald A. DeVore and George G. Lorentz. *Constructive Approximation*, volume 303 of *Grundlehren der mathematischen Wissenschaften*. Springer, Berlin, 1993. doi: 10.1007/978-3-662-02888-9.

Houssam El Cheairi, David Gamarnik, and Rahul Mazumder. Theoretical compression bounds for wide multilayer perceptrons. *arXiv preprint arXiv:2512.06288*, 2025. URL `https://arxiv.org/abs/2512.06288`.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Proceedings of the 29th Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 907–940. PMLR, 2016.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations*, 2023.

Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A. Roberts. The unreasonable ineffectiveness of the deeper layers. In *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=ngmEcEer8a`.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In *International Conference on Learning Representations*, 2016.

Boris Hanin. Universal function approximation by deep neural nets with bounded width and ReLU activations. *Mathematics*, 7(10):992, 2019. doi: 10.3390/math7100992.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. NIPS 2014 Deep Learning and Representation Learning Workshop.

Torsten Hoefler, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Journal of Machine Learning Research*, 22(241):1–124, 2021.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2): 251–257, 1991. doi: 10.1016/0893-6080(91)90009-T.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. doi: 10.1016/0893-6080(89)90020-8.

Majid Janzamin, Hanie Sedghi, and Anima Anandkumar. Beating the perils of non-convexity: Guaranteed training of neural networks using tensor methods, 2015.

Patrick Kidger and Terry Lyons. Universal approximation with deep narrow networks. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2306–2327. PMLR, 2020. URL https://proceedings.mlr.press/v125/kidger20a.html.

A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *American Mathematical Society Translations, Series 2*, 17:277–364, 1961. English translation of Uspekhi Matematicheskikh Nauk 14(2), 3–86 (1959).

Lucas Liebenwein, Cenk Baykal, Harry Lang, Dan Feldman, and Daniela Rus. Provable filter pruning for efficient neural networks. In *International Conference on Learning Representations*, 2020.

Yuhao Liu, Zilin Wang, Lei Wu, and Shaobo Zhang. Smoothness adaptivity in constant-depth neural networks: Optimal rates via smooth activations. *arXiv preprint arXiv:2602.19691*, 2026. URL https://arxiv.org/abs/2602.19691.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021. doi: 10.1137/20M134695X.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6231–6239, 2017.

Vitaly Maiorov and Allan Pinkus. Lower bounds for approximation by MLP neural networks. *Neurocomputing*, 25(1–3):81–91, 1999. doi: 10.1016/S0925-2312(98)00111-8.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. doi: 10.1073/pnas.1806579115.

Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. ShortGPT: Layers in large language models are more redundant than you expect, 2024.

Mary Phuong and Christoph Lampert. Towards understanding knowledge distillation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5142–5151. PMLR, 2019.

Allan Pinkus. *n-Widths in Approximation Theory*, volume 7 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer, Berlin, 1985. doi: 10.1007/978-3-642-69894-1.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8: 143–195, 1999. doi: 10.1017/S0962492900002919.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=XJk19XzGq2J.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020. doi: 10.1214/19-AOS1875.

Maxwell B. Stinchcombe and Halbert White. Approximating and learning unknown mappings using multilayer feedforward networks with bounded weights. In *Proceedings of the International Joint Conference on Neural Networks*, volume 3, pp. 7–16. IEEE, 1990.

Taiji Suzuki, Hiroshi Abe, Tomoya Murata, Shingo Horiuchi, Kotaro Ito, Tokuma Wachi, So Hirai, Masatoshi Yukishima, and Tomoaki Nishimura. Spectral pruning: Compressing deep neural networks via spectral analysis and its generalization error. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI-PRICAI)*, pp. 2839–2846, 2020. doi: 10.24963/ijcai.2020/393.

Matus Telgarsky. Benefits of depth in neural networks. In *Proceedings of the 29th Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 1517–1539. PMLR, 2016.

Gal Vardi, Gilad Yehudai, and Ohad Shamir. Width is less important than depth in ReLU neural networks. In *Proceedings of the 35th Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 1249–1281. PMLR, 2022. URL https://proceedings.mlr.press/v178/vardi22a.html.

Hong-Yi Wang, Di Luo, Tomaso Poggio, Isaac L. Chuang, and Liu Ziyin. A universal compression theory for lottery ticket hypothesis and neural scaling laws. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=vxkzW4ljeX.

Hongru Yang, Yingbin Liang, Xiaojie Guo, Lingfei Wu, and Zhangyang Wang. Pruning before training may improve generalization, provably, 2023.

Wenjing Yang, Ganghua Wang, Jie Ding, and Yuhong Yang. A theoretical understanding of neural network compression from sparse linear approximation. *arXiv preprint arXiv:2206.05604*, 2022. URL https://arxiv.org/abs/2206.05604.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2017. doi: 10.1016/j.neunet.2017.07.002.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 639–649. PMLR, 2018.

Shihao Zhang and Rayan Saab. Theoretical guarantees for low-rank compression of deep neural networks. *Applied and Computational Harmonic Analysis*, 2026. URL https://arxiv.org/abs/2502.02766. in press.

A NOTATIONS

Multi-index calculus. For a multivariate function on $\mathbb{R}^{d_{\text{in}}}$ we use the standard multi-index shorthand. A *multi-index* is a tuple $\alpha = (\alpha_1, \dots, \alpha_{d_{\text{in}}}) \in \mathbb{N}_0^{d_{\text{in}}}$ of non-negative integers; its *order* is $|\alpha| := \alpha_1 + \cdots + \alpha_{d_{\text{in}}}$, its *factorial* is $\alpha! := \alpha_1!\,\alpha_2! \cdots \alpha_{d_{\text{in}}}!$, the *monomial* of $x \in \mathbb{R}^{d_{\text{in}}}$ is $x^\alpha := x_1^{\alpha_1} \cdots x_{d_{\text{in}}}^{\alpha_{d_{\text{in}}}}$, and the *partial derivative* of order α is

$$\partial_x^\alpha := \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_{d_{\text{in}}}^{\alpha_{d_{\text{in}}}}}. \tag{30}$$

The Taylor polynomial of degree k of a smooth function $h : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}$ at the origin is $T_k h(x) := \sum_{|\alpha|\le k} (x^\alpha/\alpha!)\, \partial_x^\alpha h(0)$. The number of multi-indices with $|\alpha| \le k$ in dimension d_{in} is

$$D_k \;:=\; \#\{\alpha \in \mathbb{N}_0^{d_{\text{in}}} : |\alpha| \le k\} \;=\; \binom{d_{\text{in}} + k}{d_{\text{in}}}, \tag{31}$$

which is the dimension of the space of polynomials of degree at most k in d_{in} variables and the size of the basis we match.

Norms. For a vector $u \in \mathbb{R}^d$, $\|u\|$ denotes the Euclidean (ℓ_2) norm, $\|u\|_1 = \sum_i |u_i|$ the ℓ_1 norm, and $\|u\|_\infty = \max_i |u_i|$ the sup-norm. For a matrix A, $\|A\|_{\text{op}}$ is the spectral (operator-2) norm and $\|A\|_{\text{op},\infty}$ is the $\ell_\infty \to \ell_\infty$ induced norm (the maximum absolute row sum).

Matrix and submatrix notation. We use Python/NumPy-style slicing throughout. For a matrix $A \in \mathbb{R}^{m\times n}$, $A_{i,:}$ is its i-th row (as a vector), $A_{:,j}$ is its j-th column, and for index sets $I \subseteq [m]$, $J \subseteq [n]$, $A[I, J]$ is the submatrix with rows I and columns J. We write $A^\top$ for the transpose and A^+ for the Moore–Penrose pseudoinverse. The image (column space) of A is $\mathrm{Im}(A)$. The standard basis vector with a 1 in coordinate r and zeros elsewhere is e_r.

Complex-analytic objects. The open complex ball of radius r centered at the origin in $\mathbb{C}^d$ is $B_{\mathbb{C}}(0, r) := \{z \in \mathbb{C}^d : \|z\| < r\}$, and $\overline{B_{\mathbb{C}}(0, r)}$ is its closure. For $z \in \mathbb{C}$ and a set $S \subseteq \mathbb{C}$, $\mathrm{dist}(z, S) := \inf_{w\in S} |z - w|$. The Jacobian of f at z is denoted $Jf(z)$. Several holomorphic radii play distinct roles and are easy to confuse:

- ρ_f: holomorphy radius of f in the input variable (Assumption 1);
- ρ_σ: distance from the activation's input range to the boundary of σ's holomorphic domain $\mathcal{D}_\sigma$ (Eq. 5);
- ρ: the *effective* holomorphic radius of the composed map $g(\cdot, w)$, defined in Eq. (8);
- ρ_0: a working radius chosen in (R, ρ), used to apply Cauchy's inequality.

The hierarchy $R < \rho_0 < \rho \le \rho_f$ is maintained throughout (Assumption 4).

Measure-theoretic notation. The Dirac function at (v, w) is $\delta_{(v,w)}$.

B Formal theory on single-layer compression

B.1 Derivative matching controls the output

The goal of this subsection is to show that if the compressed measure $\tilde\mu$ matches all x-derivatives of $\langle v\, g(\cdot, w)\rangle_\mu$ at 0 up to order k, then $\langle v\, g(x, w)\rangle_{\tilde\mu}$ approximates $\langle v\, g(x, w)\rangle_\mu$ uniformly on Ω, with error decaying geometrically in k.

Lemma 9 (Sufficient holomorphic radius for $g(\cdot, w)$)**.** *Under Assumptions 1–4, for every $w \in \mathbb{R}^{d_{\ell-1}}$ with $\|w\| \le W_{\max}^{(\ell)}$ and every $r \in (0, \rho)$, the map $z \mapsto g(z, w)$ extends holomorphically to $B_{\mathbb{C}}(0, r)$. In particular, $g(\cdot, w)$ is holomorphic on an open neighborhood of $\overline{B_{\mathbb{C}}(0, \rho_0)}$.*

Proof. Fix $r \in (0, \rho)$. Since L_f is nondecreasing on $(0, \rho_f)$, the map $r' \mapsto r' W_{\max}^{(\ell)} L_f(r')$ is nondecreasing. The set in Eq. (8) is therefore down-closed; it is nonempty because $r' W_{\max}^{(\ell)} L_f(r') \to 0$ as $r' \to 0^+$ while $\rho_\sigma > 0$, so ρ is well defined, and down-closedness gives that the set contains all of $(0, \rho)$. In particular, $r W_{\max}^{(\ell)} L_f(r) < \rho_\sigma$. For $z \in B_{\mathbb{C}}(0, r)$, holomorphy of f and the fundamental theorem of calculus along $t \mapsto tz$ give

$$\begin{aligned} f(z) - f(0) &= \int_0^1 Jf(tz)\, z\, dt, \\ \|f(z) - f(0)\| &\le r L_f(r). \end{aligned} \tag{32}$$

Therefore

$$\begin{aligned} |w^\top f(z) - w^\top f(0)| &\le \|w\| \, \|f(z) - f(0)\| \\ &\le W_{\max}^{(\ell)} r L_f(r) < \rho_\sigma. \end{aligned} \tag{33}$$

Thus $w^\top f(z)$ lies in the disk on which σ is holomorphic. The composition $\sigma(w^\top f(z))$ is holomorphic on $B_{\mathbb{C}}(0, r)$. Since $\rho_0 < \rho$, choosing $r_1 \in (\rho_0, \rho)$ gives holomorphy near $\overline{B_{\mathbb{C}}(0, \rho_0)}$. □

Proof of Theorem 5. Fix j and $x \in \Omega$, set $s = \|x\| \le R$, and assume $s > 0$; the case $s = 0$ follows from Eq. (15) with $|\alpha| = 0$. Choose $r_1 \in (\rho_0, \rho)$. For fixed w, define $\varphi_w(t) = g(tx, w)$ for $|t| < r_1/s$. Lemma 9 makes φ_w holomorphic on a disk containing $|t| \le \rho_0/s$. Its Taylor expansion along the complex line tx is

$$\varphi_w(t) = \sum_{n=0}^{\infty} a_n(w) t^n, \qquad a_n(w) = \sum_{|\alpha|=n} \frac{x^\alpha}{\alpha!} \partial_x^\alpha g(0, w). \tag{34}$$

Indeed, by the chain rule, $a_n(w) = \frac{1}{n!} \frac{d^n}{dt^n} g(tx, w)\big|_{t=0} = \frac{1}{n!} \big[(x \cdot \nabla_x)^n g\big](0, w)$, and expanding $(x \cdot \nabla_x)^n$ by the multinomial theorem gives Eq. (34); grouping by $|\alpha| = n$ likewise gives $\sum_{n=0}^{k} a_n(w) = T_k g(x, w)$, where $T_k g(x, w) \coloneqq \sum_{|\alpha| \le k} x^\alpha \partial_x^\alpha g(0, w)/\alpha!$. On $|t| = \rho_0/s$, $|g(tx, w)| \le M_g$, so Cauchy's inequality gives

$$|a_n(w)| \le M_g (s/\rho_0)^n. \tag{35}$$

Since $s < \rho_0$,

$$|g(x, w) - T_k g(x, w)| \le M_g \sum_{n \ge k+1} (s/\rho_0)^n \le M_g \frac{\rho_0}{\rho_0 - R} \left(\frac{R}{\rho_0}\right)^{k+1} =: E_k. \tag{36}$$

Summing the tail bound over neurons yields

$$\begin{aligned} |y_j(x) - T_k y_j(x)| &\le \|v_{:,j}\|_1 E_k, \\ |\tilde{y}_j(x) - T_k \tilde{y}_j(x)| &\le \|\tilde{v}_{:,j}\|_1 E_k. \end{aligned} \tag{37}$$

Equation (15) implies $T_k y_j(x) = T_k \tilde{y}_j(x)$ after multiplying each matched derivative by $x^\alpha/\alpha!$ and summing over $|\alpha| \le k$. The triangle inequality gives Eq. (16). □

B.2 Existence of compressed weights

This subsection shows that, starting from the matrix equation $V \Phi(W^{(\ell)})^\top = W^{(\ell+1)} \Phi(W^{(\ell)})^\top$, there exists a solution V with at most D_k nonzero columns, which gives the compressed weights. The key is to apply a rank-reduction argument to the derivative-feature matrix $\Phi(W^{(\ell)})$.

Lemma 10 (Rank reduction)**.** *Let $Z = (z_1 | \cdots | z_m) \in \mathbb{R}^{D \times m}$ have rank n. Then there are distinct indices $\kappa_1, \ldots, \kappa_n \in [m]$ and a matrix $C \in \mathbb{R}^{n \times m}$ such that, with $\mathcal{K} = \{\kappa_1, \ldots, \kappa_n\}$ and $Z_{\mathcal{K}} = (z_{\kappa_1} | \cdots | z_{\kappa_n})$,*

$$Z = Z_{\mathcal{K}} C, \qquad C_{:,\kappa_r} = e_r \quad (r \in [n]), \tag{38}$$

and

$$\max_i \|C_{:,i}\|_1 \le n. \tag{39}$$

If $n = 0$, this means $\mathcal{K} = \varnothing$, $C \in \mathbb{R}^{0 \times m}$, and $Z = 0$.

Proof. The case $n = 0$ is immediate. For $n \ge 1$, choose row and column sets $I \subseteq [D]$ and $\mathcal{K} = \{\kappa_1, \ldots, \kappa_n\} \subseteq [m]$ with $|I| = |\mathcal{K}| = n$ such that $A \coloneqq Z[I, \mathcal{K}]$ is nonsingular and maximizes $|\det A|$ among all nonsingular $n \times n$ submatrices. Then $Z_{\mathcal{K}}$ has full column rank and $\mathrm{Im}(Z_{\mathcal{K}}) = \mathrm{Im}(Z)$. Set $C \coloneqq Z_{\mathcal{K}}^+ Z$. Since $Z_{\mathcal{K}} Z_{\mathcal{K}}^+$ is the projector onto $\mathrm{Im}(Z)$, $Z = Z_{\mathcal{K}} C$, and selected columns satisfy $C_{:,\kappa_r} = e_r$.

For $i \notin \mathcal{K}$, the vector $c_i = C_{:,i}$ is the unique solution of $Z_{\mathcal{K}} c_i = z_i$. Restricting to rows I gives $A c_i = Z[I, i]$. By Cramer's rule,

$$(c_i)_r = \frac{\det(A^{(r \leftarrow i)})}{\det A}, \tag{40}$$

where $A^{(r \leftarrow i)}$ replaces the r-th column of A by $Z[I, i]$. The maximum-volume choice gives $|\det(A^{(r \leftarrow i)})| \le |\det A|$, so $|(c_i)_r| \le 1$ for every r and $\|c_i\|_1 \le n$. For selected columns the bound is $\|e_r\|_1 = 1 \le n$. □

Proposition 11 (Existence of compressed weights)**.** *There exist $\mathcal{K} = \{\kappa_1, \dots, \kappa_n\} \subseteq [d_\ell]$ with $n \le D_k$ and a matrix $V \in \mathbb{R}^{d_{\ell+1} \times d_\ell}$ whose columns outside $\mathcal{K}$ are zero such that*

$$V\Phi(W^{(\ell)})^\top = W^{(\ell+1)}\Phi(W^{(\ell)})^\top. \tag{41}$$

Equivalently, the compressed measure $\tilde{\mu} = \sum_{r=1}^{n} \delta_{(\tilde{v}_{\kappa_r}, w_{\kappa_r})}$ with $\tilde{v}_{\kappa_r} \coloneqq V_{:,\kappa_r}$ satisfies the derivative-matching condition Eq. (15)*.*

Proof. Apply Lemma 10 to $Z = [\Phi(w_1)| \cdots |\Phi(w_{d_\ell})]$. For each output row j, let $a^{(j)} \in \mathbb{R}^{d_\ell}$ be the j-th row of $W^{(\ell+1)}$ as a column vector and set $\tilde{a}^{(j)} = Ca^{(j)} \in \mathbb{R}^n$. Define $V_{j,\kappa_r} = \tilde{a}_r^{(j)}$ and $V_{j,i} = 0$ for $i \notin \mathcal{K}$. Then the j-th row of $V\Phi(W^{(\ell)})^\top$ equals $(Z_{\mathcal{K}} Ca^{(j)})^\top = (Za^{(j)})^\top$, which is the j-th row of $W^{(\ell+1)}\Phi(W^{(\ell)})^\top$. Finally, Eq. (41) coincides entrywise with Eq. (15): for each $j \in [d_{\ell+1}]$ and each $|\alpha| \le k$, the corresponding entries of the two sides are $\sum_{i=1}^{d_\ell} V_{ji}\, \partial_x^\alpha g(0, w_i) = \big(\langle v\, \partial_x^\alpha g(0,w)\rangle_{\tilde{\mu}}\big)_j$ and $\sum_{i=1}^{d_\ell} W_{ji}^{(\ell+1)}\, \partial_x^\alpha g(0, w_i) = \big(\langle v\, \partial_x^\alpha g(0,w)\rangle_{\mu}\big)_j$. □

Corollary 12 (Norm inflation due to compression)**.** *With $\mathcal{K}, V$ as in Proposition 11, define $\widetilde{W}_{r,:}^{(\ell)} = w_{\kappa_r}^\top$ and $\widetilde{W}_{:,r}^{(\ell+1)} = V_{:,\kappa_r}$. Then the compressed network has width $n \le D_k$ and*

$$\|\widetilde{W}_{j,:}^{(\ell+1)}\|_1 \le nV_{\max} \le D_k V_{\max} \qquad (j \in [d_{\ell+1}]). \tag{42}$$

Proof. The width bound is Proposition 11. For the norm bound, $\widetilde{W}_{j,:}^{(\ell+1)} = Ca^{(j)}$, so

$$\begin{aligned} \|Ca^{(j)}\|_1 &\le \sum_{i=1}^{d_\ell} |a_i^{(j)}|\, \|C_{:,i}\|_1 \\ &\le n\|a^{(j)}\|_1 \le nV_{\max}. \end{aligned} \tag{43}$$

□

B.3 Rank reduction algorithm

Algorithm 1 below turns the proof of Proposition 11 into an explicit procedure.

Algorithm 1 Intermediate-layer compression via rank reduction

Require: Weight matrix $W^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ with rows $w_i^\top$; weight matrix $W^{(\ell+1)} \in \mathbb{R}^{d_{\ell+1} \times d_\ell}$; derivative-matching order $k \in \mathbb{N}_0$; derivative feature map $\Phi : \mathbb{R}^{d_{\ell-1}} \to \mathbb{R}^D$ of Eq. (18) with $D = \binom{d_{\text{in}}+k}{d_{\text{in}}}$.

Ensure: $\mathcal{K} = \{\kappa_1, \dots, \kappa_n\} \subseteq [d_\ell]$ with $n \le D$, matrices $\widetilde{W}^{(\ell)} \in \mathbb{R}^{n \times d_{\ell-1}}$ and $\widetilde{W}^{(\ell+1)} \in \mathbb{R}^{d_{\ell+1} \times n}$ such that Eq. (41) holds and Eq. (42) holds.

$Z \leftarrow \big[\, \Phi(w_1) \;\big|\; \Phi(w_2) \;\big|\; \cdots \;\big|\; \Phi(w_{d_\ell}) \,\big] \in \mathbb{R}^{D \times d_\ell}$
$n \leftarrow \operatorname{rank}(Z)$ ▷ e.g. via SVD or rank-revealing QR
if $n = 0$ **then**
  return $\mathcal{K} \leftarrow \varnothing$, $\widetilde{W}^{(\ell)} \leftarrow \mathbf{0}_{0 \times d_{\ell-1}}$, $\widetilde{W}^{(\ell+1)} \leftarrow \mathbf{0}_{d_{\ell+1} \times 0}$
end if
$(I^\star, \mathcal{K}^\star) \leftarrow \arg\max_{\substack{I \subseteq [D],\, \mathcal{K} \subseteq [d_\ell] \\ |I| = |\mathcal{K}| = n}} \big|\det Z[I, \mathcal{K}]\big|$ ▷ maximum-volume $n \times n$ submatrix
$\mathcal{K} \leftarrow \mathcal{K}^\star = \{\kappa_1, \dots, \kappa_n\}$
$Z_{\mathcal{K}} \leftarrow \big[\, \Phi(w_{\kappa_1}) \;\big|\; \cdots \;\big|\; \Phi(w_{\kappa_n}) \,\big] \in \mathbb{R}^{D \times n}$
$C \leftarrow Z_{\mathcal{K}}^{+}\, Z \in \mathbb{R}^{n \times d_\ell}$ ▷ e.g. via thin QR of $Z_{\mathcal{K}}$
$\widetilde{W}^{(\ell+1)} \leftarrow W^{(\ell+1)}\, C^\top \in \mathbb{R}^{d_{\ell+1} \times n}$
$\widetilde{W}^{(\ell)} \leftarrow \big[\, w_{\kappa_1} \;\big|\; \cdots \;\big|\; w_{\kappa_n} \,\big]^\top \in \mathbb{R}^{n \times d_{\ell-1}}$ ▷ rows of $W^{(\ell)}$ indexed by $\mathcal{K}$
return $\mathcal{K}$, $\widetilde{W}^{(\ell)}$, $\widetilde{W}^{(\ell+1)}$

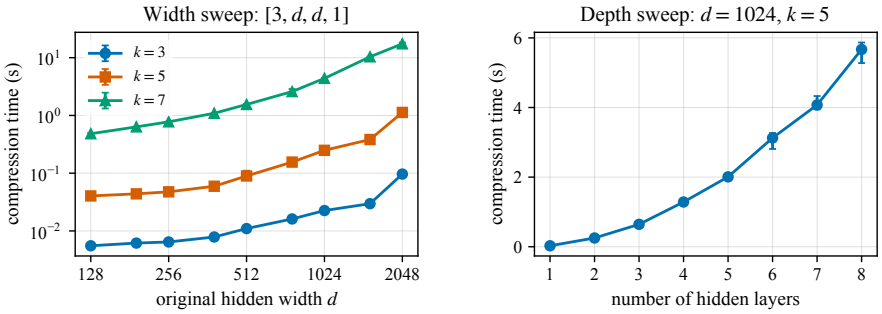

Figure 3: End-to-end runtime of backward compression with the practical pivoted-QR implementation. The timed region includes derivative-feature evaluation, basis selection, the coefficient solve, weight reconstruction for every hidden layer, and the implementation's numerical diagnostics; teacher construction and a small library warm-up are excluded. We use randomly initialized float64 sigmoid networks on CPU with one Torch/BLAS thread. Points are medians of three complete compressions and bars span their minima and maxima. Left: teachers $[3, d, d, 1]$ for $k = 3, 5, 7$. Right: teachers $[3, 1024, \ldots, 1024, 1]$ with one through eight hidden layers at $k = 5$.

Proposition 13 (Correctness of Algorithm 1)**.** *The output of Algorithm 1 satisfies Eq.* (41)*. In addition,*

$$\max_{j\in[d_{\ell+1}]} \|\widetilde{W}^{(\ell+1)}_{j,:}\|_1 \le nV_{\max} \le D_k V_{\max}. \tag{44}$$

Proof. For $n = 0$, $Z = 0$ and both sides of Eq. (41) vanish. For $n \ge 1$, Algorithm 1 implements the maximum-volume construction in Lemma 10 and the weight definition in Proposition 11; thus Eq. (41) holds. The ℓ_1 bound is exactly Corollary 12. □

Once the derivative-feature matrix Z is available, the arithmetic cost of Algorithm 1 simplifies considerably in the rank-saturated case $n = D < d_\ell$. It is

$$\underbrace{O\left(\binom{d_\ell}{D} D^3\right)}_{\text{maximum-volume search}} + \underbrace{O(d_{\ell+1} d_\ell D)}_{\text{outgoing-weight update}}. \tag{45}$$

Indeed, the rank-factorization and pseudoinverse work is $O(D^2 d_\ell)$ and is absorbed by the exhaustive-search term under $D < d_\ell$. The maximum-volume search is combinatorial and infeasible even for modest D and d_ℓ. It is needed here to obtain the uniform coefficient bound in Eq. (39); practical implementations can instead use polynomial-time rank-revealing column selection. Finding the rank in advance is also unnecessary when one directly targets the upper bound D.

The cost of forming Z by evaluating the required derivatives is harder to characterize: it depends on the input-to-layer subnetwork, activation, differentiation method, and implementation-level reuse. We therefore do not assign it an implementation-independent arithmetic bound. Figure 3 instead reports the measured wall-clock scaling of the complete practical compression pipeline used in our experiments.

B.4 The main theorem for single-layer compression

Proof of Theorem 6. By Proposition 11 and Corollary 12, there are compressed matrices with $n \le D_k$ satisfying Eq. (15) and $\max_j \|\widetilde{W}^{(\ell+1)}_{j,:}\|_1 \le D_k V_{\max}$. Theorem 5 gives

$$\begin{aligned} |y_j(x) - \tilde{y}_j(x)| &\le (1 + D_k) C_0 e^{-(k+1)\lambda} \\ &\le 2 D_k C_0 e^{-(k+1)\lambda}, \end{aligned} \tag{46}$$

because $D_k \geq 1$. Since

$$D_k = \binom{d_{\text{in}} + k}{d_{\text{in}}} = \frac{\prod_{r=1}^{d_{\text{in}}}(k+r)}{d_{\text{in}}!} \leq \frac{(k+d_{\text{in}})^{d_{\text{in}}}}{d_{\text{in}}!}, \tag{47}$$

Eq. (21) implies the right-hand side is at most ε_ℓ, proving Eq. (22).

For the asymptotics, put $A = \log(2C_0/\varepsilon_\ell)$ and $h(k) = (k+1)\lambda - d_{\text{in}} \log(k + d_{\text{in}}) + \log(d_{\text{in}}!) - A$. The function h is eventually increasing and tends to infinity, so the minimal feasible $k^\star$ exists. The feasibility inequality implies $k^\star \geq A/\lambda - O(\log A)$, while choosing $k = \lceil A/\lambda \rceil + \lceil d_{\text{in}} \log(A/\lambda + d_{\text{in}})/\lambda \rceil$ is feasible for large A. Hence $k^\star = A/\lambda + O(\log A)$, which is Eq. (24). Substituting this into $D_k \leq (k + d_{\text{in}})^{d_{\text{in}}}/d_{\text{in}}!$ gives Eq. (25). □

C MULTI-LAYER COMPOSITION

Fix a permutation $(\ell_1, \ldots, \ell_{L-1})$ of $\{1, \ldots, L-1\}$ (the *compression order*) and per-step budgets $\varepsilon_1, \ldots, \varepsilon_{L-1} > 0$. Write $F = F^{(0)}$ for the original network; let $F^{(t)}$ be obtained from $F^{(t-1)}$ by applying Algorithm 1 to layer ℓ_t with matching order $k_t := k^\star(\varepsilon_t)$ computed from the constants of $F^{(t-1)}$ via Theorem 6, and $\widetilde{F} := F^{(L-1)}$. We write $y^{(\ell+1)} := W^{(\ell+1)} h^{(\ell)}$ for the pre-activation entering layer $\ell + 1$ (so $y^{(L)} = y_{\text{out}}$, and the downstream map of layer $L-1$ is the identity). Each constant $X^{(\ell)}$ from Section 2 carries an extra superscript $X^{(\ell,t)}$ when evaluated for layer ℓ in $F^{(t)}$ (so $W_{\max}^{(\ell,t)}$, $V_{\max}^{(\ell+1,t)}$, $\rho^{(\ell,t)}$, $C_0^{(\ell,t)}$, $\lambda^{(\ell,t)}$, where the upper index of $V_{\max}$ is its *matrix* index $\ell + 1$ and the others are indexed by the layer being compressed). Throughout this appendix we treat the depth L as a fixed constant; suppressed constants in the deep theorem are allowed to depend on L.

Assumption 14 (Globally Lipschitz σ)**.** *σ is globally L_σ-Lipschitz on $\mathbb{R}$ for some finite L_σ.*

This is genuinely stronger than Assumption 1: it excludes entire activations such as $\exp$, which are admissible in Sections 2–5 via $\rho_\sigma = \infty$. Throughout this section we add Assumption 14 to the standing assumptions.

Lemma 15 (Downstream Lipschitz constant)**.** *For any $\ell \in \{1, \ldots, L-1\}$ and $t \geq 0$, the Lipschitz constant $\Lambda_{>\ell}^{(t)}$ of the map $y^{(\ell+1)} \mapsto y_{\text{out}}$ in $F^{(t)}$, in the ℓ_∞-to-ℓ_∞ sense, satisfies*

$$\Lambda_{>\ell}^{(t)} \leq L_\sigma^{L-\ell-1} \prod_{j=\ell+2}^{L} \left\| W^{(j,t)} \right\|_{\text{op},\infty}. \tag{48}$$

Here $\|A\|_{\text{op},\infty} := \sup_{u \neq 0} \|Au\|_\infty / \|u\|_\infty = \max_i \sum_j |A_{ij}|$ is the $\ell_\infty \to \ell_\infty$ induced (operator) norm, i.e. the maximum absolute row sum. In particular $\Lambda_{>L-1}^{(t)} = 1$, the downstream map of the last hidden layer being the identity.

Proof. The downstream map alternates $L - \ell - 1$ applications of the elementwise L_σ-Lipschitz activation with the linear maps $W^{(j,t)}$ for $j = \ell + 2, \ldots, L$. The induced ℓ_∞ operator norm of each matrix is $\|W^{(j,t)}\|_{\text{op},\infty}$, so composition yields Eq. (48). For $\ell = L - 1$ both the activation count and the product are empty, giving 1. □

Theorem 16 (Total error of sequential compression)**.** *Suppose Theorem 6 applies at every step t with the constants of $F^{(t-1)}$. Then*

$$\sup_{x \in \Omega} \left\| \widetilde{F}(x) - F(x) \right\|_\infty \leq \sum_{t=1}^{L-1} \Lambda_{>\ell_t}^{(t-1)} \, \varepsilon_t. \tag{49}$$

Proof. Telescope $\widetilde{F} - F = \sum_t \big(F^{(t)} - F^{(t-1)} \big)$. The upstream map of layer ℓ_t is identical in $F^{(t-1)}$ and $F^{(t)}$, so the two networks feed the same pre-activation input to layer ℓ_t, and Theorem 6 bounds the resulting perturbation at the pre-activation $y^{(\ell_t+1)}$ by ε_t uniformly on Ω. The layers downstream of $y^{(\ell_t+1)}$ coincide in $F^{(t-1)}$ and $F^{(t)}$ (so $\Lambda_{>\ell_t}^{(t-1)} = \Lambda_{>\ell_t}^{(t)}$), and Lemma 15 propagates the perturbation by $\Lambda_{>\ell_t}^{(t-1)}$. Summing gives Eq. (49). □

C.1 FORWARD VS. BACKWARD ORDER

Proposition 17 (Backward order preserves the per-step constants)**.** *Let $\ell_t = L - t$. For every t,*

$$W_{\max}^{(\ell_t,t-1)} = W_{\max}^{(\ell_t,0)}, \qquad V_{\max}^{(\ell_t+1,t-1)} \le V_{\max}^{(\ell_t+1,0)}, \qquad \rho^{(\ell_t,t-1)} = \rho^{(\ell_t,0)}. \tag{50}$$

Consequently, for a fixed working radius $\rho_0^{(\ell_t)} \in \big(R, \rho^{(\ell_t,0)}\big)$, one has $C_0^{(\ell_t,t-1)} \le C_0^{(\ell_t,0)}$ and $\lambda^{(\ell_t,t-1)} = \lambda^{(\ell_t,0)}$, and the predicted compressed width is no larger than the original-network prediction.

Proof. At step t, previous backward steps only modified layers with index greater than ℓ_t. Thus $W^{(\ell_t)}$ and the upstream map f entering layer ℓ_t are untouched, giving equality of $W_{\max}$, L_f, ρ_σ, and hence of ρ; since $\rho_0^{(\ell_t)}$ is held fixed, $\lambda = \log\big(\rho_0^{(\ell_t)}/R\big)$ is unchanged. The matrix $W^{(\ell_t+1)}$ may have been row-subsetted by the immediately previous step, so its row ℓ_1-norms cannot increase; this gives the bound on $V_{\max}^{(\ell_t+1,\cdot)}$. The claim for C_0 follows from Eq. (20). □

Proposition 18 (Forward order can inflate incoming weights)**.** *Let $\ell_t = t$, and let $k_{t-1}^\star$ be the matching order used at step $t-1$. For every $t \ge 2$,*

$$W_{\max}^{(t,t-1)} \;\le\; D_{k_{t-1}^\star}\, V_{\max}^{(t,0)}, \qquad D_{k_{t-1}^\star} = \binom{d_{\text{in}} + k_{t-1}^\star}{d_{\text{in}}}. \tag{51}$$

The right side depends recursively on previous compression widths and therefore on previous analytic constants.

Proof. At step $t-1$, Algorithm 1 replaces $W^{(t)}$ by $W^{(t,t-2)}C_{t-1}^\top$. Corollary 12 bounds the maximum row ℓ_1-norm of the result by $D_{k_{t-1}^\star}\, V_{\max}^{(t,t-2)}$. Earlier forward steps do not touch $W^{(t)}$, so $V_{\max}^{(t,t-2)} = V_{\max}^{(t,0)}$. Since $\|u\|_2 \le \|u\|_1$, Eq. (51) follows. □

Two further obstacles distinguish forward order from backward order: (i) the partial map f entering layer ℓ_t is the composition of *already-compressed* upstream layers, so the analytic constants $\rho_f^{(\ell_t,t-1)}$, $L_f^{(\ell_t,t-1)}$ entering Eq. (8) differ in general from their original values and must be re-verified; (ii) the inflated $W_{\max}^{(t,t-1)}$ can shrink $\rho^{(t,t-1)}$ via Eq. (8), hence increase $C_0^{(t,t-1)}$ and $D_{k_t^\star}$. We do not solve this recursion. Heuristically, since $W_{\max}$ enters ρ only through a logarithm, each $D_{k_t^\star}$ is expected to remain polylogarithmic in $1/\varepsilon$, but with constants strictly larger than the original-network values used in Proposition 17. The remainder of this section therefore works exclusively in backward order.

C.2 BACKWARD COMPRESSION ERROR BOUND

An a priori downstream bound. In backward order the downstream matrices have *already* been compressed when layer ℓ_t is processed, so the realized downstream constant

$$\Lambda^\star \;\coloneqq\; \max_{1\le t\le L-1} \Lambda_{>\ell_t}^{(t-1)} \;\in\; [1,\infty) \tag{52}$$

depends on the very widths produced by the budgets, which in turn depend on $\Lambda^\star$ if it is used to set the budgets.

Fix, once and for all, per-layer working radii $\rho_0^{(\ell)} \in \big(R, \rho^{(\ell,0)}\big)$ and set $\lambda^{(\ell,0)} = \log\big(\rho_0^{(\ell)}/R\big)$. With

$$C^\star \;\coloneqq\; \max_{1\le \ell\le L-1} C_0^{(\ell,0)}, \qquad \lambda^\star \;\coloneqq\; \min_{1\le \ell\le L-1} \lambda^{(\ell,0)}, \tag{53}$$

For $C^\star > 0$, define the *surrogate width* for a budget $s > 0$ by

$$\begin{aligned} \overline{D}(s) \;&\coloneqq\; \binom{d_{\text{in}} + k^\star(s)}{d_{\text{in}}}, \\ k^\star(s) \;&\coloneqq\; \min\Big\{ k \in \mathbb{N}_0 : (k+1)\lambda^\star \ge \log(2C^\star/s) \\ &\qquad\qquad + d_{\text{in}} \log(k + d_{\text{in}}) - \log(d_{\text{in}}!) \Big\}. \end{aligned} \tag{54}$$

Because the feasibility condition Eq. (21) is monotone in (C_0, λ) and $C^\star \geq C_0^{(\ell,t)}$, $\lambda^\star \leq \lambda^{(\ell,t)}$ for every layer and step (Proposition 17), $\overline{D}(s)$ upper-bounds the compressed width produced at *any* step whose budget is s. By Eqs. (24)–(25),

$$\overline{D}(s) \;=\; O\!\left(\frac{1}{d_{\mathrm{in}}!}\left(\frac{\log(2C^\star/s)}{\lambda^\star}\right)^{d_{\mathrm{in}}}\right) \qquad (s \to 0^+). \tag{55}$$

Write $L_\sigma^+ \coloneqq \max\{1, L_\sigma\}$ and $\Lambda_0^\star \coloneqq \prod_{j=2}^{L} \max\{1, \|W^{(j,0)}\|_{\mathrm{op},\infty}\}$, and define

$$\psi(\Lambda) \;\coloneqq\; \big(L_\sigma^+\big)^{L-2}\, \Lambda_0^\star\, \overline{D}\!\left(\frac{\varepsilon}{(L-1)\,\Lambda}\right)^{L-2}. \tag{56}$$

Lemma 19 (A priori downstream bound)**.** *For every finite L there exists $\overline{\Lambda} \geq 1$ with $\psi(\overline{\Lambda}) \leq \overline{\Lambda}$. Taking $\overline{\Lambda}$ to be the least such value, one has, as $\varepsilon \to 0^+$ with L and d_{in} fixed,*

$$\log \overline{\Lambda} \;=\; O\big(d_{\mathrm{in}}\, L\, \log\log(1/\varepsilon)\big). \tag{57}$$

Proof. By Eq. (55), for fixed ε and $\Lambda \to \infty$, $\overline{D}\big(\varepsilon/((L-1)\Lambda)\big) = O\big((\log\Lambda)^{d_{\mathrm{in}}}\big)$, so $\psi(\Lambda) = O\big((\log\Lambda)^{(L-2)d_{\mathrm{in}}}\big) = o(\Lambda)$. Hence $\psi(\Lambda)/\Lambda \to 0$, and there is $\overline{\Lambda}$ with $\psi(\overline{\Lambda}) \leq \overline{\Lambda}$. (For $L = 2$ the exponent is 0 and $\psi \equiv \Lambda_0^\star$, so $\overline{\Lambda} = \max\{1, \Lambda_0^\star\}$ works.)

For the rate, it remains to consider $L \geq 3$. Put

$$A = \log\!\left(\frac{2C^\star(L-1)}{\varepsilon}\right), \qquad m = (L-2)d_{\mathrm{in}}. \tag{58}$$

By Eq. (55), there is a constant $B = B\big(L, \Lambda_0^\star, L_\sigma, \lambda^\star, C^\star\big)$, independent of ε and u, such that, for all sufficiently small ε and every $u \geq 0$,

$$\log \psi(e^u) \;\leq\; B + m\log(A+u). \tag{59}$$

Choose a constant $B_+ \geq \max\{B, 0\}$ and set $u_A = B_+ + m\log(2A)$. Since $u_A \leq A$ for all sufficiently large A, Eq. (59) gives

$$\log \psi(e^{u_A}) \;\leq\; B + m\log(A+u_A) \;\leq\; B + m\log(2A) \;\leq\; u_A. \tag{60}$$

Thus e^{u_A} is feasible, and the least feasible value satisfies

$$\log \overline{\Lambda} \;\leq\; u_A \;=\; B_+ + (L-2)d_{\mathrm{in}}\log(2A) \;=\; O\big(d_{\mathrm{in}} L \log\log(1/\varepsilon)\big), \tag{61}$$

which proves Eq. (57). □

Feasibility and width. We now show that the uniform allocation against $\overline{\Lambda}$ is feasible and yields polylogarithmic width.

Proposition 20 (Feasibility and width of the uniform allocation)**.** *Fix $\overline{\Lambda}$ as in Lemma 19 and set*

$$\varepsilon_t \;=\; \frac{\varepsilon}{(L-1)\,\overline{\Lambda}} \qquad (t = 1, \ldots, L-1). \tag{62}$$

Compressing in backward order $\ell_t = L - t$ with these budgets, every realized downstream constant obeys $\Lambda_{>\ell_t}^{(t-1)} \leq \overline{\Lambda}$. Consequently the allocation is feasible,

$$\sum_{t=1}^{L-1} \Lambda_{>\ell_t}^{(t-1)}\, \varepsilon_t \;\leq\; \varepsilon, \tag{63}$$

and the retained width at step t satisfies

$$\begin{aligned} d'_{\ell_t} &\leq D_{k_t^\star} \leq \overline{D}(\varepsilon_t), \\ \overline{D}(\varepsilon_t) &= O\left(\frac{1}{d_{\mathrm{in}}!}\left(\frac{\log\big(2(L-1)\overline{\Lambda}C^\star/\varepsilon\big)}{\lambda^\star}\right)^{d_{\mathrm{in}}}\right). \end{aligned} \tag{64}$$

Proof. Because the budgets ε_t are fixed in advance, the width produced at step t depends only on ε_t and on the realized constants $C_0^{(\ell_t,t-1)}, \lambda^{(\ell_t,t-1)}$, which by Proposition 17 (with the fixed choice $\rho_0^{(\ell_t)}$) satisfy $C_0^{(\ell_t,t-1)} \le C^\star$ and $\lambda^{(\ell_t,t-1)} \ge \lambda^\star$. Hence $D_{k_t^\star} \le \overline{D}(\varepsilon_t)$ for every t; Corollary 12 also gives $d'_{\ell_t} \le D_{k_t^\star}$, and the asymptotic in Eq. (64) is Eq. (55) at $s = \varepsilon_t$.

For the downstream constant, fix a step t and a downstream matrix index $j \in \{\ell_t + 2, \dots, L\}$. In backward order, layer j is compressed at step $L - j < t$ (row-subsetting $W^{(j)}$, which does not increase its maximum row ℓ_1-norm), and layer $j-1$ is compressed at step $L - j + 1 \le t - 1$ (reweighting $W^{(j)}$). Corollary 12 applied at that reweighting gives

$$\big\|W^{(j,t-1)}\big\|_{\mathrm{op},\infty} \;\le\; D_{k_{j-1}^\star}\,\big\|W^{(j,0)}\big\|_{\mathrm{op},\infty} \;\le\; \overline{D}(\varepsilon_t)\,\max\big\{1, \|W^{(j,0)}\|_{\mathrm{op},\infty}\big\}, \tag{65}$$

using $D_{k_{j-1}^\star} \le \overline{D}(\varepsilon_t)$ (the budget is uniform). Substituting into Lemma 15 and bounding $L_\sigma^{L-\ell_t-1} \le (L_\sigma^+)^{L-2}$ and $L - \ell_t - 1 \le L - 2$,

$$\Lambda_{>\ell_t}^{(t-1)} \;\le\; (L_\sigma^+)^{L-2}\,\Lambda_0^\star\,\overline{D}(\varepsilon_t)^{\,L-2} \;=\; \psi(\overline{\Lambda}) \;\le\; \overline{\Lambda}, \tag{66}$$

the last step by the defining inequality of $\overline{\Lambda}$. Then $\sum_{t=1}^{L-1} \Lambda_{>\ell_t}^{(t-1)} \varepsilon_t \le \overline{\Lambda}\,(L-1)\,\varepsilon_t = \varepsilon$, which is Eq. (63). □

Remark 21 (Robustness of the allocation)**.** Because $D_{k_t^\star}$ depends on ε_t only through $\log(1/\varepsilon_t)$, the width bound Eq. (64) is unchanged, up to a constant factor, for any allocation $\varepsilon_t = p_t\,\varepsilon/\overline{\Lambda}$ with $\sum_t p_t = 1$ and $p_t \in [c_1/L, c_2/L]$ for absolute $c_1, c_2 > 0$. One may, for instance, spend a slightly larger budget on layers with larger $C_0^{(\ell,0)}$ or smaller $\lambda^{(\ell,0)}$ without changing the asymptotic compressed width. This allocation-insensitivity (for fixed L) is precisely why optimizing the split is unnecessary; the uniform choice Eq. (62) is simply the most convenient feasible one.

The deep theorem.

Theorem 22 (Polylog deep-MLP compression)**.** *Fix one original teacher network $F^{(0)}$ for which Assumptions 1–4 hold at every hidden layer and Assumption 14 holds. Compress in backward order $\ell_t = L - t$ with the uniform allocation Eq.* (62)*, where $C^\star, \lambda^\star$ are the original-teacher constants in Eq.* (53) *and $\overline{\Lambda}$ is given by Lemma 19. Then the compressed network satisfies*

$$\sup_{x\in\Omega}\big\|\widetilde{F}(x) - F(x)\big\|_\infty \;\le\; \varepsilon, \tag{67}$$

and every retained hidden width obeys

$$\begin{aligned} d'_\ell &\le \overline{D}\!\left(\frac{\varepsilon}{(L-1)\overline{\Lambda}}\right), \\ \overline{D}\!\left(\frac{\varepsilon}{(L-1)\overline{\Lambda}}\right) &= O\!\left(\frac{1}{d_{\mathrm{in}}!}\left(\frac{\log\!\big(2(L-1)\overline{\Lambda}C^\star/\varepsilon\big)}{\lambda^\star}\right)^{d_{\mathrm{in}}}\right), \qquad \ell = 1, \dots, L-1. \end{aligned} \tag{68}$$

Since $\log\overline{\Lambda} = O(d_{\mathrm{in}} L \log\log(1/\varepsilon))$ by Lemma 19, the accuracy-driven cap is $O((\log(1/\varepsilon))^{d_{\mathrm{in}}})$ for this fixed teacher network.

Proof. Backward order leaves the upstream map of layer ℓ_t untouched, so Assumptions 1 and 4 continue to hold for $F^{(t-1)}$ at layer ℓ_t, and Theorem 6 applies at every step; Proposition 17 certifies that its constants are bounded by the original-network values, so $C^\star, \lambda^\star$ are valid uniform bounds. Proposition 20 provides the feasibility bound Eq. (63), so Theorem 16 yields Eq. (67); its retained-width clause Eq. (64) gives Eq. (68). The final estimate substitutes the rate of Lemma 19. □

Because the input and output layers are not compressed ($d'_0 = d_{\mathrm{in}}$, $d'_L = d_{\mathrm{out}}$), the total parameter count obeys

$$\sum_{\ell=1}^{L} d'_\ell\, d'_{\ell-1} \;=\; O\!\left(L\left(\max_{1\le\ell\le L-1} d'_\ell\right)^2 + (d_{\mathrm{in}} + d_{\mathrm{out}})\max_{1\le\ell\le L-1} d'_\ell\right), \tag{69}$$

For fixed $L, d_{\mathrm{in}}, d_{\mathrm{out}}$ and the fixed teacher's constants, substituting the accuracy-driven bounds yields a polylogarithmic upper bound in $1/\varepsilon$. Equation (69) records the explicit dependence on L.

D NON-VACUITY CONDITION

The compression theorem is valid for every covered teacher network, but it certifies a strict width reduction only when the retained dimension is smaller than the original width. This appendix makes that finite-width condition explicit. We write m for the admissible effective input dimension used by the derivative coordinates; the baseline case is $m = d_{\mathrm{in}}$. Define

$$D_k^{(m)} \;\coloneqq\; \binom{m+k}{m}. \tag{70}$$

D.1 MATCHING ORDER AND REALIZED RANK

Before the relaxation used in Eq. (21), the proof of Theorem 6 gives the error bound

$$\sup_{x\in\Omega} \|y(x) - \tilde{y}(x)\|_\infty \;\le\; (1 + D_k^{(m)}) C_0 e^{-(k+1)\lambda}. \tag{71}$$

Accordingly, define the sharper finite-scale certification order

$$k_{\mathrm{cert}} \;\coloneqq\; \min\left\{k \in \mathbb{N}_0 : (1 + D_k^{(m)}) C_0 e^{-(k+1)\lambda} \le \varepsilon_\ell\right\}. \tag{72}$$

The set is nonempty because the exponential term eventually dominates the polynomial $D_k^{(m)}$. The theorem-level $k^\star$ instead uses $2D_k^{(m)}$, followed by the upper bound $D_k^{(m)} \le (k+m)^m/m!$, to obtain the convenient explicit condition Eq. (21). Hence, $k_{\mathrm{cert}} \le k^\star$; the two have the same asymptotic order but differ at constant factor level.

Let

$$r_k \;\coloneqq\; \operatorname{rank} \Phi_k(W^{(\ell)}) \;\le\; \min\{D_k^{(m)}, d_\ell\} \tag{73}$$

be the realized derivative-feature rank. The rank-basis construction retains r_k neurons. Thus

$$D_{k_{\mathrm{cert}}}^{(m)} < d_\ell \tag{74}$$

is an a priori sufficient condition for strict compression. Condition (74) is not necessary: feature dependencies can make $r_k \ll D_k^{(m)}$. The compression error can remain safe when Eq. (74) fails; what becomes vacuous is only the guaranteed width reduction.

D.2 A CLOSED-FORM SUFFICIENT THRESHOLD

For $d_\ell \ge 2$, a convenient sufficient condition for Eq. (74) is

$$\boxed{\lambda\left[(m!\, d_\ell)^{1/m} - m\right] \;\ge\; \log\!\left(\frac{(1+d_\ell)C_0}{\varepsilon_\ell}\right)}. \tag{75}$$

To verify it, put

$$a \;\coloneqq\; \frac{1}{\lambda}\log\!\left(\frac{(1+d_\ell)C_0}{\varepsilon_\ell}\right). \tag{76}$$

If $a > 0$, choose $\widehat{k} = \lceil a \rceil - 1$. Equation (75) gives

$$\widehat{k} + m < a + m \le (m!\, d_\ell)^{1/m}, \tag{77}$$

so

$$D_{\widehat{k}}^{(m)} = \frac{\prod_{j=1}^m (\widehat{k}+j)}{m!} \le \frac{(\widehat{k}+m)^m}{m!} < d_\ell. \tag{78}$$

Moreover,

$$(1 + D_{\widehat{k}}^{(m)}) C_0 e^{-(\widehat{k}+1)\lambda} \le (1+d_\ell) C_0 e^{-\lambda\lceil a\rceil} \le \varepsilon_\ell. \tag{79}$$

If $a \le 0$, the direct choice $\widehat{k} = 0$ has $D_0^{(m)} = 1 < d_\ell$ and also satisfies Eq. (71). Therefore, Eq. (75) indeed certifies a strictly narrower layer. It is intentionally only sufficient.

E INFORMATION-THEORETIC PERSPECTIVES

This appendix records information-theoretic evidence for the exponent in our retained-width bound.

E.1 KOLMOGOROV-WIDTH

Let K be a compact set strictly inside a complex domain $D \subset \mathbb{C}^m$, and let

$$\mathcal{A}_M(D) \;\coloneqq\; \{f \in H^\infty(D) : \|f\|_{H^\infty(D)} \le M\}\big|_K \tag{80}$$

be the restrictions to K of a bounded holomorphic ball. For standard nested-domain regularity conditions, the Kolmogorov widths of this restriction class satisfy

$$d_N\big(\mathcal{A}_M(D); C(K)\big) \;=\; M\exp\big(-\Theta(N^{1/m})\big), \tag{81}$$

with domain-dependent constants; see the general theory of n-widths (Pinkus, 1985) and sharp holomorphic restriction asymptotics (Bandtlow & Nivoche, 2022). Inverting Eq. (81) gives the benchmark dimension

$$N_\varepsilon \;=\; \Theta\big(\big[\log(M/\varepsilon)\big]^m\big)\,. \tag{82}$$

Thus the same logarithmic power as our derivative-feature count appears for a canonical analytic function class. This comparison indicates that the exponent m is natural rather than a peculiarity of the rank-reduction algorithm. It is not a compression lower bound for one fixed teacher: obtaining such a result would require showing that the relevant teacher-generated class contains a comparably large analytic ball.

E.2 METRIC-ENTROPY

For prototypical analytic classes on nested cubes or polydisks, classical ε-entropy estimates have the form

$$\log \mathcal{N}\big(\varepsilon, \mathcal{A}_M(D), \|\cdot\|_\infty\big) \;=\; \Theta\Big(\big[\log(M/\varepsilon)\big]^{m+1}\Big) \tag{83}$$

under the corresponding geometric regularity conditions (Kolmogorov & Tikhomirov, 1961). A stable model with P bounded-precision real parameters has, heuristically, covering entropy of order at most $P\log(1/\varepsilon)$. Comparing with Eq. (83) suggests $P \gtrsim [\log(1/\varepsilon)]^m$. The word *heuristically* is essential: without bounded ranges, finite precision, or regularity of the parameter-to-function map, a raw parameter count is not an information-theoretic code-length bound. The entropy comparison is therefore supporting evidence, not a theorem about arbitrary neural parameterizations.

E.3 WHY NO ACTIVATION-UNIFORM LOWER BOUND IS CLAIMED

There is a complementary obstruction to a universal width lower bound. Maiorov & Pinkus (1999) construct a particular analytic, strictly monotone sigmoidal activation for which two-hidden-layer networks with fixed finite hidden widths are dense in continuous functions on compacta. Consequently, no lower bound based only on hidden width can hold uniformly over all analytic sigmoidal activations. A matching lower bound for the present setting would need additional restrictions—for example, a fixed nonpathological activation together with quantitative bounds on weights, precision, or parameter stability.

F CONSISTENCY WITH CLASSICAL APPROXIMATION RATES

There is no contradiction between our polylogarithmic fixed-teacher statement and classical polynomial approximation rates. Our theorem compresses one fixed teacher network, for which C_0 is a fixed constant. Universal approximation and minimax results instead concern an accuracy-indexed family: the approximating network $\mathcal{N}_1(\varepsilon)$ changes as ε changes. The following example shows explicitly that our finite weight range assumptions cannot hold uniformly along such a family, whereas they naturally hold for any fixed network.

F.1 A CONCRETE EXAMPLE

Consider scalar-output bilayer MLPs. In the bias-free notation, the derivative features $\partial_x^\alpha g(0, w)$ are weighted degree-$|\alpha|$ monomials in w, so matching all derivatives through order k reduces to matching the corresponding polynomial moments of the neuron weights. Let $\mathcal{P}_k$ be the linear space of polynomials in x of total degree at most k, and define

$$E_k(h) \;:=\; \inf_{p\in\mathcal{P}_k} \|h - p\|_{L^\infty(\Omega)}. \tag{84}$$

For a one-dimensional teacher output y, the Taylor-remainder argument in the proof of Theorem 5 gives

$$E_k(y) \;\le\; C_0 e^{-\lambda(k+1)}, \qquad C_0 \;:=\; V_{\max} M_g \frac{\rho_0}{\rho_0 - R}, \qquad \lambda \;:=\; \log(\rho_0/R). \tag{85}$$

Fix an integer $n \ge 1$ and

$$\Omega \;:=\; [-1, 1]^{d_{\text{in}}}, \qquad R = \sqrt{d_{\text{in}}}, \qquad f(x) \;:=\; \frac{1}{n!}(x_1)_+^n, \tag{86}$$

where $(t)_+^n = t^n$ for $t > 0$ and is zero otherwise. Three properties are relevant: $0 \in \Omega$; the target lies in the unit ball of $W^{n,\infty}(\Omega)$ because its weak derivative $\partial_1^n f = \mathbf{1}_{\{x_1>0\}}$ is bounded, while $f \notin C^n$; and the singular set $\{x_1 = 0\}$ meets the interior of Ω. Bernstein's saturation theorem gives

$$\begin{aligned} E_k(f) &\ge \inf_{p\in\mathcal{P}_k^{(1\text{D})}} \left\| \frac{1}{n!}(t)_+^n - p(t) \right\|_{L^\infty([-1,1])} \\ &\ge c_n k^{-n}, \end{aligned} \tag{87}$$

for a Bernstein saturation constant $c_n > 0$ (DeVore & Lorentz, 1993), e.g., $c_1 \approx 0.298$ and $c_2 \approx 0.188$.

F.2 FINDING A UNIVERSAL APPROXIMATION FAMILY $\mathcal{N}_1(\varepsilon)$

At the first hidden layer the upstream map is $f_{\text{in}} = \text{id}$, hence $L_f \equiv 1$ and $\rho = \rho_\sigma / W_{\max}^{(1)}$. Assumption 4 is therefore equivalent to

$$W_{\max}^{(1)} < \frac{\rho_\sigma}{R} = \frac{\rho_\sigma}{\sqrt{d_{\text{in}}}} =: W. \tag{88}$$

Existence of the family. Because Eq. (88) caps the input weights, we need universality inside a bounded-input-weight class. For a superanalytic activation to which their result applies, Theorem 2.6 of Stinchcombe & White (1990) gives density of

$$\text{span}\big\{\sigma(w \cdot x + b) : \|(w, b)\|_\infty \le B\big\} \tag{89}$$

in $C(\Omega)$; Corollary 3.5 of Pinkus (1999) sharpens this to parameters in an arbitrarily small neighborhood of the origin. Choose once and for all $W_0 < W$ and such a fixed neighborhood. Then, for every $\varepsilon > 0$, there is a finite network

$$\mathcal{N}_1(\varepsilon)(x) = \sum_{i=1}^{q(\varepsilon)} v_i(\varepsilon)\, \sigma\big(w_i(\varepsilon)^\top x + b_i(\varepsilon)\big), \qquad \|w_i(\varepsilon)\| \le W_0, \tag{90}$$

such that

$$\|\mathcal{N}_1(\varepsilon) - f\|_{L^\infty(\Omega)} \le \varepsilon; \tag{91}$$

here $q(\varepsilon)$ is the number of hidden units and is free to grow with accuracy.

Which constants move with ε. The fixed margin $W_0 < W$ makes the effective radius $\rho[\varepsilon]$ uniformly larger than R. We may therefore choose one $\rho_0 > R$ for the whole family, so $\lambda = \log(\rho_0/R) > 0$ is fixed; the bounded incoming affine parameters also give a uniform upper bound on $M_g[\varepsilon]$. The remaining free factor in Eq. (85) is

$$V_{\max}[\varepsilon] = \sum_{i=1}^{q(\varepsilon)} |v_i(\varepsilon)|. \tag{92}$$

A uniform bound on $V_{\max}[\varepsilon]$ would give $C_0 = O(1)$ and hence the polylogarithmic retained width uniformly over this family. Even polynomial growth in $1/\varepsilon$ would leave $\log C_0 = O(\log(1/\varepsilon))$. We next show that this family cannot have either behavior: the uniform-in-ε form of Assumption 3 must fail much more decisively.

F.3 Why $V_{\max}$ must diverge

Combining Eq. (87), the approximation property of $\mathcal{N}_1(\varepsilon)$, and Eq. (85) gives, for all sufficiently large integers k,

$$\begin{aligned} c_n k^{-n} &\le E_k(f) \\ &\le \|f - \mathcal{N}_1(\varepsilon)\|_\infty + E_k(\mathcal{N}_1(\varepsilon)) \\ &\le \varepsilon + C_0[\mathcal{N}_1(\varepsilon)] e^{-\lambda(k+1)}. \end{aligned} \tag{93}$$

For sufficiently small ε, take

$$k_\varepsilon \;:=\; \left\lfloor \left(\frac{c_n}{2\varepsilon}\right)^{1/n} \right\rfloor. \tag{94}$$

Then $c_n k_\varepsilon^{-n} \ge 2\varepsilon$ and $k_\varepsilon + 1 > (c_n/(2\varepsilon))^{1/n}$. Equation (93) yields

$$\begin{aligned} C_0[\mathcal{N}_1(\varepsilon)] &\ge \varepsilon e^{\lambda(k_\varepsilon+1)} \\ &\ge \varepsilon \exp\left[\lambda \left(\frac{c_n}{2\varepsilon}\right)^{1/n}\right]. \end{aligned} \tag{95}$$

Consequently,

$$\log C_0[\mathcal{N}_1(\varepsilon)] = \Omega(\varepsilon^{-1/n}). \tag{96}$$

Every factor of C_0 other than $V_{\max}$ has a uniform upper bound in this construction. Hence

$$\log V_{\max}[\varepsilon] = \Omega(\varepsilon^{-1/n}), \tag{97}$$

so the outgoing-mass bound cannot hold uniformly along the accuracy-indexed family, although it remains finite for each fixed teacher.

F.4 Consequence: no contradiction, but agreement

If the derivative-matching certificate is required to approximate $\mathcal{N}_1(\varepsilon)$ by a compressed network $\mathcal{N}_2(\varepsilon)$ to error at most ε, its matching order must in particular make $C_0[\mathcal{N}_1(\varepsilon)]e^{-\lambda(k+1)} \le \varepsilon$. Equation (95) therefore forces

$$k \ge \left(\frac{c_n}{2\varepsilon}\right)^{1/n} + O(1), \tag{98}$$

and the retained derivative dimension obeys

$$D_k = \binom{d_{\text{in}} + k}{d_{\text{in}}} = \Omega(\varepsilon^{-d_{\text{in}}/n}). \tag{99}$$

Thus the same ε-exponent as the classical Sobolev rate reappears once the accuracy dependence of the teacher family is included.

Finally, we remark that our Assumptions 2 and 3, limiting $V_{\max}, W_{\max} < \infty$, naturally stand for one fixed network. Even if a family of networks is concerned, they arise naturally due to the weight regularization terms we typically have in training.

G Experimental details

We use the target in Eq. (29) on the Euclidean unit disk. Each teacher network is trained by unconstrained Adam on the same 20,000 Sobol inputs for seeds $0, \dots, 4$. Compression is then applied to the frozen trained weights for $k = 0, \dots, 10$, with no fine-tuning or seed selection after compression. The architecture in Fig. 2 is $[2, 512, 512, 512, 1]$, and all three hidden layers are compressed in backward order to the same $D_k = \binom{k+2}{2}$. Each frozen random initialization is an architecture- and seed-matched control.

Task error is MSE against q on $150{,}000$ held-out Sobol points. Function fidelity is the largest teacher–compression discrepancy on those points or in projected local searches initialized at the largest sampled discrepancies.

For the three-hidden-layer teachers, layer 1 has $\rho_1 = \pi/W_{\max}^{(1)} > 1$ for every seed and therefore passes the radius check. At layer 2 we first compute the exact distance to the nearest upstream sigmoid pole,

$$\rho_f^{(2)} = \min_j \frac{\sqrt{(b_j^{(1)})^2 + \pi^2}}{\|W_{j,:}^{(1)}\|_2}, \tag{100}$$

and then sample the complex-input Jacobian norm. At $R = 1$, define

$$\widehat{q}_2 = \frac{R W_{\max}^{(2)} \widehat{L}_{h^{(1)}}(R)}{\pi} \tag{101}$$

For every seed, $\widehat{q}_2 \geq 1$, directly witnessing failure of Assumption 4. More quantitatively, the first sampled violating radii for seeds $0, \ldots, 4$ are $(0.300, 0.287, 0.248, 0.290, 0.261)$. Since $\widehat{L}_{h^{(1)}}(r) \leq L_{h^{(1)}}(r)$, these give ρ_2 upper bounds at the same values and hence $R/\rho_2 \geq (3.33, 3.48, 4.03, 3.45, 3.83)$, respectively.

Seed	Teacher MSE	ρ_1	$\widehat{q}_2$	First task-preserving D_k (reduction)
0	2.3490×10^{-4}	1.2632	7.87	66 (57.9×)
1	3.6973×10^{-4}	1.2829	7.00	21 (522.7×)
2	2.4447×10^{-4}	1.0715	14.46	36 (187.7×)
3	2.3616×10^{-4}	1.2011	8.32	28 (303.6×)
4	3.0791×10^{-4}	1.0520	13.81	21 (522.7×)

Table 1: Per-seed diagnostics after backward compression of all three hidden layers. The original network has $527{,}361$ parameters. Layer 1 passes the radius check, while the sampled layer-2 diagnostic fails for every seed.

The three-hidden-layer teachers have median MSE 2.445×10^{-4}, median width-66 fidelity error 2.305×10^{-3}, and random-control median 2.60×10^{-11}. Test/train MSE ratios range from 0.99953 to 0.99980, and validation/train ratios from 0.99911 to 1.00342, so there is no generalization gap indicative of overfitting. Seeds 3 and 4 select step 2250 because both training and validation errors rise at the final checkpoint; this synchronized fluctuation is optimization instability rather than overfitting.