Sharpness, Stability, and Step-Size Scaling in Deep Polynomial Networks

Alexandru Crăciun
Technical University of Munich
a.craciun@tum.de

Debarghya Ghoshdastidar
Technical University of Munich
Munich Center for Machine Learning
ghoshdas@cit.tum.de

Abstract

Predicting the maximum stable learning rate of a neural network from its architecture alone has remained outside the reach of current theory, except for deep linear and shallow scalar models, where exact sharpness expressions are available. We close this gap for deep polynomial networks by proving the first computable, architecture-explicit lower bound on the largest Hessian eigenvalue at any interpolating minimum. The bound factorizes into three independently computable terms: a data-geometry factor, an architectural-capacity factor, and a label-energy factor. Combining our bound with the dynamical framework of Chemnitz and Engel [2025] and the stability conditions of Cohen et al. [2022], we derive critical step-size upper bounds for a range of optimizers from vanilla gradient descent to adaptive versions like Adam. The maximum stable learning rate for deep polynomial networks scales as $(\text{activation degree})^{(-2\times\text{depth})}$, placing realistic polynomial network training in the Edge of Stability regime. Our derivation shows that this scaling is a Hessian-level consequence of the network's algebraic structure, not a property of any specific optimizer or dataset. With identity activation, our bound recovers the exact deep linear sharpness of Mulayoff and Michaeli [2020].

1 Introduction

The loss landscape of a neural network is the central object relating the architecture, the optimization algorithm, and the learned function. Once the data, the architecture, and the loss are fixed, everything the optimizer has access to is encoded in this landscape, and the behavior of any gradient based method is governed by its local geometry. The first-order geometry, captured by the gradient, tells the optimizer where to move; the second-order geometry, captured by the Hessian, tells it how fast it can move without becoming unstable, how narrow or wide the basins of attraction are, and how sensitive the solution is to perturbation. Much of what makes deep learning work is determined by the Hessian.

Three lines of evidence make this concrete and all of them are concerned with the largest eigenvalue of the Hessian matrix $\lambda_{\max}$, which we call the *sharpness*:

Sharpness-Aware Optimization. The sharpness–generalization connection [Hochreiter and Schmidhuber, 1997] has influenced optimizer design: the sharpness is regularized by sharpness-aware minimization [Foret et al., 2021] and a host of related practical methods that have become standard in modern training pipelines; understanding architectural limits on sharpness transfers directly into bounds on what such methods can achieve.

Implicit Bias. At the level of implicit bias, gradient descent with label noise [Blanc et al., 2020] and large step-size [Damian et al., 2023] provably reduces sharpness along the manifold of interpolating global minima; the practical efficacy of this implicit regularization, however, depends on whether the architecture admits flat minima at all.

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

Edge of Stability. At the level of optimization, the step-size at which gradient descent (GD) remains stable is set by the sharpness: the classical bound $\eta < 2/\lambda_{\max}$ is tight, and violating it leads either to divergence or, as systematically documented by Cohen et al. [2022b], to the *Edge of Stability (EoS)*, a regime in which the iterates oscillate around minima whose curvature is precisely calibrated to the step-size, namely $\lambda_{\max} \approx 2/\eta$. Cohen et al. [2022a] extended the phenomenon across the optimizer family practitioners actually deploy (Adam, GD with momentum and preconditioning), establishing optimizer-specific stability conditions involving the sharpness as their central quantity.

These three phenomena rely on the same architectural object: $\lambda_{\max}$ at minima. Predicting how this quantity depends on the architecture is therefore a basic question, but it is also surprisingly difficult. The Hessian is high-dimensional, its spectrum is expensive to compute, and closed-form characterizations are rare outside of restrictive settings. For deep linear networks, a precise picture has emerged: Mulayoff and Michaeli [2020], and more recently Kamber and Parhi [2025], give exact closed-form expressions for $\lambda_{\max}$ at arbitrary minima of deep matrix factorization, and Marion and Chizat [2024] give a lower bound scaling linearly with depth for deep linear regression. For genuinely nonlinear networks, the corresponding picture is almost entirely missing: Arjevani et al. [2026] showed that one cannot simply transfer the linear network analysis, since the correct geometry is *data-weighted* rather than Euclidean and the equivalences that drive the linear case break down.

Lower bound on the sharpness. This paper closes that gap for arbitrary polynomial networks. Polynomial networks are a natural first step beyond the linear case, for three concrete reasons. First, deep linear networks are deployed nowhere, yet their exact sharpness theory [Mulayoff and Michaeli, 2020, Kamber and Parhi, 2025] has influenced how the community reasons about flat minima and step-size selection. Polynomial networks are the minimal family that adds the essential nonlinearities of deep networks (overparametrization, scaling symmetries, depth-induced ill-conditioning) while keeping the landscape exactly algebraic: their functional space is a well-studied algebraic variety [Kileel et al., 2019, Kubjas et al., 2024], and they admit algebraic tools that extend cleanly from the linear theory. Questions that remain empirical for ReLU networks therefore become provable for polynomial networks, and, to our knowledge, our results are the first optimization-side quantitative results in the neuroalgebraic-geometry program [Marchetti et al., 2025]. Second, the r^{-2h} step-size ceiling we derive is a first-principles, Hessian-level analogue of the depth scaling of maximal-update learning rates reported for ReLU networks by Jelassi et al. [2023]. In our derivation this scaling originates in the degree of homogeneity κ of the parametrization map (Section 2.1), which indicates what to look for in other homogeneous architectures. Third, polynomial networks appear in deployed systems for privacy-preserving inference [Gilad-Bachrach et al., 2016, Hesamifard et al., 2019], where higher-degree variants suffer numerical instability and gradient explosion during training [Hossain and Rahman, 2025]. This instability is a consequence of the step-size ceiling we derive. Our main result is a *computable* lower bound on $\lambda_{\max}$:

Theorem 1 (**Efficiently Computable Bound on the Sharpness).** *Let $h \geq 1$ be the depth and $r \geq 1$ the activation degree of a deep polynomial network p_θ with arbitrary layer widths $d_0, \ldots, d_h$. Fix any data and label matrices $X \in \mathbb{R}^{d_0 \times N}, Y \in \mathbb{R}^{d_h \times N}$ with $N \geq 1$ data points. Then, at any interpolating minimum θ^* of the regression-setting empirical risk $L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \|y_i - p_\theta(x_i)\|_2^2$, the sharpness is lower bounded as*

$$\lambda_{\max}\big(\nabla^2 L(\theta^*)\big) \geq \underbrace{\frac{2}{N}\left(\min_{1\leq i\leq N} \|x_i\|_2^{2r^{h-1}} - (N-1)\mu^{r^{h-1}}\right)}_{\textit{data-geometry factor}} \cdot \underbrace{\left(\kappa r^{\rho/\kappa}\delta\right)}_{\textit{architecture factor}} \cdot \underbrace{\max_{1\leq i\leq N}\left(\frac{\|y_i\|_2^2}{\|x_i\|_2^{2r^{h-1}}}\right)^{\frac{\kappa-1}{\kappa}}}_{\textit{label-energy factor}}.$$

In the theorem above, $\kappa = \frac{(r^h-1)}{(r-1)}$, $\rho = \sum_{j=1}^{h-1} j r^j$, and $\delta = \prod_{i=1}^{h-1} d_i^{\frac{1-r}{r}\frac{r^{h-i}}{\kappa}}$ are architecture constants, (x_i, y_i) form a data-label pair, and $\mu \coloneqq \max_{i\neq j} |\langle x_i, x_j\rangle|$ is the data coherence. The bound factorizes into three independent components: (i) a *data-geometry* term depending only on the data Gram matrix $X^\top X$, (ii) an *architectural-capacity* term depending only on r, h, and the *hidden* widths, and (iii) a *label-energy* term measuring the hardest normalized point in the training set. This factorization makes each contribution to the sharpness independently traceable to its source. The data-geometry factor can be sharpened by keeping the Gershgorin row sums instead of the coherence μ (Section 2.1). Using Theorem 3 and specializing at $r = 1$ *exactly* recovers the linear closed form of Mulayoff and Michaeli [2020], providing a strict generalization of one of the only sharpness theorems available for linear networks.

We leverage this bound using the dynamical framework of Wu et al. [2018] and Chemnitz and Engel [2025] and the stability conditions in Cohen et al. [2022a] to derive critical step-size upper bounds for four optimizers: full-batch gradient descent (GD), exponential moving average heavy ball (EmaHB), together with its preconditioned version (PrecondEmaHB), and Adam; for details, see Appendix C.

Corollary 2 (Upper Bound on the Critical Step-Size). *Let θ^* be an interpolating minimum of the empirical loss L and let $\lambda^*(r, h, d, X, Y)$ denote the lower bound given Theorem 3. Then θ^* is stable only if the step-size satisfies the following bound, for each of the optimizers below:*

$$\textit{GD: } \eta < \frac{2}{\lambda^*}, \qquad \textit{EmaHB: } \eta < \frac{2(1+\beta_1)}{\lambda^*(1-\beta_1)},$$

$$\textit{PrecondEmaHB: } \eta < \frac{2\lambda_{\max}(\mathbf{P})(1+\beta_1)}{\lambda^*(1-\beta_1)}, \qquad \textit{Adam: } \eta < \frac{2\epsilon\,(1+1\beta_1)}{\lambda^*\,(1-\beta_1)},$$

where $\beta_1, \epsilon, \mathbf{P}$ are hyperparameters of the optimizers.

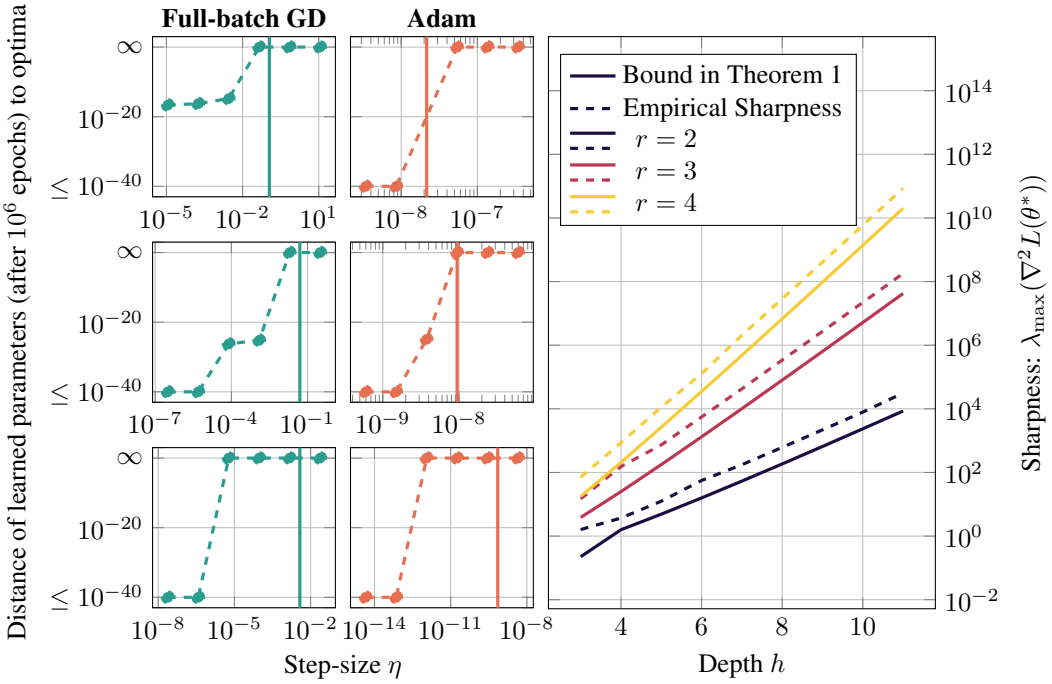


Figure 1: **Left figure:** distance between the parameters (after 10^6 epochs) and the manifold of optimal interpolators for full-batch GD and Adam. The network architectures have: **(top)** $r = 3$, widths $2-4-2$, **(middle)** $r = 2$, widths $2-4-4-2$, and **(bottom)** $r = 2$, widths $2-4-4-4-2$. We initialize in a ball of radius 10^{-30} around an optimal interpolator, hence larger distances indicate divergence or entry into the Edge of Stability regime. Corollary 2 provides the step-size thresholds (solid vertical lines) above which convergence is theoretically impossible.
Right figure: we plot the sharpness lower bound from Theorem 1 against the depth of the network as solid lines. The *empirical* sharpness of the flattest optimum is plotted as dashed lines. The network architecture has widths $50-5-\ldots-5-1$ with varying h and r. The bound in Theorem 1 is off by approximately a factor of 5, i.e., $\lambda_{\text{max,empirical}} \approx 5 \cdot \lambda_{\text{max,lower bound}}$.

Across all optimizers, combining Theorem 3 and Corollary 2, the ceiling on the step-size shrinks as r^{-2h} jointly in depth and activation degree. Throughout, this scaling refers to the normalization $\|x_i\|_2 = 1$ for all i and $\max_i \|y_i\|_2 = 1$, which is also the one used in our experiments; the normalization only sets the units in which the Hessian is measured, since rescaling the data changes the sharpness and our bound by exactly the same factor (Section 2.1). This yields an analogue of the maximal-update law in deep ReLU networks [Jelassi et al., 2023], here derived from the Hessian rather than the forward-pass scaling. Figure 1 illustrates the consequence: across architectures and optimizers, increasing the step-size beyond our predicted ceiling separates the regime in which training remains stable and approaches the manifold of minima from the regime in which it cannot.

Direct implications. Combining the lower bound on sharpness with the upper bound on the critical step-size has three implications worth stating directly. First, below an explicit width threshold, Edge of Stability is *forced*, not optional, for polynomial networks at any reasonable depth. With a common hidden width d, the architecture factor of Theorem 1 scales as $r^{2h}d^{-(r-1)/r}$, so the step-size required to remain in the convergent regime shrinks exponentially in h unless $d \gtrsim r^{2hr/(r-1)}$, that is, unless the width grows exponentially with the depth. Below this threshold, any step-size that is not exponentially small in h rules out convergence to an interpolating minimum, and training takes place in the Edge of Stability regime or beyond it. Second, our bound provides a quantitative answer to the question raised empirically by Lewkowycz et al. [2020]: *"Can an architecture-dependent maximal learning rate beyond which convergence fails be derived from first principles?"*. This has been answered analytically for the deep-linear case by Marion and Chizat [2024]. We provide the architecture-dependent maximal learning rate for deep polynomial networks, generalizing the deep-linear analog to a fully nonlinear deep family. Third, polynomial networks are deployed in real systems: most prominently in privacy-preserving inference under homomorphic encryption [Gilad-Bachrach et al., 2016, Hesamifard et al., 2019], where higher-degree variants are known to be unstable to train [Hossain and Rahman, 2025]. Our bound contributes a quantitative explanation: stable training requires step-size $\eta \lesssim r^{-2h}$ across the analyzed optimizers, with the prefactor differing by optimizer.

We summarize our technical contributions as follows:

- **A computable lower bound on sharpness.** The first architecture-explicit lower bound on $\lambda_{\max}(\nabla^2 L(\theta^*))$ at interpolating minima of an arbitrary polynomial network, factorizing into independently computable data, architecture, and label terms (cf. Table 1). Its non-simplified form (Theorem 3) recovers the exact Mulayoff and Michaeli [2020] sharpness in the linear ($r = 1$) limit.
- **The r^{-2h} scaling law.** We use the neural networks–dynamical systems framework to derive bounds on the maximal step-size beyond which convergence fails for polynomial networks. The bound shrinks as r^{-2h} uniformly across the multiple optimizers we analyze.
- **Algebraic toolkit for the polynomial parametrization map.** A closed-form expression for the differential of the polynomial network parametrization map. Until now, this has been an unresolved gap in the neuroalgebraic-geometry program. The expression can be used to approach second-order questions about polynomial networks, such as Hessian-based analyses of the loss landscape, computing Gauss-Newton spectra and conditioning [Pennington and Bahri, 2017, Zhao et al., 2024] or neural tangent kernels of polynomial networks [Zhu et al., 2022].

Before proceeding further, we state several clarifying remarks explicitly. We do not claim that our bound implies generalization properties: recent work [Wen et al., 2023, Andriushchenko et al., 2023b,a] has substantially weakened the flatness–generalization connection, and we treat sharpness purely as an optimization quantity. We treat full-batch deterministic dynamics; we do not address SGD-specific stability thresholds [Andreyev and Beneventano, 2024]. Our bound is only a lower bound; closing the gap to a matching upper bound, known for the linear case [Kamber and Parhi, 2025], remains open. Finally, our results apply to monomial activations; extension to ReLU and related architectures requires different machinery [Mulayoff et al., 2021, Nacson et al., 2023].

Section 2 introduces notation and contains our main results: the sharpness lower bound (Section 2.1) and the algebraic apparatus driving its proof (Section 2.2). Section 3 surveys the four lines of work our contribution bridges. Section 4 discusses limitations, open questions, and concludes.

2 Main Results

This section presents the paper's main theorems. We begin with the central result of the paper: a closed-form lower bound on the largest Hessian eigenvalue of the empirical loss at any interpolating minimum, expressed as a product of three independent factors. The bound is established through a projection argument that bypasses the data-weighted geometry obstruction identified by Arjevani et al. [2026] and that requires only a closed-form expression for the differential of the polynomial parametrization map. We develop this differential in the second subsection and provide an explicit formula. The exact derivation of upper bounds on the maximal stable step-sizes from Corollary 2 is provided in Appendix C. The polynomial network results at $r = 1$ recover the exact deep linear sharpness expression of Mulayoff and Michaeli [2020]. Full proofs of all results are in the appendices.

Notation and Setting. We write $h \geq 1$ for the network depth, $r \geq 1$ for the monomial activation degree $\sigma : z \mapsto z^r$, and $d = (d_0, d_1, \ldots, d_h)$ for the layer widths, with d_0, d_h the input, output dimensions, respectively. The trainable parameters are $\theta = (W_1, \ldots, W_h) \in \Theta$ with $W_k \in \mathbb{R}^{d_k \times d_{k-1}}$ in the parameter space $\Theta = \prod_{i=1}^{h} \mathbb{R}^{d_i \times d_{i-1}}$. A polynomial neural network p_θ is then defined as

$$p_\theta(x) := W_h \sigma\big(W_{h-1}\sigma(\cdots \sigma(W_2 \sigma(W_1 x)) \cdots)\big),$$

with the activation acting component-wise. For clarity of exposition, we treat the bias-free setting but our results extend directly to polynomial networks with non-zero biases. The data and label matrices are $X \in \mathbb{R}^{d_0 \times N}$ and $Y \in \mathbb{R}^{d_h \times N}$, with $N \geq 1$. With quadratic loss, the average empirical risk is

$$L(\theta) := \frac{1}{N} \sum_{i=1}^{N} \|y_i - p_\theta(x_i)\|_2^2.$$

Three architecture-dependent constants recur throughout the section:

$$\kappa := \sum_{i=1}^{h} r^{h-i} = \frac{r^h - 1}{r - 1}, \qquad \rho := \sum_{i=1}^{h-1} i\, r^i, \qquad \delta := \prod_{i=1}^{h-1} d_i^{\frac{1-r}{r} \cdot \frac{r^{h-i}}{\kappa}}.$$

Spectral conventions: $\sigma_{\min}^{+}(M)$ denotes the smallest positive singular value of M, $\lambda_{\min}^{+}(M)$ the smallest positive eigenvalue of a symmetric M, and $(M)^{\circ k}$ the entrywise (Hadamard) k-th power.

2.1 Sharpness Lower Bound at Interpolating Minima

The central result of the paper is a closed-form lower bound on the largest Hessian eigenvalue of the empirical loss at any interpolating minimum, computable directly from the data matrices and three architecture-dependent constants. The bound factorizes into three independent components, each corresponding to a distinct source of curvature: the geometry of the data through a Hadamard-power Gram matrix, the architecture through a closed-form constant in the depth, activation degree, and hidden widths, and the labels through the optimal functional interpolator.

Theorem 3 (Sharpness lower bound at interpolating minima). *Let $h, r \geq 1$, let $d = (d_0, d_1, \ldots, d_h)$ be the layer widths, and let $X \in \mathbb{R}^{d_0 \times N}$ and $Y \in \mathbb{R}^{d_h \times N}$ be a dataset for which there exists θ^* with $L(\theta^*) = 0$. Let $X^{\uparrow r^{h-1}}$ be the lifted data matrix with columns $x_i^{\otimes r^{h-1}}$ and let $\mathcal{X} := \{x \in \mathbb{R}^{d_0} : x^{\otimes r^{h-1}} \in \mathrm{range}(X^{\uparrow r^{h-1}})\}$. Then, with κ, ρ, δ as in the notation,*

$$\lambda_{\max}\big(\nabla^2 L(\theta^*)\big) \;\geq\; \frac{2}{N}\, \lambda_{\min}^{+}\big((X^\top X)^{\circ r^{h-1}}\big) \cdot \Big(\kappa\, r^{\rho/\kappa}\, \delta\Big) \cdot \sup_{\substack{\|u\|=\|x\|=1 \\ x \in \mathcal{X}}} \langle u, p_{\theta^*}(x)\rangle^{2\frac{\kappa-1}{\kappa}}.$$

The supremum ranges only over inputs whose lift lies in the lifted data span, since the Hessian of the empirical loss sees the inputs only through this span (see Appendix B). The set $\mathcal{X}$ contains all the normalized data points $\hat{x}_i := x_i / \|x_i\|_2$, on $\mathcal{X}$ all interpolating minima represent the same function (Lemma 13), and $\mathcal{X} = \mathbb{R}^{d_0}$ whenever the lifted data span $\mathrm{Sym}_{r^{h-1}}(\mathbb{R}^{d_0})$.

The $r = 1$ recovery. The bound recovers the deep linear sharpness expression of Mulayoff and Michaeli [2020] at $r = 1$. In this limit the Khatri-Rao power degenerates: $r^{h-1} = 1$, $\kappa = h$, $\rho = 0$, $\delta = 1$, and $(X^\top X)^{\circ 1} = X^\top X$. The comparison with Mulayoff and Michaeli [2020] assumes that X has full row rank, so that $(XX^\top)^{-1}$ exists; then $\mathrm{range}(X) = \mathbb{R}^{d_0}$, hence $\mathcal{X} = \mathbb{R}^{d_0}$, and the restricted supremum of Theorem 3 coincides with the supremum over the unit sphere. The interpolation hypothesis in the linear case forces $\sup_{\|u\|_2=\|x\|_2=1} \langle u, p_{\theta^*}(x)\rangle^{2(1-\frac{1}{h})}$ to equal the operator norm $\sigma_{\max}(YX^\top(XX^\top)^{-1})^{2(1-\frac{1}{h})}$ at the flattest minimum, and for whitened data $\lambda_{\min}^{+}(X^\top X) = 1$. Mulayoff and Michaeli [2020] don't average over the data, thus the N^{-1} term drops. The bound collapses to

$$\lambda_{\max}\big(\nabla^2 L(\theta^*)\big) \;\geq\; 2h\, \sigma_{\max}\big(YX^\top(XX^\top)^{-1}\big)^{2(1-\frac{1}{h})},$$

which matches the exact value derived by Mulayoff and Michaeli [2020] at the flattest minimum of a deep linear network. The bound is therefore tight at the linear limit, providing a strict generalization of one of the only sharpness theorems available for arbitrary-depth networks.

Depth-degree scaling. Treating the bound as a function of h and r at fixed widths and dataset reveals its qualitative behavior. The factor $\kappa\, r^{\rho/\kappa}\delta$ grows as r^{2h} for $r \geq 2$, since $\kappa \asymp r^h/(r-1)$ and $r^{\rho/\kappa} \sim r^{h-r/(r-1)}$ in the deep limit. The width factor δ approaches a depth-independent constant determined by the hidden layers closest to the input (see below), the data-geometry factor $\lambda_{\min}^{+}((X^\top X)^{\circ r^{h-1}})$ is strictly positive, and the label-energy factor is bounded under any reasonable normalization of the data (see the normalization paragraph below). The averaging coefficient $2/N$ is architecture-independent. The net consequence is that the sharpness lower bound grows as r^{2h} in the depth-degree pair, with each added monomial layer multiplying the bound by approximately r^2. As we develop in Appendix C, this scaling translates directly into an upper bound on the maximum stable step-size of $\eta_{\max} \lesssim r^{-2h}$ across all optimizers we consider.

Interplay of depth, degree, and width. The hidden widths enter only through δ, and with a negative exponent: width plays a flattening role, while depth increases the bound exponentially. With a common hidden width $d_1 = \cdots = d_{h-1} = d$, one has $\delta = d^{-\frac{r-1}{r}\cdot\frac{\kappa-1}{\kappa}}$, hence $\kappa r^{\rho/\kappa}\delta \sim r^{2h} d^{-(r-1)/r}$ in the deep limit. Offsetting a single added layer therefore requires multiplying the width by approximately $r^{2r/(r-1)}$, and keeping the architecture factor bounded requires $d \gtrsim r^{2hr/(r-1)}$, a width exponential in depth. Since Theorem 3 is a lower bound, whether the minima of wide networks are actually flatter would require a matching upper bound, which remains open for $r \geq 2$. Appendix B.5 discusses the roles of κ, ρ, and δ in detail.

Normalization. Rescaling the data changes the sharpness and our bounds by exactly the same factor. Writing $L_{X,Y}$ for the empirical risk on the dataset (X, Y), under $(X, Y) \mapsto (cX, c^{r^{h-1}}Y)$, $c > 0$, the interpolating minima are unchanged and

$$\lambda_{\max}(\nabla^2 L_{cX,c^{r^{h-1}}Y}(\theta^*)) = c^{2r^{h-1}}\lambda_{\max}(\nabla^2 L_{X,Y}(\theta^*)),$$

while the bounds in Theorems 1 and 3 pick up the same factor $c^{2r^{h-1}}$. The normalization therefore only sets the scale in which the Hessian is measured; the r^{2h} law for the sharpness and the r^{-2h} law for the step-size refer to the convention $\|x_i\|_2 = 1$ for all i and $\max_i \|y_i\|_2 = 1$ of our experiments.

Computational simplifications. Theorem 3 as stated in this section is different from the one in the introduction. As stated in this section computing the smallest positive eigenvalue of the Hadamard Gram matrix, $\lambda_{\min}^{+}((X^\top X)^{\circ r^{h-1}})$ and the supremum of the inner product $\sup\langle u, p_{\theta^*}(x)\rangle$ is computationally expensive. However, we can avoid both of these problems as we now describe.

Lower bounding the Hadamard Gram matrix eigenvalue. The matrix $(X^\top X)^{\circ r^{h-1}}$ has entries $\langle x_i, x_j\rangle^{r^{h-1}}$. Gershgorin's theorem [Horn and Johnson, 1985, Theorem 6.1.1] gives the row-sum bound below, and bounding each off-diagonal sum by $(N-1)$ times the maximum off diagonal entry, with the data coherence $\mu := \max_{i\neq j} |\langle x_i, x_j\rangle|$, yields the coarser but more interpretable coherence bound used in Theorem 1:

$$\lambda_{\min}^{+}((X^\top X)^{\circ r^{h-1}}) \geq \min_{1\leq i\leq N}(\|x_i\|_2^{2r^{h-1}} - \sum_{j\neq i} |\langle x_i, x_j\rangle|^{r^{h-1}}) \geq \min_i \|x_i\|_2^{2r^{h-1}} - (N-1)\mu^{r^{h-1}}.$$

If the row-sum bound is positive, the matrix is positive definite and its smallest eigenvalue is $\lambda_{\min}^{+}$; otherwise both bounds hold trivially. Theorem 1 remains valid with the sharper row-sum bound as data-geometry factor, which requires the same Gram matrix $X^\top X$ to compute. The coherence bound is positive if and only if $(N-1)(\mu/\min_i \|x_i\|_2^2)^{r^{h-1}} < 1$. Hence, for $N \geq 2$, it remains non-vacuous at large depth if and only if $\mu < \min_i \|x_i\|_2^2$, in which case its ratio to $\min_i \|x_i\|_2^{2r^{h-1}}$ converges to 1 exponentially fast in r^{h-1}. For unit-norm data, this is the condition $\mu < 1$, i.e., no two data points are parallel.

Lower bounding the label-energy factor. The $\sup_{\|u\|_2=\|x\|_2=1,\, x\in\mathcal{X}}\langle u, p_{\theta^*}(x)\rangle^{2\frac{\kappa-1}{\kappa}}$ factor is the supremum of the inner product between the network output and a unit vector, over all unit inputs in $\mathcal{X}$ and all unit vectors. It can be lower bounded by restricting x to the normalized data points $\hat{x}_i = x_i/\|x_i\|_2$, which lie in $\mathcal{X}$, setting $u = p_{\theta^*}(x_i)/\|p_{\theta^*}(x_i)\|_2 = y_i/\|y_i\|_2$ (where the last equality comes from the interpolation hypothesis), and taking the maximum of the resulting ratio over all the datapoints. Using the r^{h-1}-homogeneity of the polynomial map $p_\theta(cx) = c^{r^{h-1}}p_\theta(x)$ we get

$$\sup_{\substack{\|u\|_2=\|x\|_2=1\\ x\in\mathcal{X}}} \langle u, p_{\theta^*}(x)\rangle \geq \max_{1\leq i\leq N} \frac{\langle y_i, p_{\theta^*}(\hat{x}_i)\rangle}{\|y_i\|_2} = \max_{1\leq i\leq N}\left(\frac{\|y_i\|_2}{\|x_i\|_2^{r^{h-1}}}\right).$$

Random design. The input dimension d_0 enters the bound only through the data-geometry factor, and for isotropic random inputs the coherence concentrates.

Corollary 4 (Random design). *Let $N \geq 2$, let $x_1, \ldots, x_N$ be drawn i.i.d. from the uniform distribution on the unit sphere, and let $y_i := p_{\theta_T}(x_i)$ be generated by a fixed teacher θ_T. Fix $\gamma > 0$. If $d_0 \geq 8(\gamma+2)N^{2/r^{h-1}} \log N$, then, with probability at least $1 - N^{-\gamma}$, at any interpolating minimum θ^*,*

$$\lambda_{\max}(\nabla^2 L(\theta^*)) \geq \frac{1}{N} \cdot \left(\kappa r^{\rho/\kappa}\delta\right) \cdot \max_{1\leq i\leq N} \|y_i\|_2^{2\frac{\kappa-1}{\kappa}}.$$

The proof (Appendix B.4) combines a spherical-cap estimate with a union bound to obtain $1 - (N-1)\mu^{r^{h-1}} \geq 1/2$. The dimension requirement weakens rapidly with depth, and as d_0 grows the bound concentrates on the architecture and label-energy factors, so the r^{-2h} law is cleanest in high dimension; general, non-isotropic designs are left for future work.

Two caveats. We make two caveats explicit. First, the bound applies at interpolating minima, where the residual term in the Hessian decomposition vanishes; at non-interpolating critical points the second fundamental form contribution must be handled separately and the projection argument breaks down. It can be shown there exist arbitrarily flat *strictly local* minima for deep polynomial networks, and it is an open question why first-order optimizers avoid these. Second, the depth-degree scaling is structural. The network map p_θ is κ-homogeneous in the parameters and the Hessian inherits this homogeneity (Appendix B.5); the bound therefore captures the correct order of magnitude of the sharpness, even if the other constant might not be tight. Closing the gap to a matching upper bound, which is known for the deep-linear case [Kamber and Parhi, 2025], remains open for $r \geq 2$.

Beyond the quadratic loss. The proof of Theorem 3 uses the quadratic loss only through the Gauss-Newton form of the Hessian at interpolation. For a loss $\ell(\hat{y}, y)$ with $\nabla_{\hat{y}}\ell(y, y) = 0$ that attains an interpolating minimum at finite parameters, the Hessian at that minimum is the Gauss-Newton form weighted by $\nabla^2_{\hat{y}}\ell(y_i, y_i)$. If $c := \min_i \lambda_{\min}(\nabla^2_{\hat{y}}\ell(y_i, y_i)) > 0$, the proof carries through with the quadratic-loss constant 2 replaced by c, and Theorems 1 and 3 hold with $2/N$ replaced by c/N.

2.2 Differential of the Network Parametrization Map

Associated to any neural network architecture, there exists a parametrization map which, for a set of parameters θ produces the function p_θ that the network represents; for polynomial networks, following the convention of Kileel et al. [2019], we denote this parametrization map by $\mu_{r,d}$. We have that $\mu_{r,d}(\theta) = p_\theta$. Generically, $\mu_{r,d}$ maps into the space of functions $C(\mathbb{R}^{d_0}, \mathbb{R}^{d_h})$, however, for polynomial networks, we can identify points $\mu_{r,d}(\theta)$ with symmetric tensors in $\mathrm{Sym}_{r^{h-1}}(\mathbb{R}^{d_0})^{d_h}$.

We now develop a closed-form expression for the differential of the polynomial network parametrization map $\mu_{r,d}$, which is the main analytic result of the paper and is used in the proof of Theorem 3. The formula is stated in Theorem 5 below and its verification at $r = 1$ confirms that it recovers the standard product rule for deep linear networks. For a differentiable function $f : \mathbb{R}^m \to \mathbb{R}^n$, by $D_\theta f(U)$ we denote its differential at θ acting on U; $x \otimes y$ denotes the tensor product of x and y and $x^{\otimes m}$ is the m-times tensor product of x with itself.

Kileel et al. [2019, Section 3.2] established that any polynomial network of depth h, monomial activation degree r, and layer widths $d = (d_0, \ldots, d_h)$ can be written in the form $p_\theta(x) = \mu_{r,d}(\theta)\, x^{\otimes r^{h-1}}$, where the parametrization map is given explicitly by

$$\mu_{r,d}(\theta) = \mathrm{SymRow}\Big(W_h\big(W_{h-1}\cdots(W_2 W_1^{\bullet r})^{\bullet r}\cdots\big)^{\bullet r}\Big).$$

Here the row-wise Khatri-Rao power $W^{\bullet r} \in \mathbb{R}^{a\times b^r}$ replaces each row w_j of $W \in \mathbb{R}^{a\times b}$ by $\mathrm{vec}(w_j^{\otimes r})$ and SymRow projects each row, viewed as an element of $(\mathbb{R}^{d_0})^{\otimes r^{h-1}}$, onto the fully symmetric subspace. We refer to Kileel et al. [2019] for the construction of this representation.

The differential of the parametrization map admits a closed-form expression that we now describe. Two recursive constructions are needed to state it. The intermediate states are defined by $M_1 := W_1$ and $M_k := W_k M_{k-1}^{\bullet r}$ for $k \geq 2$, so that $M_h \in \mathbb{R}^{d_h \times d_0^{r^{h-1}}}$ and $\mu_{r,d}(\theta) = \mathrm{SymRow}(M_h)$. The single-layer variation at layer i is $V_1 := U_1$ and $V_i := U_i M_{i-1}^{\bullet r}$ for $i \geq 2$, and the transmission propagating V_i from layer i to k is defined by $\Phi_i^i := V_i$ and $\Phi_i^k := W_k\big(\Phi_i^{k-1} \odot M_{k-1}^{\bullet(r-1)}\big)$ for $k > i$.

Theorem 5 (Differential of the network parametrization map). *Let $\theta = (W_1, \ldots, W_h)$ and let $U = (U_1, \ldots, U_h)$ be a tangent direction. Then the differential $\mathcal{K}(U) := D_\theta \mu_{r,d}(U)$ is given by*

$$\mathcal{K}(U) \;=\; U_h \,\mathrm{SymRow}\big(M_{h-1}^{\bullet r}\big) \;+\; W_h \sum_{i=1}^{h-1} r^{h-i}\,\mathrm{SymRow}\big(\Phi_i^{h-1} \odot M_{h-1}^{\bullet(r-1)}\big).$$

Here and above, the row-wise Kronecker product $A \odot B$ tensors corresponding rows of A and B.

Verification at $r = 1$. At the linear limit, the Khatri-Rao power degenerates: $W^{\bullet 1} = W$ and $W^{\bullet 0}$ is the all-ones column vector, so the intermediate states reduce to $M_k = W_k W_{k-1} \cdots W_1$, the single-path transmissions reduce to $\Phi_i^{h-1} = W_{h-1} \cdots W_{i+1}\, U_i\, W_{i-1} \cdots W_1$, and the depth-counter weights r^{h-i} collapse to one. The formula in Theorem 5 reduces to the standard product rule

$$D_{(W_1 \cdots W_h)}\mu(U_1, \ldots, U_h) \;=\; \sum_{i=1}^{h} W_h \cdots W_{i+1}\, U_i\, W_{i-1} \cdots W_1,$$

recovering the differential of the deep-linear parametrization [see Trager et al., 2020, Section 3.1]. This verification confirms that the Khatri-Rao machinery and the depth-counter weights r^{h-i} are the nonlinear corrections needed to extend the linear formula to monomial activations of arbitrary degree.

Networks with bias. The parametrization map of Kileel et al. [2019] treats the bias-free case. Adding a bias vector $b_k \in \mathbb{R}^{d_k}$ extends the construction by treating the augmented weight matrices $\widetilde{W}_k := [W_k \mid b_k] \in \mathbb{R}^{d_k \times (d_{k-1}+1)}$ acting on augmented activations, and the differential formula of Theorem 5 carries through with modification that the widths d_i are replaced by $d_i + 1$.

3 Related work

Our work bridges four lines of research that have remained largely disconnected: the exact spectral analysis of deep linear networks, the neuroalgebraic geometry of polynomial models, the dynamical systems theory behind Edge of Stability, and the multi-optimizer stability conditions that have emerged from the adaptive-optimizer literature.

Sharpness of deep linear networks. Mulayoff and Michaeli [2020] gave the first closed-form expression for the maximum Hessian eigenvalue at the *flattest* global minimum of a deep linear network. Our bound (Theorem 3) recovers their expression exactly at $r = 1$. Kamber and Parhi [2025] extend this analysis from the flattest minimum to *every* minimizer. Their result is the natural deep linear analog of what we prove for polynomial networks; the comparison at $r = 1$ is exact at the flattest minimum but loses information away from it, and closing this gap for $r \geq 2$ is open.

The closest nonlinear precedent is the line of work due to Mulayoff et al. [2021], which extends the deep linear sharpness–stability analysis to single hidden layer univariate ReLU networks. The follow-up by Nacson et al. [2023] generalizes this to multivariate ReLU. These works share with ours the goal of converting an algebraic description of the function class into a quantitative sharpness statement, but the mechanics of ReLU and monomial activations are disjoint. Singh et al. [2025] characterize the rank of the Hessian for deep linear and several shallow nonlinear networks, providing tight upper bounds on its kernel as a function of layer widths.

Polynomial networks, tensor decomposition, and neuroalgebraic geometry. The shallow ($h = 2$, $r = 2$) special case of our setting is the quadratic activation network, whose landscape was characterized by Soltanolkotabi et al. [2019]: in the over-parametrized regime every local minimum is global, all saddles are strict, and gradient descent converges linearly. Their methods do not extend to depth $h \geq 3$ or to higher monomial degree. The most quantitatively detailed Hessian computation in the polynomial-network landscape is Arjevani and Vinograd [2023], who construct Puiseux series families of critical points for order-r symmetric tensor decomposition (the shallow $h = 2$ instance of our parametrization) and provide rigorous analytic estimates for the objective values and *full Hessian spectrum* at these critical points. Arjevani et al. [2026] prove an Eckart–Young-type theorem for quadratic shallow polynomial networks under Gaussian data, characterizing all critical points and full Hessian signatures of the data loss under a teacher-induced data distribution. Their main conclusion, that the data-dependent landscape is governed by a *data-weighted* norm with no Frobenius substitute, highlights why extending the linear analysis to the non-linear setting is non-trivial.

Setting	Result on the Sharpness	Reference
Deep Linear	Exact expression for the *smallest* sharpness	Mulayoff and Michaeli [2020, Theorem 1]
	Lower bound at global minima	Marion and Chizat [2024, Theorem 1]
	Expression for the *rank of the Hessian*	Singh et al. [2021, Theorem 3]
	Exact expression for *any* global minimum	Kamber and Parhi [2025, Theorem 3.1]
Shallow ReLU	Implicit lower bound	Mulayoff et al. [2021, Lemma 4]
	Exact expression for scalar input	Singh et al. [2025, Theorem 3.4]
Deep ReLU	Implicit upper and lower bound	Nacson et al. [2023, Lemmas 2–3]
Shallow Polynomial	Exact expression	Arjevani and Vinograd [2023]
Deep Polynomial	Lower bound at interpolating minima	Our work — Theorem 3

Table 1: Known bounds/expressions on the sharpness.

The algebraic-geometric description of polynomial network function spaces was initiated by Kileel et al. [2019]. Kubjas et al. [2024] gave semialgebraic descriptions of such functional spaces for several width regimes, introduced the *learning degree*, and conjectured non-defectivity for non-increasing widths. Finkel et al. [2025] prove an *activation-degree threshold*, above which the functional space is non-defective, and show that equi-width networks are non-defective for all $r \geq 2$. Usevich et al. [2026] prove generic identifiability of deep polynomial networks under non-increasing widths and improve the activation-degree bound. Massarenti and Mella [2025] give the sharpest Alexander–Hirschowitz-type statement available for neurovarieties, fully resolving when single-output functional spaces attain their expected dimension and proving global identifiability. The Marchetti et al. [2025] survey consolidates these threads under the heading of *neuroalgebraic geometry*, advocating algebraic-geometric tools for the study of neural network function spaces. To our knowledge, no work in this program has previously addressed optimization-side quantitative results: convergence rates, sharpness, or learning-rate scaling, and our work is the first to do so.

Edge of Stability and dynamical systems analysis. The empirical Edge of Stability phenomenon: that $\lambda_{\max}$ progressively grows during training and then hovers just above $2/\eta$, was systematically documented by Cohen et al. [2022b]. Its earliest theoretical formulation is the *catapult* mechanism of Lewkowycz et al. [2020], which identifies three phases while training: a lazy phase $\eta < 2/\lambda_0$, a catapult phase $2/\lambda_0 < \eta < \eta_{\max}$ in which the curvature is *actively reduced* by the dynamics until $\lambda_{\max} < 2/\eta$, and a divergent phase. The mechanism behind the catapult was made precise by Damian et al. [2023]. Their analysis is essentially equivalent to projected gradient descent on the constraint $\lambda_{\max}(\theta) \leq 2/\eta$, giving a *dynamic upper bound* that complements our static lower bound. The most recent state-of-the-art predictive theory of EoS dynamics is the *central flows* framework of Cohen et al. [2024], which approximates the GD trajectory by a deterministic ODE that captures sharpness evolution with high empirical accuracy. Their framework predicts sharpness *trajectories* during training; ours provides an architectural *limit* on the sharpness, and the two are complementary.

A parallel line of work has used dynamical systems theory to formalize the link between sharpness and step-size stability. Wu et al. [2018] introduced the dynamical-stability framework that has become standard in this literature. Ahn et al. [2022] extended the analysis to non-convex settings. Chemnitz and Engel [2025] gave a definitive characterization for stochastic gradient descent in overparametrized learning; we adopt their formalism in our derivation relating sharpness to critical step-size (Appendix C). MacDonald et al. [2025] analyze the dynamics of GD beyond the Edge of Stability for deep linear networks, showing that for $\eta > 2/\lambda^*_{\max}$ convergence to global minima becomes impossible and *provide the first rigorous explanation of the behavior of GD in this regime: the dynamics instead stabilize in a periodic trajectory around the flattest minimum*. Marion and Chizat [2024] derive an upper bound on the critical step-size for linear scalar networks; our bound provides the corresponding answer for polynomial networks. Beneventano and Woodworth [2025] prove that for a depth-2 scalar linear network, GD with $\eta \approx 2/\lambda_{\max}$ converges linearly to a minimum with strictly lower norm and strictly lower $\lambda_{\max}$ than the gradient-flow limit.

Sharpness across optimizers. Cohen et al. [2022a] extended the EoS phenomenon from vanilla gradient descent to multiple optimizers. Their stability conditions are the inputs for Corollary 2: substituting our architectural lower bound on $\lambda_{\max}$ into each condition yields an explicit critical step-size depending on the optimizer. Long and Bartlett [2024] showed that sharpness-aware minimization (SAM) likewise operates at an Edge of Stability regime, reinforcing the centrality of $\lambda_{\max}$ across the modern optimizer landscape. For stochastic gradient descent, Andreyev and Beneventano [2024] introduced the *Edge of Stochastic Stability* phenomenon, showing that minibatch SGD does not settle at $\lambda_{\max} = 2/\eta$ but at a different threshold; this is one of the key reasons we treat the SGD case only conjecturally. Closing this gap quantitatively for polynomial networks remains open.

The connection between sharpness and generalization, originally proposed by Hochreiter and Schmidhuber [1997] and revived empirically by Keskar et al. [2017], motivates the explicit regularizer of Foret et al. [2021] and the implicit-bias results of Blanc et al. [2020] and Damian et al. [2023], which show that GD with label noise or large step-size is itself a sharpness-reducing flow on the manifold of interpolating minima. Our results give an architecture-imposed *limit* on this minimization: for polynomial networks of depth h and degree r, no minimum interpolating dataset is flatter than the bound in our main theorem, and the implicit bias toward flat minima can therefore push $\lambda_{\max}$ no lower than r^{2h} in scale. We note, however, that recent work [Wen et al., 2023, Andriushchenko et al., 2023b] has substantially weakened the flatness–generalization connection, and we treat sharpness throughout as an optimization quantity rather than a generalization proxy.

4 Discussion and Conclusion

We have established a closed-form lower bound on the maximum Hessian eigenvalue at any interpolating minimum of a polynomial network (Theorem 3), factorized into independent data-geometry, architectural-capacity, and label-energy terms, and we have shown that this bound translates into critical step-size upper bounds for multiple optimizers, including full-batch gradient descent and Adam. Besides the intractable bound in Theorem 3, we also provide computationally friendly version of it (Theorem 1). The bound in Theorem 3 recovers the deep linear sharpness expression of Mulayoff and Michaeli [2020] exactly, and its key consequence is that the maximum stable step size shrinks as r^{-2h}, placing realistic polynomial network training squarely in the Edge of Stability regime.

Several limitations bound the scope of these statements. The bound is a lower bound only, and the gap to a matching upper bound, available for the deep linear case through Kamber and Parhi [2025], remains open for $r \geq 2$. The depth-degree scaling is structural, since the parametrization map is κ-homogeneous in the parameters, with $\kappa \sim r^h/(r-1)$, but the precise multiplicative constants are unlikely to be tight. As illustrated in the right panel of Figure 1, our bound is off by about a factor of 5, however, this still translates into meaningful upper bounds for the critical step-size (cf. left panel of Figure 1). We suspect this gap between the empirical sharpness and the bound of Theorem 1 is due to the data-geometry factor. We leave the improvement of this bound as an open question.

Our bound holds true only at an interpolating minimum, *without any further restrictions on the data*. Extending our results to non-interpolating minima is open and to solve this we believe an in-depth investigation of residual factor in the structure of the Hessian is needed. For the same reason, our results cover the quadratic loss and, more generally, any loss that attains an interpolating minimum at finite parameters and is locally strongly convex there, but not losses whose infimum is approached only as $\|\theta\| \to \infty$, such as the exponential loss or the cross-entropy loss with hard labels. Such losses admit no finite interpolating minimum, and the sharpness at infinity requires a different analysis.

Finally, our results apply to monomial activations; the extension to general polynomial activations is considerably more involved and left for future work. Extension to ReLU and related architectures requires different machinery, exemplified by Mulayoff et al. [2021] and Nacson et al. [2023]. We do not treat the stochastic case, and characterizing the SGD-specific stability threshold is open.

Two implications are worth stating explicitly. First, our results give an architectural floor on the implicit-bias: for polynomial networks of depth h and degree r, no minimum interpolating a generic dataset is flatter than the bound in Theorem 3, and the implicit bias toward flat minima can therefore push $\lambda_{\max}$ no lower than r^{2h} in scale. For polynomial networks of depth h and activation degree r, the maximum stable step size scales as r^{-2h} across all standard first-order optimizers, placing realistic training squarely in the Edge of Stability regime, and this scaling is a Hessian-level consequence of the network's algebraic structure rather than a property of any specific optimizer or dataset.

Acknowledgments and Disclosure of Funding

This work has been partly supported by the German Research Foundation (DFG) through the Priority Program SPP 2298 / project GH 257/2-2, and by the German Academic Exchange Service (DAAD) through project 57761786.

References

Kwangjun Ahn, Jingzhao Zhang, and Suvrit Sra. Understanding the unstable convergence of gradient descent. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, Proceedings of Machine Learning Research, pages 247–257. PMLR, 2022.

Arseniy Andreyev and Pierfrancesco Beneventano. Edge of stochastic stability: Revisiting the edge of stability for SGD. *CoRR*, abs/2412.20553, 2024. doi: 10.48550/ARXIV.2412.20553.

Maksym Andriushchenko, Dara Bahri, Hossein Mobahi, and Nicolas Flammarion. Sharpness-aware minimization leads to low-rank features. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023a.

Maksym Andriushchenko, Francesco Croce, Maximilian Müller, Matthias Hein, and Nicolas Flammarion. A modern look at the relationship between sharpness and generalization. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pages 840–902. PMLR, 2023b.

Yossi Arjevani and Gal Vinograd. Symmetry & critical points for symmetric tensor decomposition problems. *CoRR*, abs/2306.07886, 2023. doi: 10.48550/ARXIV.2306.07886.

Yossi Arjevani, Joan Bruna, Joe Kileel, Elzbieta Polak, and Matthew Trager. Geometry and Optimization of Shallow Polynomial Networks. *SIAM Journal on Applied Algebra and Geometry*, 10 (2):174–209, June 2026. ISSN 2470-6566. doi: 10.1137/25M1732994.

Keith Ball. An elementary introduction to modern convex geometry. In Silvio Levy, editor, *Flavors of Geometry*, volume 31 of *Mathematical Sciences Research Institute Publications*, pages 1–58. Cambridge University Press, Cambridge, 1997.

Pierfrancesco Beneventano and Blake Woodworth. Gradient Descent Converges Linearly to Flatter Minima than Gradient Flow in Shallow Linear Networks, January 2025.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an Ornstein-Uhlenbeck like process. In Jacob D. Abernethy and Shivani Agarwal, editors, *Conference on Learning Theory, COLT 2020, 9-12 July 2020, Virtual Event [Graz, Austria]*, Proceedings of Machine Learning Research, pages 483–513. PMLR, 2020.

Dennis Chemnitz and Maximilian Engel. Characterizing Dynamical Stability of Stochastic Gradient Descent in Overparameterized Learning. *Journal of Machine Learning Research*, 26(134):1–46, 2025.

Jeremy Cohen, Alex Damian, Ameet Talwalkar, J. Zico Kolter, and Jason D. Lee. Understanding Optimization in Deep Learning with Central Flows. In *The Thirteenth International Conference on Learning Representations*, October 2024.

Jeremy M. Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E. Dahl, and Justin Gilmer. Adaptive Gradient Methods at the Edge of Stability, July 2022a.

Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient Descent on Neural Networks Typically Occurs at the Edge of Stability, November 2022b.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Bella Finkel, Jose Israel Rodriguez, Chenxi Wu, and Thomas Yahl. Activation degree thresholds and expressiveness of polynomial neural networks. *Algebraic Statistics*, 16(2):113–130, September 2025. ISSN 2693-3004, 2693-2997. doi: 10.2140/astat.2025.16.113.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 201–210, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `https://proceedings.mlr.press/v48/gilad-bachrach16.html`.

Ehsan Hesamifard, Hassan Takabi, and Mehdi Ghasemi. Deep Neural Networks Classification over Encrypted Data. In *Proceedings of the Ninth ACM Conference on Data and Application Security and Privacy*, pages 97–108, Richardson Texas USA, March 2019. ACM. ISBN 978-1-4503-6099-9. doi: 10.1145/3292006.3300044.

Sepp Hochreiter and Jürgen Schmidhuber. Flat Minima. *Neural Computation*, 9(1):1–42, January 1997. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1997.9.1.1.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1 edition, December 1985. ISBN 978-0-521-38632-6 978-0-521-30586-0 978-0-511-81081-7. doi: 10.1017/CBO9780511810817.

Forsad Al Hossain and Tauhidur Rahman. A Training Framework for Optimal and Stable Training of Polynomial Neural Networks, May 2025.

Samy Jelassi, Boris Hanin, Ziwei Ji, Sashank J. Reddi, Srinadh Bhojanapalli, and Sanjiv Kumar. Depth dependence of μP learning rates in ReLU mlps. *CoRR*, abs/2305.07810, 2023. doi: 10.48550/ARXIV.2305.07810.

Anil Kamber and Rahul Parhi. Sharpness of minima in deep matrix factorization: Exact expressions. *CoRR*, abs/2509.25783, 2025. doi: 10.48550/ARXIV.2509.25783.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*, 2017.

Joe Kileel, Matthew Trager, and Joan Bruna. On the expressive power of deep polynomial neural networks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10310–10319, 2019.

Kaie Kubjas, Jiayi Li, and Maximilian Wiesmann. Geometry of polynomial neural networks. *Algebraic Statistics*, 15(2):295–328, December 2024. ISSN 2693-3004, 2693-2997. doi: 10.2140/astat.2024.15.295.

Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: The catapult mechanism, March 2020.

Philip M. Long and Peter L. Bartlett. Sharpness-aware minimization and the edge of stability. *Journal of Machine Learning Research*, 25(179):1–20, 2024.

Lachlan MacDonald, Hancheng Min, Leandro Palma, Salma Tarmoun, Ziqing Xu, and Rene Vidal. Convergence rates for gradient descent on the edge of stability for overparametrised least squares. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Giovanni Luca Marchetti, Vahid Shahverdi, Stefano Mereta, Matthew Trager, and Kathlén Kohn. An Invitation to Neuroalgebraic Geometry, January 2025.

Pierre Marion and Lénaïc Chizat. Deep linear networks for regression are implicitly regularized towards flat minima. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

A. Massarenti and M. Mella. The Alexander-Hirschowitz theorem for neurovarieties, November 2025.

Rotem Mulayoff and Tomer Michaeli. Unique Properties of Flat Minima in Deep Networks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7108–7118. PMLR, November 2020.

Rotem Mulayoff, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability: A view from function space. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual*, pages 17749–17761, 2021.

Mor Shpigel Nacson, Rotem Mulayoff, Greg Ongie, Tomer Michaeli, and Daniel Soudry. The implicit bias of minima stability in multivariate shallow ReLU networks. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2798–2806. PMLR, August 2017.

Sidak Pal Singh, Gregor Bachmann, and Thomas Hofmann. Analytic insights into structure and rank of neural network hessian maps. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 23914–23927. Curran Associates, Inc., 2021.

Sidak Pal Singh, Weronika Ormaniec, and Thomas Hofmann. Cracking the Hessian: Closed-Form Hessian Spectra for Fundamental Neural Networks, October 2025.

Mahdi Soltanolkotabi, Adel Javanmard, and Jason D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2019. doi: 10.1109/TIT.2018.2854560.

Matthew Trager, Kathlén Kohn, and Joan Bruna. Pure and Spurious Critical Points: A Geometric Study of Linear Networks, April 2020.

Konstantin Usevich, Ricardo Borsoi, Clara Dérand, and Marianne Clausel. Identifiability of deep polynomial neural networks, 2026. URL `https://arxiv.org/abs/2506.17093`.

Kaiyue Wen, Zhiyuan Li, and Tengyu Ma. Sharpness minimization algorithms do not only minimize sharpness to achieve better generalization. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Lei Wu, Chao Ma, and Weinan E. How sgd selects the global minima in over-parameterized learning: a dynamical stability perspective. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in neural information processing systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/6651526b6fb8f29a00507de6a49ce30f-Paper.pdf`.

Jim Zhao, Sidak Pal Singh, and Aurélien Lucchi. Theoretical characterisation of the gauss newton conditioning in neural networks. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/d067d16e3e5fe8fa8a3e62909907659a-Abstract-Conference.html.

Zhenyu Zhu, Fanghui Liu, Grigorios Chrysos, and Volkan Cevher. Generalization properties of NAS under activation and skip connection search. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/952b691c116bf753daafa6ce274e81bb-Abstract-Conference.html.

A Mathematics of the Parametrization Map

This appendix establishes the closed-form differential of the polynomial-network parametrization map stated as Theorem 5 in the main text. The proof proceeds in two stages. Three preparatory lemmas in Section A.1 develop the calculus of the row-wise Khatri-Rao product and its interaction with the symmetrization operator. Section A.2 assembles the lemmas into a proof of Theorem 5 by induction on depth.

We use throughout the notation introduced in Section 2.2 of the main text. We recall the recursive constructions for convenience. The intermediate states are $M_1 := W_1$ and $M_k := W_k M_{k-1}^{\bullet r}$ for $k \geq 2$, the single-layer source variation injected at layer i is $V_1 := U_1$ and $V_i := U_i M_{i-1}^{\bullet r}$ for $i \geq 2$, and the single-path transmission propagating V_i from layer i to layer k is $\Phi_i^i := V_i$ and $\Phi_i^k := W_k\big(\Phi_i^{k-1} \odot M_{k-1}^{\bullet(r-1)}\big)$ for $k > i$. We also recall the notation $\mathcal{K}(U) := D_\theta \mu_{r,d}(U)$ for the differential of the parametrization map, and we write $\mathcal{K}_h$ when emphasizing the depth dependence in inductive arguments.

A.1 Preparatory Lemmas

The first lemma computes the differential of the row-wise Khatri-Rao power $W \mapsto W^{\bullet r}$.

Lemma 6 (Derivative of the Khatri-Rao power)**.** *For $W \in \mathbb{R}^{a\times b}$, the map $W \mapsto W^{\bullet r}$ has differential*

$$D_W\big(W \mapsto W^{\bullet r}\big)(U) \;=\; \sum_{j=1}^{r} W^{\bullet(j-1)} \odot U \odot W^{\bullet(r-j)}.$$

Proof. The map acts row by row: row i of $W^{\bullet r}$ is $g(w_i) := \mathrm{vec}\big(w_i^{\otimes r}\big)$, where w_i is the i-th row of W. The differential of g at w_i in direction u_i (the i-th row of U) is, by the multilinearity of the tensor power,

$$D_{w_i} g(u_i) \;=\; \sum_{j=1}^{r} w_i^{\otimes(j-1)} \otimes u_i \otimes w_i^{\otimes(r-j)}.$$

Stacking these row derivatives produces a matrix whose i-th row is the vectorization of the displayed sum, and recalling the operator $\odot$ as the Hadamard product, i.e., taking the row-wise tensor product, identifies this stacked matrix as $\sum_{j=1}^{r} W^{\bullet(j-1)} \odot U \odot W^{\bullet(r-j)}$. □

The second lemma collapses the symmetrization of the sum that appears in Lemma 6 into a single term times a multiplicity factor of r. This is the structural identity that produces the depth-counter weights r^{h-i} in Theorem 5.

Lemma 7 (Symmetrized Khatri-Rao average)**.** *For $W \in \mathbb{R}^{a\times b}$ and U of the same shape,*

$$\mathrm{SymRow}\left(\sum_{j=1}^{r} W^{\bullet(j-1)} \odot U \odot W^{\bullet(r-j)}\right) \;=\; r\,\mathrm{SymRow}\big(U \odot W^{\bullet(r-1)}\big).$$

Proof. The symmetrization operator SymRow acts row by row, projecting each row, viewed as an element of $V^{\otimes r}$ for $V = \mathbb{R}^b$, onto the fully symmetric subspace $\mathrm{Sym}^r V$. On the i-th row, the inner sum reads

$$\sum_{j=1}^{r} w_i^{\otimes(j-1)} \otimes u_i \otimes w_i^{\otimes(r-j)},$$

and applying the symmetrization to each summand yields the symmetric tensor $u_i \odot w_i^{\odot(r-1)}$, where $\odot$ on vectors denotes the symmetric tensor product. This expression is independent of the slot index j at which u_i was inserted, since the symmetric tensor product is invariant under permutation of factors. Summing over $j = 1, \ldots, r$ produces a factor of r, giving

$$\sum_{j=1}^{r} \mathrm{Sym}\big(w_i^{\otimes(j-1)} \otimes u_i \otimes w_i^{\otimes(r-j)}\big) \;=\; r\,u_i \odot w_i^{\odot(r-1)}.$$

Stacking row-wise and recognizing the right-hand side as the i-th row of $r\,U \odot W^{\bullet(r-1)}$ after symmetrization completes the proof. □

The third lemma states that two matrices with equal symmetric components remain equivalent after right-tensoring with a third matrix and symmetrizing. This invariance allows the inductive argument in Section A.2 to substitute the inductive hypothesis under symmetrization without disturbing intermediate factors.

Lemma 8 (Symmetrization invariance under right $\odot$-tensoring)**.** *Let A, B be matrices with the same row-wise tensor type, and let C be a matrix with a compatible row-wise tensor type. If $\mathrm{SymRow}(A) = \mathrm{SymRow}(B)$, then*

$$\mathrm{SymRow}(A \odot C) \;=\; \mathrm{SymRow}(B \odot C).$$

Proof. The operations act row by row, so it suffices to prove the analogous statement at the level of tensors. Let $a, b \in V^{\otimes p}$ satisfy $\mathrm{Sym}_p(a) = \mathrm{Sym}_p(b)$, and let $c \in V^{\otimes q}$. We show that $\mathrm{Sym}_{p+q}(a \otimes c) = \mathrm{Sym}_{p+q}(b \otimes c)$, where Sym_{p+q} denotes the symmetrization over all $p+q$ slots and Sym_p denotes the symmetrization over the first p slots.

Let $\mathrm{Sym}_{p+q}^{(1\dots p)}$ be the partial symmetrization acting on the first p slots of a $(p+q)$-tensor, leaving the remaining q slots unchanged. The fully symmetric subspace of $V^{\otimes(p+q)}$ is contained in the range of $\mathrm{Sym}_{p+q}^{(1\dots p)}$, so the composition satisfies $\mathrm{Sym}_{p+q} \circ \mathrm{Sym}_{p+q}^{(1\dots p)} = \mathrm{Sym}_{p+q}$. Applying this to $a \otimes c$,

$$\mathrm{Sym}_{p+q}(a \otimes c) \;=\; \mathrm{Sym}_{p+q}\big(\mathrm{Sym}_{p+q}^{(1\dots p)}(a \otimes c)\big) \;=\; \mathrm{Sym}_{p+q}\big(\mathrm{Sym}_p(a) \otimes c\big).$$

The same argument applied to b gives $\mathrm{Sym}_{p+q}(b \otimes c) = \mathrm{Sym}_{p+q}(\mathrm{Sym}_p(b) \otimes c)$, and the assumed equality $\mathrm{Sym}_p(a) = \mathrm{Sym}_p(b)$ produces the conclusion. □

A.2 Proof of Theorem 5

We now assemble the lemmas into a proof of the differential formula.

Proof of Theorem 5. Let $\Delta M_k := D_\theta M_k(U_1, \dots, U_k)$ denote the unsymmetrized total differential of the intermediate state M_k. Combining the matrix product rule with the formula for the differential of the Khatri-Rao power (Lemma 6) yields the recurrence

$$\Delta M_1 = U_1, \qquad \Delta M_k = U_k M_{k-1}^{\bullet r} + W_k \sum_{j=1}^{r} M_{k-1}^{\bullet(j-1)} \odot \Delta M_{k-1} \odot M_{k-1}^{\bullet(r-j)} \quad (k \geq 2).$$

By the linearity of the derivative, we decompose $\Delta M_k = \sum_{i=1}^{k} \Delta_i M_k$, where $\Delta_i M_k$ is the contribution obtained by setting all entries of U to zero except U_i. Each isolated branch starts at depth i as $\Delta_i M_i = V_i$ and propagates upward through the recurrence

$$\Delta_i M_k \;=\; W_k \sum_{j=1}^{r} M_{k-1}^{\bullet(j-1)} \odot \Delta_i M_{k-1} \odot M_{k-1}^{\bullet(r-j)}, \qquad k > i.$$

We claim that for every $i \in \{1, \dots, h\}$,

$$\mathrm{SymRow}\big(\Delta_i M_h\big) \;=\; r^{h-i}\,\mathrm{SymRow}\big(\Phi_i^h\big), \qquad \Phi_i^h := \begin{cases} V_h & \text{if } i = h, \\ \Phi_i^{h-1} \odot M_{h-1}^{\bullet(r-1)} & \text{if } i < h, \end{cases} \tag{1}$$

where the case $i < h$ is understood without a leading W_h factor at the topmost step; the W_h factor is restored when summing branches in the final assembly. We prove (1) by induction on the depth h.

Base case. For $h = 1$, only $i = 1$ is relevant. Since $\Delta_1 M_1 = V_1$, $\Phi_1^1 = V_1$, and $r^0 = 1$, the identity (1) reads $\mathrm{SymRow}(V_1) = \mathrm{SymRow}(V_1)$, which holds trivially.

Inductive step. Assume (1) holds at depth $h-1$. We treat the case $i = h$ and the case $i < h$ separately.

For $i = h$, the recurrence gives $\Delta_h M_h = U_h M_{h-1}^{\bullet r} = V_h$ and $\Phi_h^h = V_h$, with multiplicity $r^{h-h} = 1$, so (1) holds.

For $i \in \{1, \dots, h-1\}$, applying the recurrence and the linearity of SymRow,

$$\mathrm{SymRow}(\Delta_i M_h) \;=\; W_h\,\mathrm{SymRow}\left(\sum_{j=1}^{r} M_{h-1}^{\bullet(j-1)} \odot \Delta_i M_{h-1} \odot M_{h-1}^{\bullet(r-j)}\right),$$

where we have used that left multiplication by the row-acting matrix W_h commutes with SymRow. Lemma 7 collapses the inner sum:

$$\mathrm{SymRow}\left(\sum_{j=1}^{r} M_{h-1}^{\bullet(j-1)} \odot \Delta_i M_{h-1} \odot M_{h-1}^{\bullet(r-j)}\right) \;=\; r\,\mathrm{SymRow}\big(\Delta_i M_{h-1} \odot M_{h-1}^{\bullet(r-1)}\big).$$

The induction hypothesis gives $\mathrm{SymRow}(\Delta_i M_{h-1}) = r^{h-1-i}\,\mathrm{SymRow}(\Phi_i^{h-1})$, and Lemma 8 lifts this equality through right tensoring by $M_{h-1}^{\bullet(r-1)}$:

$$\mathrm{SymRow}\big(\Delta_i M_{h-1} \odot M_{h-1}^{\bullet(r-1)}\big) \;=\; r^{h-1-i}\,\mathrm{SymRow}\big(\Phi_i^{h-1} \odot M_{h-1}^{\bullet(r-1)}\big).$$

Combining the two previous equations and recognizing the right-hand side as $r^{h-i}\,\mathrm{SymRow}(\Phi_i^h)$ proves (1) at depth h for $i < h$.

Assembly. The differential derivative of the parametrization map is

$$\mathcal{K}(U) \;=\; D_\theta \mu_{r,d}(U) \;=\; \mathrm{SymRow}(\Delta M_h) \;=\; \sum_{i=1}^{h} \mathrm{SymRow}(\Delta_i M_h),$$

using the linearity of SymRow in the first equality and the branch decomposition in the second. Substituting the result (1) for each branch and separating the $i = h$ term from the $i < h$ terms,

$$\mathcal{K}(U) \;=\; \mathrm{SymRow}(V_h) \;+\; \sum_{i=1}^{h-1} r^{h-i}\, W_h\,\mathrm{SymRow}\big(\Phi_i^{h-1} \odot M_{h-1}^{\bullet(r-1)}\big),$$

where the leading W_h has been factored out of the second term using its commutativity with SymRow. Recognizing $V_h = U_h M_{h-1}^{\bullet r}$ and $\mathrm{SymRow}(V_h) = U_h\,\mathrm{SymRow}(M_{h-1}^{\bullet r})$ yields the formula stated in Theorem 5. □

B Proof of the Sharpness Lower Bound

This appendix establishes the sharpness lower bound stated as Theorem 3 in the main text. The proof proceeds in three stages. Section B.1 isolates the per-point lower bound, which controls the supremum of the Hessian quadratic form when restricted to a single training input. Section B.2 carries out the column-space projection argument that converts the bound on the lifted data matrix into a bound on the Hadamard Gram of the unlifted data. Section B.3 records the properties of the lifted data span on which the label-energy supremum is taken and assembles the two reductions into a proof of the main theorem. Section B.4 proves the random-design Corollary 4, and Section B.5 contains the full versions of the remarks following Theorem 3 in the main text.

We use throughout the notation introduced in Section 2 of the main text and recalled in Appendix A. We additionally introduce the following auxiliary quantities. For an input $x \in \mathbb{R}^{d_0}$, the forward-pass states are $z_0 := x$, $z_1 := W_1 z_0$, and $z_k := W_k z_{k-1}^{\circ r}$ for $k \geq 2$, where $\circ r$ denotes the entrywise r-th power. The diagonal Jacobians are $D_k := r\,\mathrm{diag}\big(z_{k-1}^{\circ(r-1)}\big)$, the backpropagated Jacobians are $A_h^h := I_{d_h}$ and $A_i^h := W_h D_h \cdots W_{i+1} D_{i+1}$ for $i \in \{1, \dots, h-1\}$, and the layerwise inputs are $b_1^h := z_0$ and $b_i^h := z_{i-1}^{\circ r}$ for $i \geq 2$.

B.1 Per-Point Lower Bound

We begin with a forward-pass identity that expresses the differential evaluated on a symmetric rank-one tensor as a sum of layerwise terms.

Lemma 9 (Forward-pass identity)**.** *Let $x \in \mathbb{R}^{d_0}$ and let $v_h := x^{\otimes r^{h-1}}$. Then*

$$\mathcal{K}_h(U)\, v_h \;=\; \sum_{i=1}^{h} A_i^h\, U_i\, b_i^h.$$

Proof. By induction on the depth h.

Base case. For $h = 1$, $\mathcal{K}_1(U_1) = U_1$ by Theorem 5, and $v_1 = x = z_0 = b_1^1$ with $A_1^1 = I_{d_1}$. The identity reads $U_1 x = A_1^1 U_1 b_1^1$, which is immediate.

Inductive step. Assume the identity holds at depth h. To pass to depth $h+1$, observe that $v_{h+1} = x^{\otimes r^h} = (x^{\otimes r^{h-1}})^{\otimes r} = v_h^{\otimes r}$. Theorem 5 applied at depth $h+1$ gives

$$\mathcal{K}_{h+1}(U)\, v_{h+1} = U_{h+1}\, \mathrm{SymRow}\big(M_h^{\bullet r}\big)\, v_h^{\otimes r} \;+\; W_{h+1} \sum_{i=1}^{h} r^{h+1-i}\, \mathrm{SymRow}\big(\Phi_i^h \odot M_h^{\bullet (r-1)}\big)\, v_h^{\otimes r}.$$

Two row-wise tensor contractions are needed.

For the first term, row j of $M_h^{\bullet r}$ is $\mathrm{vec}\big((M_h)_{j,:}^{\otimes r}\big)$, which contracted with $v_h^{\otimes r}$ yields $\big((M_h)_{j,:}\, v_h\big)^r$ since the symmetric tensor power is invariant under symmetrization of the contracting tensor. Stacking over j gives $\mathrm{SymRow}(M_h^{\bullet r})\, v_h^{\otimes r} = (M_h\, v_h)^{\circ r}$. We claim that $M_h\, v_h = z_h$. This follows by an inner induction: $M_1\, v_1 = W_1\, x = z_1$, and assuming $M_{k-1}\, v_{k-1} = z_{k-1}$, one has $M_k\, v_k = W_k M_{k-1}^{\bullet r} v_k = W_k\, z_{k-1}^{\circ r} = z_k$ by the same row-wise contraction argument applied at depth k. Hence the first term equals $U_{h+1}\, z_h^{\circ r} = A_{h+1}^{h+1}\, U_{h+1}\, b_{h+1}^{h+1}$.

For the second term, splitting the row-wise contraction gives

$$\big(\Phi_i^h \odot M_h^{\bullet (r-1)}\big)\, v_h^{\otimes r} \;=\; (\Phi_i^h\, v_h) \circ \big(M_h^{\bullet (r-1)}\, v_h^{\otimes (r-1)}\big) \;=\; (\Phi_i^h\, v_h) \circ z_h^{\circ (r-1)},$$

where the second equality applies the same argument as above to $r-1$ slots. Symmetrizing row by row and recognizing $\mathrm{diag}(z_h^{\circ (r-1)})$ as a row-wise scaling,

$$\mathrm{SymRow}\big(\Phi_i^h \odot M_h^{\bullet (r-1)}\big)\, v_h^{\otimes r} \;=\; \mathrm{diag}\big(z_h^{\circ (r-1)}\big)\, (\Phi_i^h\, v_h).$$

Therefore the second term equals

$$W_{h+1} \sum_{i=1}^{h} r^{h+1-i}\, \mathrm{diag}\big(z_h^{\circ (r-1)}\big)\, (\Phi_i^h\, v_h) \;=\; W_{h+1} D_{h+1} \sum_{i=1}^{h} r^{h-i}\, \Phi_i^h\, v_h,$$

where we have absorbed one factor of r into D_{h+1} via $D_{h+1} = r\, \mathrm{diag}(z_h^{\circ (r-1)})$ and matched the multiplicities $r^{h+1-i} = r \cdot r^{h-i}$.

The sum $\sum_{i=1}^{h} r^{h-i}\, \Phi_i^h\, v_h$ is recognized through Theorem 5 applied at depth h. Specifically, contracting Theorem 5 at depth h with v_h,

$$\mathcal{K}_h(U_1, \ldots, U_h)\, v_h \;=\; U_h\, \mathrm{SymRow}\big(M_{h-1}^{\bullet r}\big)\, v_h \;+\; W_h \sum_{i=1}^{h-1} r^{h-i}\, \mathrm{SymRow}\big(\Phi_i^{h-1} \odot M_{h-1}^{\bullet (r-1)}\big)\, v_h.$$

The same row-wise contraction arguments as above identify the right-hand side as $\sum_{i=1}^{h} r^{h-i}\, \Phi_i^h\, v_h$ after a bookkeeping rearrangement using $\Phi_i^h = W_h\big(\Phi_i^{h-1} \odot M_{h-1}^{\bullet (r-1)}\big)$ for $i < h$ and $\Phi_h^h = V_h$. Thus

$$W_{h+1} D_{h+1} \sum_{i=1}^{h} r^{h-i}\, \Phi_i^h\, v_h \;=\; W_{h+1} D_{h+1}\, \mathcal{K}_h(U_1, \ldots, U_h)\, v_h.$$

Combining the two terms,

$$\mathcal{K}_{h+1}(U)\, v_{h+1} \;=\; A_{h+1}^{h+1}\, U_{h+1}\, b_{h+1}^{h+1} \;+\; W_{h+1} D_{h+1}\, \mathcal{K}_h(U_1, \ldots, U_h)\, v_h.$$

The induction hypothesis substitutes $\mathcal{K}_h(U_1, \ldots, U_h)\, v_h = \sum_{i=1}^{h} A_i^h\, U_i\, b_i^h$. For $i \le h$, the definition $A_i^{h+1} = W_{h+1} D_{h+1} A_i^h$ holds by inspection of the backpropagated Jacobian recursion, and $b_i^{h+1} =$

b_i^h since the layerwise input at index i does not depend on layers above i. Summing the depth-$h+1$ top-layer term and the substituted inductive sum yields

$$\mathcal{K}_{h+1}(U)\, v_{h+1} \;=\; \sum_{i=1}^{h+1} A_i^{h+1}\, U_i\, b_i^{h+1},$$

completing the induction. □

We now use the forward-pass identity to lower-bound the Hessian quadratic form when contracted with a rank-one symmetric test tensor.

Proposition 10 (Per-point lower bound)**.** *Let θ^* be an interpolating minimum of the empirical loss, and fix unit vectors $u \in \mathbb{R}^{d_h}$ and $x \in \mathbb{R}^{d_0}$. Then*

$$\sup_{\|U\|=1} \left(u^\top \mathcal{K}(U)\, x^{\otimes r^{h-1}}\right)^2 \;\geq\; \kappa\, r^{\rho/\kappa}\, \delta \left(u^\top \mu_{r,d}(\theta^*)\, x^{\otimes r^{h-1}}\right)^{2(\kappa-1)/\kappa}.$$

Proof. Set $v_h := x^{\otimes r^{h-1}}$ and $Q := uv_h^\top$, a unit-Frobenius rank-one symmetric tensor in the codomain of $\mu_{r,d}$. We work with the dual operator $\mathcal{T} := \mathcal{K}\mathcal{K}^*$ acting on the codomain, and lower-bound its Rayleigh quotient at Q.

Step 1: dual-operator reformulation. The supremum of interest is the squared operator norm of the linear functional $\mathcal{A} : U \mapsto u^\top \mathcal{K}(U) v_h = \langle Q, \mathcal{K}(U)\rangle$ on the unit sphere of tangent space, which, using the C^*-algebra equality becomes:

$$\sup_{\|U\|=1} \left(u^\top \mathcal{K}(U)\, v_h\right)^2 = \|\mathcal{A}\|_{\mathrm{op}}^2 = \lambda_{\max}(\mathcal{A}\mathcal{A}^*) = \sup_{|e|=1} |\mathcal{A}\mathcal{A}^*(e)|,$$

where $e \in \mathbb{R}$. To compute the adjoint $\mathcal{A}^*$, observe the equality

$$\langle \mathcal{A}^*(e), U\rangle = \langle e, \mathcal{A}(U)\rangle = e\langle Q, \mathcal{K}(U)\rangle = \langle e\mathcal{K}^*(Q), U\rangle.$$

Hence, the dual operator acting on any vector $e \in \mathbb{R}$ is $\mathcal{A}\mathcal{A}^*(e) = e\langle Q, \mathcal{K}\mathcal{K}^*(Q)\rangle$. Taking the supremum over the unit ball in $\mathbb{R}$ means just multiplying by the identity and we get

$$\sup_{\|U\|=1} (u^\top \mathcal{K}(U) v_h)^2 = \langle Q, \mathcal{T}(Q)\rangle.$$

Step 2: computing the adjoint. For any tangent direction U, Lemma 9 gives $\mathcal{K}(U)v_h = \sum_{i=1}^h A_i^h\, U_i\, b_i^h$, so

$$\langle Q, \mathcal{K}(U)\rangle \;=\; u^\top \mathcal{K}(U) v_h \;=\; \sum_{i=1}^h u^\top A_i^h\, U_i\, b_i^h \;=\; \sum_{i=1}^h \big\langle (A_i^h)^\top u\, (b_i^h)^\top,\, U_i \big\rangle.$$

By the definition of the adjoint, $\langle Q, \mathcal{K}(U)\rangle = \langle \mathcal{K}^*(Q), U\rangle$, and matching components gives the closed form

$$\mathcal{K}^*(Q) \;=\; \big((A_1^h)^\top u\, (b_1^h)^\top,\, (A_2^h)^\top u\, (b_2^h)^\top,\, \ldots,\, (A_h^h)^\top u\, (b_h^h)^\top\big).$$

Step 3: evaluating the Rayleigh quotient. The Rayleigh quotient factorizes across layers:

$$\langle Q,\, \mathcal{T}(Q)\rangle \;=\; \sum_{i=1}^h \big\|(A_i^h)^\top u\, (b_i^h)^\top\big\|_F^2 \;=\; \sum_{i=1}^h \|(A_i^h)^\top u\|_2^2 \cdot \|b_i^h\|_2^2,$$

where the last equality uses that the Frobenius norm of an outer product factorizes as the product of vector norms. Setting $\alpha_i := \|(A_i^h)^\top u\|_2$ and $\beta_i := \|b_i^h\|_2$,

$$\langle Q,\, \mathcal{T}(Q)\rangle \;=\; \sum_{i=1}^h \alpha_i^2 \beta_i^2. \tag{2}$$

Step 4: weighted AM-GM. Set $T_i := \alpha_i^2 \beta_i^2$ and weights $w_i := r^{h-i}$, with $\sum_{i=1}^h w_i = \kappa$. The weighted arithmetic-geometric mean inequality

$$\sum_{i=1}^h T_i \;=\; \sum_{i=1}^h w_i \cdot \frac{T_i}{w_i} \;\geq\; \kappa \prod_{i=1}^h \Big(\frac{T_i}{w_i}\Big)^{w_i/\kappa}$$

rearranges to

$$\sum_{i=1}^h T_i \;\geq\; \kappa \left(\prod_{i=1}^h w_i^{-w_i}\right)^{1/\kappa} \left(\prod_{i=1}^h T_i^{w_i}\right)^{1/\kappa}. \tag{3}$$

We compute the two product factors separately.

The weight product. Since $w_h = 1$, the factor $w_h^{-w_h} = 1$ contributes trivially. For $i \in \{1, \dots, h-1\}$,

$$\prod_{i=1}^{h-1} w_i^{-w_i} \;=\; \prod_{i=1}^{h-1} (r^{h-i})^{-r^{h-i}} \;=\; r^{-\sum_{i=1}^{h-1} (h-i)\, r^{h-i}} \;=\; r^{-\sum_{j=1}^{h-1} j\, r^j} \;=\; r^{-\rho}.$$

Hence the first factor in (3) is $r^{-\rho/\kappa}$.

The objective product. Using the index shift $w_{i+1} = w_i / r$ for $i \in \{1, \dots, h-1\}$, we reorganize the product to expose the boundary terms:

$$\prod_{i=1}^h T_i^{w_i} \;=\; \prod_{i=1}^h \alpha_i^{2w_i}\, \beta_i^{2w_i} \;=\; \alpha_h^{2w_h}\, \beta_1^{2w_1} \prod_{i=1}^{h-1} \alpha_i^{2w_i}\, \beta_{i+1}^{2w_{i+1}} \;=\; \alpha_h^{2w_h}\, \beta_1^{2w_1} \prod_{i=1}^{h-1} \big(\alpha_i\, \beta_{i+1}^{1/r}\big)^{2w_i}.$$

The boundary values are $\beta_1 = \|z_0\|_2 = \|x\|_2 = 1$ and $\alpha_h = \|I_{d_h}\, u\|_2 = 1$, which eliminate the first two factors.

Step 5: the propagation identity. For $i \in \{1, \dots, h-1\}$, the backpropagated Jacobian satisfies $A_i^h\, z_i = r^{h-i}\, z_h$. To see this, expand at the layer just above i:

$$A_i^h\, z_i = W_h D_h \cdots W_{i+1}\, D_{i+1}\, z_i = W_h D_h \cdots W_{i+1}\, r\, \mathrm{diag}\big(z_i^{\circ(r-1)}\big)\, z_i = r\, W_h D_h \cdots W_{i+1}\, z_i^{\circ r}.$$

Recognizing $W_{i+1}\, z_i^{\circ r} = z_{i+1}$ gives $A_i^h\, z_i = r\, A_{i+1}^h\, z_{i+1}$, and iterating yields $A_i^h\, z_i = r^{h-i}\, z_h$ as claimed.

By Cauchy-Schwarz,

$$\alpha_i\, \|z_i\|_2 \;=\; \|(A_i^h)^\top u\|_2 \cdot \|z_i\|_2 \;\geq\; u^\top A_i^h\, z_i \;=\; r^{h-i}\, u^\top z_h.$$

By definition, $\beta_{i+1} = \|z_i^{\circ r}\|_2 = \|z_i\|_{2r}^r$, hence $\beta_{i+1}^{1/r} = \|z_i\|_{2r}$. The L_p-norm monotonicity inequality on $\mathbb{R}^{d_i}$ gives

$$\|z_i\|_{2r} \;\geq\; d_i^{(1-r)/(2r)}\, \|z_i\|_2.$$

Combining,

$$\alpha_i\, \beta_{i+1}^{1/r} \;\geq\; d_i^{(1-r)/(2r)}\, \alpha_i\, \|z_i\|_2 \;\geq\; r^{h-i}\, d_i^{(1-r)/(2r)}\, u^\top z_h. \tag{4}$$

Step 6: assembling the objective product. Squaring (4) and raising to the w_i-th power,

$$\big(\alpha_i\, \beta_{i+1}^{1/r}\big)^{2w_i} \;\geq\; r^{2(h-i)\, w_i}\, d_i^{(1-r)\, w_i/r}\, (u^\top z_h)^{2w_i}.$$

Taking the product over $i = 1, \dots, h-1$ and using $\sum_{i=1}^{h-1} 2(h-i)\, w_i = 2\rho$ together with $\sum_{i=1}^{h-1} w_i = \kappa - 1$,

$$\prod_{i=1}^{h-1} \big(\alpha_i\, \beta_{i+1}^{1/r}\big)^{2w_i} \;\geq\; r^{2\rho} \left(\prod_{i=1}^{h-1} d_i^{(1-r)\, w_i/r}\right) (u^\top z_h)^{2(\kappa-1)}.$$

Raising to the $1/\kappa$ power and recognizing the d_i exponents as the definition of δ,

$$\left(\prod_{i=1}^h T_i^{w_i}\right)^{1/\kappa} \;\geq\; r^{2\rho/\kappa}\, \delta\, (u^\top z_h)^{2(\kappa-1)/\kappa}.$$

Step 7: combining the factors. Substituting into (3),

$$\sum_{i=1}^{h} T_i \;\geq\; \kappa\, r^{-\rho/\kappa}\, r^{2\rho/\kappa}\, \delta\, (u^\top z_h)^{2(\kappa-1)/\kappa} \;=\; \kappa\, r^{\rho/\kappa}\, \delta\, (u^\top z_h)^{2(\kappa-1)/\kappa}.$$

The forward-pass relation established in the proof of Lemma 9 gives $z_h = M_h\, v_h$, and combining with $\mu_{r,d}(\theta^*) = \mathrm{SymRow}(M_h)$ together with the symmetry of v_h yields $u^\top z_h = u^\top \mu_{r,d}(\theta^*)\, x^{\otimes r^{h-1}}$. Together with (2),

$$\sup_{\|U\|=1} \big(u^\top \mathcal{K}(U)\, v_h\big)^2 \;=\; \langle Q, \mathcal{T}(Q)\rangle \;=\; \sum_{i=1}^{h} T_i \;\geq\; \kappa\, r^{\rho/\kappa}\, \delta\, \big(u^\top \mu_{r,d}(\theta^*)\, x^{\otimes r^{h-1}}\big)^{2(\kappa-1)/\kappa},$$

which is the claim. □

B.2 Column-Space Projection

We now establish the column-space projection step that converts the Frobenius norm bound on the lifted Gram matrix into a bound on the unlifted Hadamard Gram.

Lemma 11 (Hadamard-Gram identity for the lifted data matrix)**.** *Let $X \in \mathbb{R}^{d_0 \times N}$ and let $X^{\uparrow r^{h-1}} \in \mathbb{R}^{d_0^{r^{h-1}} \times N}$ have columns $x_i^{\otimes r^{h-1}}$. Then*

$$\big(X^{\uparrow r^{h-1}}\big)^\top X^{\uparrow r^{h-1}} \;=\; (X^\top X)^{\circ r^{h-1}}.$$

Proof. The (i,j) entry of the left-hand side is

$$\big\langle x_i^{\otimes r^{h-1}},\, x_j^{\otimes r^{h-1}} \big\rangle \;=\; \langle x_i, x_j\rangle^{r^{h-1}},$$

using the multilinearity of the inner product on tensor products. The right-hand side has (i,j) entry $(X^\top X)_{(i,j)}^{r^{h-1}} = \langle x_i, x_j\rangle^{r^{h-1}}$ by the definition of the Hadamard power. □

Lemma 12 (Column-space projection)**.** *Let $P \in \mathbb{R}^{m\times N}$ be a matrix and let $\Pi_P := PP^+$ denote the orthogonal projector onto the column space of P. For any matrix $\mathcal{K}$ of compatible dimensions,*

$$\|\mathcal{K}\, P\|_F^2 \;\geq\; \sigma_{\min}^+(P)^2\, \|\mathcal{K}\, \Pi_P\|_F^2.$$

Proof. The inequality $PP^\top \succeq \sigma_{\min}^+(P)^2\, \Pi_P$ holds because $PP^\top$ acts as zero on $(\mathrm{range}\, P)^\perp$ and as a positive operator with smallest eigenvalue $\sigma_{\min}^+(P)^2$ on $\mathrm{range}\, P$, which Π_P does not exceed in either subspace. Hence

$$\|\mathcal{K}\, P\|_F^2 \;=\; \mathrm{tr}\big(\mathcal{K}\, PP^\top\, \mathcal{K}^\top\big) \;\geq\; \sigma_{\min}^+(P)^2\, \mathrm{tr}\big(\mathcal{K}\, \Pi_P\, \mathcal{K}^\top\big) \;=\; \sigma_{\min}^+(P)^2\, \|\mathcal{K}\, \Pi_P\|_F^2,$$

where the inequality uses the monotonicity of the trace under the Loewner order on positive semidefinite matrices and the cyclic property of trace. □

B.3 Proof of Theorem 3

We assemble the per-point lower bound and the column-space projection into a proof of the main theorem. We begin with a lemma which connects the set $\mathcal{X} := \{x \in \mathbb{R}^{d_0} : x^{\otimes r^{h-1}} \in \mathrm{range}(X^{\uparrow r^{h-1}})\}$ with the column-space projection in the previous section and with the training data.

Lemma 13 (**Lifted data span**)**.** *Let $P := X^{\uparrow r^{h-1}}$ and $\mathcal{X}$ as above.*

1. *For every $x \in \mathcal{X}$ with $\|x\|_2 = 1$, the lifted input $v := x^{\otimes r^{h-1}}$ satisfies $\Pi_P v = v$ and $\|v\|_2 = 1$.*
2. *For every i with $x_i \neq 0$, the normalized data point $\hat{x}_i := x_i/\|x_i\|_2$ belongs to $\mathcal{X}$.*
3. *At every interpolating minimum θ^* and for every $x \in \mathcal{X}$, $p_{\theta^*}(x) = YP^+ x^{\otimes r^{h-1}}$. In particular, the label-energy factor of Theorem 3 depends only on the data (X, Y) and not on the choice of interpolating minimum.*

Proof. 1. Since Π_P is the orthogonal projector onto $\operatorname{range} P$ and $v \in \operatorname{range} P$ by the definition of $\mathcal{X}$, we have $\Pi_P v = v$. By the multilinearity of the inner product on tensor products $\|v\|_2^2 = \langle x, x\rangle^{r^{h-1}} = \|x\|_2^{2r^{h-1}} = 1$.

2. We have $\hat{x}_i^{\otimes r^{h-1}} = \|x_i\|_2^{-r^{h-1}} x_i^{\otimes r^{h-1}} = \|x_i\|_2^{-r^{h-1}} P e_i \in \operatorname{range} P$, where e_i is the i-th standard basis vector of $\mathbb{R}^N$.

3. For $x \in \mathcal{X}$ the lifted input lies in $\operatorname{range} P$ so $x^{\otimes r^{h-1}} = \Pi_P x^{\otimes r^{h-1}} = PP^+ x^{\otimes r^{h-1}}$. Interpolation gives $\mu_{r,d}(\theta^*)P = Y$, whence $p_{\theta^*}(x) = \mu_{r,d}(\theta^*) x^{\otimes r^{h-1}} = \mu_{r,d}(\theta^*) PP^+ x^{\otimes r^{h-1}} = YP^+ x^{\otimes r^{h-1}}$.

□

Proof of Theorem 3. Let θ^* be an interpolating minimum of the averaged empirical loss $L(\theta) = N^{-1} \|Y - \mu_{r,d}(\theta)\, X^{\uparrow r^{h-1}}\|_F^2$, and write $P := X^{\uparrow r^{h-1}}$ throughout this proof for the lifted data matrix.

Step 1: Hessian collapse. The residual $Y - \mu_{r,d}(\theta^*)\, P$ vanishes by interpolation, so the second derivative of L collapses to twice its Gauss-Newton component. For any tangent direction U,

$$\nabla^2 L(\theta^*)[U, U] \;=\; \frac{2}{N} \left\| \mathcal{K}(U)\, P \right\|_F^2,$$

where $\mathcal{K} := D_{\theta^*} \mu_{r,d}$.

Step 2: column-space projection. Lemma 12 gives

$$\left\| \mathcal{K}(U)\, P \right\|_F^2 \;\geq\; \sigma_{\min}^+(P)^2 \left\| \mathcal{K}(U)\, \Pi_P \right\|_F^2.$$

Lemma 11 identifies $\sigma_{\min}^+(P)^2 = \lambda_{\min}^+\big(P^\top P\big) = \lambda_{\min}^+\big((X^\top X)^{\circ r^{h-1}}\big)$.

Step 3: pointwise restriction. The Frobenius norm dominates the operator norm:

$$\left\| \mathcal{K}(U)\, \Pi_P \right\|_F^2 \;\geq\; \sup_{\|v\|_2 = 1} \left\| \mathcal{K}(U)\, \Pi_P\, v \right\|_2^2.$$

By Lemma 13, every unit input $x \in \mathcal{X}$ defines a unit vector $v := x^{\otimes r^{h-1}}$ that lies in $\operatorname{range} P$, so $\Pi_P v = v$. Restricting the supremum to this family and using $\|w\|_2^2 \geq (u^\top w)^2$ for any unit vector $u \in \mathbb{R}^{d_h}$,

$$\sup_{\|v\|_2 = 1} \left\| \mathcal{K}(U)\, \Pi_P\, v \right\|_2^2 \;\geq\; \sup_{\substack{\|u\|_2 = \|x\|_2 = 1 \\ x \in \mathcal{X}}} \left(u^\top \mathcal{K}(U)\, x^{\otimes r^{h-1}} \right)^2.$$

Step 4: pushing the supremum. Suprema commute, hence

$$\sup_{\|U\|=1} \sup_{\substack{\|u\|_2 = \|x\|_2 = 1 \\ x \in \mathcal{X}}} (u^\top \mathcal{K}(U) x^{\otimes r^{h-1}})^2 = \sup_{\substack{\|u\|_2 = \|x\|_2 = 1 \\ x \in \mathcal{X}}} \sup_{\|U\|=1} (u^\top \mathcal{K}(U) x^{\otimes r^{h-1}})^2.$$

Step 5: invocation of the per-point bound. Proposition 10 holds for arbitrary unit vectors u and x, so it can be invoked pointwise at every pair in the family of Step 3 rather than only at the training pairs. Together with $p_{\theta^*}(x) = \mu_{r,d}(\theta^*) x^{\otimes r^{h-1}}$, it gives

$$\sup_{\|U\|=1} (u^\top \mathcal{K}(U) x^{\otimes r^{h-1}})^2 \geq \kappa r^{\rho/\kappa} \delta (u^\top \mu_{r,d}(\theta^*) x^{\otimes r^{h-1}})^{2(\kappa-1)/\kappa} = \kappa r^{\rho/\kappa} \delta \langle u, p_{\theta^*}(x)\rangle^{2\frac{\kappa-1}{\kappa}}.$$

Step 6: assembly. Combining Steps 1–5 with $\lambda_{\max}(\nabla^2 L(\theta^*)) = \sup_{\|U\|=1} \nabla^2 L(\theta^*)[U, U]$,

$$\lambda_{\max}(\nabla^2 L(\theta^*)) \geq \frac{2}{N} \lambda_{\min}^+((X^\top X)^{\circ r^{h-1}}) \cdot \left(\kappa r^{\rho/\kappa} \delta \right) \sup_{\substack{\|u\|_2 = \|x\|_2 = 1 \\ x \in \mathcal{X}}} \langle u, p_{\theta^*}(x)\rangle^{2\frac{\kappa-1}{\kappa}},$$

which is the bound in Theorem 3.

Step 7: the datapoint specialization. By Lemma 13, each $\hat{x}_i = x_i/\|x_i\|_2$ belongs to $\mathcal{X}$. For every i with $y_i \neq 0$, set $u_i := y_i/\|y_i\|_2$; data points with $y_i = 0$ contribute zero to the maximum below. Interpolation gives $\mu_{r,d}(\theta^*)x_i^{\otimes r^{h-1}} = y_i$, so

$$\langle u_i, p_{\theta^*}(\hat{x}_i)\rangle = u_i^\top \mu_{r,d}(\theta^*)\hat{x}_i^{\otimes r^{h-1}} = \frac{y_i^\top y_i}{\|y_i\|_2\|x_i\|_2^{r^{h-1}}} = \frac{\|y_i\|_2}{\|x_i\|_2^{r^{h-1}}}.$$

Evaluating the supremum of Step 6 at the pairs $(u_i, \hat{x}_i)$ and maximizing over i

$$\lambda_{\max}(\nabla^2 L(\theta^*)) \geq \frac{2}{N}\lambda_{\min}^+((X^\top X)^{\circ r^{h-1}}) \cdot \left(\kappa r^{\rho/\kappa}\delta\right) \max_{1\leq i\leq N}\left(\frac{\|y_i\|_2}{\|x_i\|_2^{r^{h-1}}}\right)^{2\frac{\kappa-1}{\kappa}},$$

which is the computable form of the label-energy factor used in Theorem 1. □

B.4 Proof of Corollary 4

The bound treats the two dimensions of the data differently: the number of data points N enters explicitly through $2/N$ and $(N-1)\mu^{r^{h-1}}$, while the input dimension d_0 enters only through the data-geometry factor, and the architecture factor contains the hidden widths alone. For isotropic random inputs, the coherence concentrates and the data-geometry factor of Theorem 1 is bounded below by a constant with high probability.

The proof below combines a spherical-cap estimate with a union bound over pairs of data points to obtain $\mu < \sqrt{2(\gamma+2)\log N/d_0}$, so that $1-(N-1)\mu^{r^{h-1}} \geq \frac{1}{2}$ in the data-geometry factor of Theorem 1. The dimension requirement weakens rapidly with depth, since the exponent $2/r^{h-1}$ decays geometrically in h. The high-dimensional message is therefore favorable: as d_0 grows, the coherence vanishes, the term $1-(N-1)\mu^{r^{h-1}}$ tends to 1, and the bound concentrates on the architecture and label-energy factors, so the r^{-2h} law is cleanest in high dimension. A careful analysis under general, non-isotropic designs requires more care and is left for future work.

Proof of Corollary 4. Write $k := r^{h-1}$. The teacher θ_T interpolates the data, so Theorem 1 applies. Since $\|x_i\|_2 = 1$ for all i, its data-geometry factor equals $\frac{2}{N}\big(1-(N-1)\mu^k\big)$ and its label-energy factor equals $\max_i \|y_i\|_2^{2(\kappa-1)/\kappa}$. It therefore suffices to show that $(N-1)\mu^k \leq \frac{1}{2}$ with probability at least $1-N^{-\gamma}$.

Fix $i \neq j$. Conditionally on x_j, the rotation invariance of the uniform distribution on $\mathbb{S}^{d_0-1}$ implies that $\langle x_i, x_j\rangle$ is distributed as the first coordinate of a uniform point on $\mathbb{S}^{d_0-1}$, and the spherical-cap estimate [Ball, 1997, Lemma 2.2] gives, for $0 \leq t < 1$,

$$\mathbb{P}\big(|\langle x_i, x_j\rangle| \geq t\big) \;\leq\; 2\exp\big(-d_0 t^2/2\big).$$

A union bound over the $N(N-1)/2$ pairs gives $\mathbb{P}(\mu \geq t) \leq N^2\exp(-d_0t^2/2)$. We choose $t^2 := 2(\gamma+2)\log N/d_0$. The assumption on d_0 gives $t^2 \leq \frac{1}{4}N^{-2/k} < 1$, and $\mathbb{P}(\mu \geq t) \leq N^2\,N^{-(\gamma+2)} = N^{-\gamma}$. On the complementary event,

$$(N-1)\,\mu^k \;<\; (N-1)\,(t^2)^{k/2} \;\leq\; (N-1)\Big(\tfrac{1}{4}N^{-2/k}\Big)^{k/2} \;=\; \frac{N-1}{N}\,2^{-k} \;<\; \frac{1}{2},$$

so the data-geometry factor is at least $\frac{2}{N}\cdot\frac{1}{2} = \frac{1}{N}$, and Theorem 1 gives the claim. □

B.5 Further Remarks on the Sharpness Bound

This section contains the full versions of the remarks following Theorem 3 in Section 2.1.

Interplay of depth, degree, and width. The three architecture constants play distinct roles. Write $w_i := r^{h-i}$ for the degree of homogeneity of p_θ in the weight matrix W_i alone. Then $\kappa = \sum_{i=1}^h w_i$ is the total degree of homogeneity of p_θ in the parameters, and $r^{\rho/\kappa} = \prod_{i=1}^h w_i^{w_i/\kappa}$ is the geometric mean of the layerwise degrees, weighted by w_i/κ. Both are of order r^h, which produces the r^{2h} growth. The hidden widths enter only through δ, and with a negative exponent: width plays a

flattening role, while depth sharpens the bound exponentially. The exponent of d_i in δ is $-\frac{r-1}{r}\cdot\frac{r^{h-i}}{\kappa}$, which decays geometrically in i, so δ is governed by the hidden layers closest to the input. With a common hidden width $d_1 = \cdots = d_{h-1} = d$, the identity $\sum_{i=1}^{h-1} r^{h-i} = \kappa - 1$ gives

$$\delta \;=\; d^{-\frac{r-1}{r}\cdot\frac{\kappa-1}{\kappa}}, \qquad \text{hence} \qquad \kappa\, r^{\rho/\kappa}\,\delta \;\asymp\; r^{2h}\, d^{-\frac{r-1}{r}}$$

in the deep limit: exponential growth in depth against sub-linear polynomial decay in width. Offsetting a single added layer therefore requires multiplying the width by approximately $r^{2r/(r-1)}$ (that is, by 16 for $r = 2$ and by 27 for $r = 3$), and keeping the architecture factor bounded as the depth grows requires $d \gtrsim r^{2hr/(r-1)}$, a width exponential in depth. At $r = 1$ the exponent vanishes, $\delta = 1$, and the width plays no role, as in the deep linear case. Since Theorem 3 is a lower bound, large width lowers the floor on the sharpness; whether the minima of wide networks are actually flatter would require a matching upper bound, which remains open for $r \geq 2$.

Normalization. The r^{2h} law is stated at fixed data and labels, while the data-geometry factor contains terms such as $\|x_i\|_2^{2r^{h-1}}$, which grow much faster than r^{2h} under rescaling of the inputs. These terms, however, enter both sides of the inequality by exactly the same amount. Write $L_{X,Y}$ for the empirical risk on the dataset (X, Y). Under $(X, Y) \mapsto (cX, c^{r^{h-1}}Y)$, $c > 0$, the r^{h-1}-homogeneity $p_\theta(cx) = c^{r^{h-1}} p_\theta(x)$ leaves the set of interpolating minima unchanged, and $L_{cX,c^{r^{h-1}}Y}(\theta) = c^{2r^{h-1}} L_{X,Y}(\theta)$ for every θ, hence

$$\lambda_{\max}\big(\nabla^2 L_{cX,c^{r^{h-1}}Y}(\theta^*)\big) \;=\; c^{2r^{h-1}}\, \lambda_{\max}\big(\nabla^2 L_{X,Y}(\theta^*)\big).$$

The bounds of Theorems 1 and 3 transform by the same factor $c^{2r^{h-1}}$: the terms $\lambda^+_{\min}((X^\top X)^{\circ r^{h-1}})$, $\min_i \|x_i\|_2^{2r^{h-1}}$, and $(N-1)\mu^{r^{h-1}}$ each pick up $c^{2r^{h-1}}$, while the label-energy factor is invariant. The normalization therefore sets the units in which the Hessian is measured, and the ratio between the bound and the true sharpness does not depend on it. Throughout, the r^{2h} law for the sharpness and the r^{-2h} law for the step-size refer to the convention $\|x_i\|_2 = 1$ for all i and $\max_i \|y_i\|_2 = 1$, which is also the convention of our experiments (Appendix D).

Homogeneity of the Hessian. For an input x, write $g_x(\theta) := p_\theta(x)$. Scaling a single layer $W_i \mapsto tW_i$, $t > 0$, scales the pre-activation of layer i by t, and each of the subsequent $h-i$ monomial activations raises the scaling factor to the r-th power; hence g_x is homogeneous of degree r^{h-i} in W_i, and scaling all layers jointly gives $g_x(t\theta) = t^\kappa g_x(\theta)$ with $\kappa = \sum_{i=1}^h r^{h-i}$. Differentiating this identity in θ gives $D_{t\theta}\, g_x(U) = t^{\kappa-1} D_\theta\, g_x(U)$ for every direction U. At an interpolating minimum the Hessian reduces to its Gauss-Newton component (Step 1 in Appendix B.3),

$$\nabla^2 L(\theta^*)[U, U] \;=\; \frac{2}{N}\sum_{i=1}^{N} \big\| D_{\theta^*}\, g_{x_i}(U)\big\|_2^2,$$

and the right-hand side, as a function of the base point θ^*, is homogeneous of degree $2(\kappa - 1)$. Since $t\theta^*$ is an interpolating minimum for the rescaled labels $t^\kappa Y$, and $\theta \mapsto t\theta$ is a bijection between the interpolating minima for Y and for $t^\kappa Y$, we obtain, writing $s := t^\kappa$,

$$\lambda_{\max}\big(\nabla^2 L_{X,sY}(s^{1/\kappa}\theta^*)\big) \;=\; s^{2\frac{\kappa-1}{\kappa}}\, \lambda_{\max}\big(\nabla^2 L_{X,Y}(\theta^*)\big).$$

In Theorems 1 and 3 the labels enter only through the label-energy factor, which picks up exactly the same factor $s^{2\frac{\kappa-1}{\kappa}}$, so both sides of the bound transform identically under rescaling of the labels. Conversely, the homogeneity dictates this exponent: a bound of the form $C \cdot E(Y)^a$, with $C > 0$ independent of Y and $E(sY) = s^2 E(Y) > 0$, can hold for all $s > 0$ only if $a = \frac{\kappa-1}{\kappa}$, since otherwise it is violated as $s \to 0$ or as $s \to \infty$. The layerwise degrees r^{h-i} are also the weights of the weighted AM-GM step in the proof of Proposition 10, which is how the homogeneity structure enters the architecture factor $\kappa\, r^{\rho/\kappa}\delta$.

Beyond the quadratic loss. The proof of Theorem 3 uses the quadratic loss only through the Gauss-Newton form of the Hessian at interpolation, and therefore extends to losses that attain an interpolating minimum at finite parameters and are locally strongly convex there. Let $\ell(\hat{y}, y)$ be twice differentiable in its first argument with $\nabla_{\hat{y}}\ell(y, y) = 0$, let $L_\ell(\theta) := \frac{1}{N}\sum_{i=1}^N \ell(p_\theta(x_i), y_i)$, and let

θ^* be an interpolating minimum. The term containing the second derivatives of g_{x_i} is multiplied by $\nabla_{\hat{y}}\ell(y_i, y_i) = 0$, and the Hessian reduces to the Gauss-Newton form

$$\nabla^2 L_\ell(\theta^*)[U, U] \;=\; \frac{1}{N}\sum_{i=1}^{N} \big(D_{\theta^*}\, g_{x_i}(U)\big)^\top\, \nabla^2_{\hat{y}}\ell(y_i, y_i)\, D_{\theta^*}\, g_{x_i}(U).$$

If $c := \min_i \lambda_{\min}\big(\nabla^2_{\hat{y}}\ell(y_i, y_i)\big) > 0$, every step of the proof of Theorem 3 carries through with the quadratic-loss constant 2 replaced by c, since the argument uses only this quadratic form together with the differential of Theorem 5. Consequently, Theorems 1 and 3, and with them Corollary 2, hold with the prefactor $2/N$ replaced by c/N. For the quadratic loss, $\nabla^2_{\hat{y}}\ell = 2I$ and $c = 2$; for the Huber loss, $c = 1$.

C Critical Step-Sizes for Optimizer

This appendix establishes Corollary 2 of the main text. Section C.1 introduces the Chemnitz and Engel [2025] framework as it applies to deterministic first-order optimization, including the precise formulation of the linearization and the spectral-radius criterion for local stability. Section C.2 introduces the stability bounds computed by Cohen et al. [2022a] and uses them to derive the bounds in Corollary 2.

We use throughout the notation introduced in Section 2 of the main text.

C.1 The Chemnitz-Engel Framework for Deterministic Optimizers

A first-order optimizer with step size $\eta > 0$ acting on parameter space $\Theta = \prod_{k=1}^{h} \mathbb{R}^{d_k \times d_{k-1}}$ is a smooth map

$$G_\eta : \Theta \times \mathcal{B} \to \Theta \times \mathcal{B},$$

where $\mathcal{B}$ is an auxiliary state space encoding the optimizer's internal memory. For methods without auxiliary state, such as full-batch gradient descent, $\mathcal{B}$ is trivial and G_η acts on Θ alone. A point $(\theta^*; \beta^*)$ is a fixed point of G_η if $G_\eta(\theta^*; \beta^*) = (\theta^*; \beta^*)$, which corresponds to θ^* being a critical point of the loss with the auxiliary state at its optimizer-specific steady-state value.

Definition 14 (Local linear stability)**.** A fixed point $(\theta^*; \beta^*)$ of G_η is locally linearly stable if for any $\epsilon > 0$ there exists a $\delta > 0$ and an $N > 0$ such that for all $n > N$ and all $(\theta^0; \beta^0)$ such that $\|(\theta^0; \beta^0) - (\theta^*; \beta^*)\| < \delta$ we have $\|G^n_\eta(\theta^0; \beta^0) - (\theta^*; \beta^*)\| < \epsilon$. In other words, all iterates of G_η starting from a sufficiently small neighborhood of $(\theta^*; \beta^*)$ remain in an arbitrarily small neighborhood of $(\theta^*; \beta^*)$ for all sufficiently large n.

Chemnitz and Engel [2025] have shown local linear stability of a fixed point $(\theta^*; \beta^*)$ can be determined by examining the eigenvalues of the Jacobian $DG_\eta(\theta^*; \beta^*)$ of G_η at $(\theta^*; \beta^*)$. The necessary condition is that the there exists at least one eigenvalue of the Jacobian with magnitude less than one, and all the others are either less than it or equal to one in absolute value. Thus, we denote by $\rho^{\neq 1}(DG_\eta)$ the greatest eigenvalue of the Jacobian in absolute value that is not equal to one.

Lemma 15 (Chemnitz and Engel [2025, Section 2.3])**.** *A point $(\theta^*; \beta^*)$ is local linearly stable if $\rho^{\neq 1}(DG_\eta) < 1$; if $\rho^{\neq 1}(DG_\eta) > 1$ then it is unstable: there exist initial conditions arbitrarily close to $(\theta^*; \beta^*)$ whose iterates under G_η do not converge to $(\theta^*; \beta^*)$.*

C.2 Optimizer Corollaries

For each of the four optimizers, we recall the local linear stability bounds derived in Cohen et al. [2022b,a] and use them to derive the step-size ceilings stated in Corollary 2.

C.2.1 Full-Batch Gradient Descent

Cohen et al. [2022b] show that a critical point θ^* is locally linearly stable under full-batch gradient descent with step size η if $\lambda_{\max}(\nabla^2 L(\theta^*) \leq \frac{2}{\eta}$. This gives the upper bound on the critical step-size

$$\eta \;<\; \eta_{\text{crit}} \coloneqq \frac{2}{\lambda^*}.$$

C.3 Exponential-Moving-Average Heavy Ball

Cohen et al. [2022a] show that a critical point θ^* is locally linearly stable under full-batch gradient descent with step size η if

$$\lambda_{\max}\big(\nabla^2 L(\theta^*)\big) \;<\; \frac{2+2\beta_1}{\eta\,(1-\beta_1)}.$$

Solving for η,

$$\eta_{\mathrm{crit}}(\lambda;\beta_1) \;=\; \frac{2+2\beta_1}{(1-\beta_1)\,\lambda}.$$

At $\beta_1 = 0$ this recovers the gradient-descent ceiling $2/\lambda^*$.

C.4 Preconditioned Heavy Ball

Cohen et al. [2022a] extend their analysis to this setting as well, showing that the stability condition becomes

$$\lambda_{\max}\big(P^{-1}\,\nabla^2 L(\theta^*)\big) \;<\; \frac{2+2\beta_1}{\eta\,(1-\beta_1)}.$$

The condition is stated in terms of the preconditioned Hessian. To express it as a scalar inequality in $\lambda_{\max}\big(\nabla^2 L(\theta^*)\big)$, we use Ostrowski's inequality.

Lemma 16 (Ostrowski Horn and Johnson [1985, Theorem 4.5.9])**.** *Let H be symmetric and P symmetric positive definite, both of size $n \times n$. Then*

$$\lambda_{\max}(P^{-1}H) \;\geq\; \frac{\lambda_{\max}(H)}{\lambda_{\max}(P)}.$$

Proof. The matrix $P^{-1}H$ is similar to $P^{-1/2}HP^{-1/2}$, which is symmetric, so the two share spectra. By the variational characterization,

$$\lambda_{\max}(P^{-1/2}HP^{-1/2}) \;=\; \max_{u\neq 0}\frac{u^\top P^{-1/2}HP^{-1/2}u}{u^\top u} \;=\; \max_{v\neq 0}\frac{v^\top Hv}{v^\top Pv},$$

where $v := P^{-1/2}u$ ranges over nonzero vectors. The denominator is at most $\lambda_{\max}(P)\,v^\top v$, so

$$\lambda_{\max}(P^{-1/2}HP^{-1/2}) \;\geq\; \max_{v\neq 0}\frac{v^\top Hv}{\lambda_{\max}(P)\,v^\top v} \;=\; \frac{\lambda_{\max}(H)}{\lambda_{\max}(P)}.\;\square$$

□

Applying Lemma 16 with $H = \nabla^2 L(\theta^*)$ and the architectural lower bound $\lambda^* \leq \lambda_{\max}\big(\nabla^2 L(\theta^*)\big)$,

$$\frac{\lambda^*}{\lambda_{\max}(P)} \;\leq\; \lambda_{\max}\big(P^{-1}\,\nabla^2 L(\theta^*)\big).$$

Combining with the Cohen et al. [2022a] condition gives the required scalar inequality:

$$\rho<1 \;\Longrightarrow\; \frac{\lambda^*}{\lambda_{\max}(P)} \;<\; \frac{2+2\beta_1}{\eta\,(1-\beta_1)} \;\Longleftrightarrow\; \eta \;<\; \frac{\lambda_{\max}(P)\,(2+2\beta_1)}{(1-\beta_1)\,\lambda^*}.$$

The threshold is then

$$\eta_{\mathrm{crit}}(\lambda;\beta_1,P) = \lambda_{\max}(P)\,(2+2\beta_1)/((1-\beta_1)\,\lambda).$$

At $P = I$ this reduces to the unpreconditioned heavy-ball ceiling.

C.5 Adam at Interpolation

Adam maintains a first-moment EMA m_t and a second-moment EMA v_t, with update

$$m_{t+1} \;=\; \beta_1\, m_t + (1-\beta_1)\,\nabla L(\theta_t),\; v_{t+1} \;=\; \beta_2\, v_t + (1-\beta_2)\,\nabla L(\theta_t)^{\circ 2},\; \theta_{t+1} \;=\; \theta_t - \eta\,\frac{m_{t+1}}{\sqrt{v_{t+1}}+\epsilon}.$$

Hyperparameters are $\beta_1, \beta_2 \in [0, 1)$ and $\epsilon > 0$.

At an interpolating minimum, the gradient vanishes, and the steady-state values are $m^* = 0$, $v^* = 0$. The effective preconditioner at the steady state is $\left(\sqrt{v^*}+\epsilon\right)^{-1} = \epsilon^{-1}\, I$. The function $v \mapsto 1/(\sqrt{v}+\epsilon)$ is smooth at $v = 0$ for $\epsilon > 0$, with Taylor expansion $\epsilon^{-1} + O(v^{1/2})$, so the linearization of the Adam update at the fixed point coincides with the linearization of preconditioned heavy ball with $P = \epsilon\, I$ to leading order. Higher-order corrections from the v-derivative do not affect the stability threshold.

Substituting $\lambda_{\max}(P) = \epsilon$ into the preconditioned heavy-ball ceiling produces the threshold

$$\eta_{\mathrm{crit}}(\lambda; \beta_1, \epsilon) \;=\; \frac{\epsilon\,(2+2\beta_1)}{(1-\beta_1)\,\lambda}.$$

This is an asymptotic statement: it characterizes the linear stability of an interpolating fixed point of Adam in the limit $v_t \to 0$. It does not bound the maximum step size at which Adam reaches such a fixed point from a generic initialization, since during the transient v_t is generally far from zero and the effective preconditioner is correspondingly larger.

C.6 Verification of the r^{-2h} Scaling

Substituting the architectural lower bound from Theorem 3 into each of the four ceilings produces a step-size upper bound that, for fixed widths and dataset, scales as the inverse of the architectural component $\kappa\, r^{\rho/\kappa}\, \delta$ of the bound. The depth-degree scaling paragraph of Section 2.1 establishes that this component grows as r^{2h} in the deep limit, so the step-size ceiling shrinks as r^{-2h}, with each added monomial layer multiplying the ceiling by approximately r^{-2}. The optimizer-specific prefactors $(2+2\beta_1)/(1-\beta_1)$, $\lambda_{\max}(P)\,(2+2\beta_1)/(1-\beta_1)$, and $\epsilon\,(2+2\beta_1)/(1-\beta_1)$ are independent of depth and degree, so the depth-degree dependence is preserved across the four optimizers.

This completes the proof of Corollary 2.

D Experimental Details

This appendix describes the experimental protocol used to validate Theorem 1 and Corollary 2. We organize the appendix around the quantitative claims being tested: the step-size ceiling per optimizer (Section D.5) and the tightness of the sharpness lower bound against the empirical flattest-minimum sharpness (Section D.7).

D.1 Overall protocol

For each experimental configuration (r, h, d), we follow a six-step protocol. We sample a target parameter θ^* from a Glorot-normal initialization; we sample N training inputs $\{x_i\}_{i=1}^N$ with $N = 2\,\dim \mathcal{V}_{r,d}$, exceeding the threshold for unique functional identifiability[Usevich et al., 2026]; we compute the labels $y_i := p_{\theta^*}(x_i)$, guaranteeing that θ^* is an interpolating minimum and that the global minimizer variety

$$\mathcal{M}^* := \{\theta : p_\theta(x_i) = y_i \text{ for } i = 1, \ldots, N\}$$

is uniquely determined as a function-space object; we compute the critical step-size η_{crit} from the bound of Theorem 1 using the optimizer-specific formula in Corollary 2; we run the optimizer on a logarithmic step-size grid spanning $[\eta_{\mathrm{crit}}/10^8,\; 10^2\eta_{\mathrm{crit}}]$, sweeping from the largest to the smallest step-size; and we record the distance from the trained parameter to $\mathcal{M}^*$ via a Gauss-Newton KKT projection. The resulting distance-versus-step-size curves are the primary experimental signal.

The protocol normalizes data and labels so that the bound's data-geometry and label-energy factors take controlled values, isolating the architectural component as the primary quantity. Specifically, the

input matrix X is column-normalized so that $\|x_i\|_2 = 1$ for all i, and after computing $Y = p_{\theta^*}(X)$ the output-layer weight matrix W_h^* is rescaled by $\max_i \|y_i\|_2$ so that $\max_i \|y_i\|_2 = 1$; the labels are then recomputed from the rescaled weights. This rescaling does not alter the variety $\mathcal{M}^*$ as a set, but reduces the label-energy factor in the bound to 1 and keeps the per-layer activations bounded across the depth-degree sweep.

D.2 Architecture and parameter generation

The architecture tuple is `arch` $= (d_0, d_1, \ldots, d_h)$ with d_0 the input dimension, d_h the output dimension, and $d_1 = \cdots = d_{h-1}$ the common hidden width. The trainable parameters are bias-free linear layers $W_k \in \mathbb{R}^{d_k \times d_{k-1}}$ implemented via Flux's `Dense(...; bias = false)`; the activation is the entrywise monomial z^r on hidden layers and the identity on the output layer.

The target parameter θ^* is drawn entrywise from the Glorot-normal distribution with variance $2/(d_{k-1} + d_k)$ at each layer, using a fixed `seed_theta_star`. The Glorot scale keeps the per-layer activations bounded under the polynomial squarings. The activation degree r is restricted to integer values ≥ 2. The depth h ranges over $\{2, 3\}$.

D.3 Data generation

Inputs are sampled as $X \sim \mathcal{N}(0, I_{d_0})^{\otimes N}$, then column-normalized:

$$x_i \;\longleftarrow\; x_i/\|x_i\|_2, \qquad i = 1, \ldots, N.$$

The sample size is $N = 2 \cdot \min\big(d_h \sum_{i=1}^{h}(d_i - 1)\, d_i,\; d_h \binom{d_0 + r^{h-1} - 1}{r^{h-1}}\big)$, twice the upper bound on $\dim \mathcal{V}_{r,d}$ established by Kileel et al. [2019]. The factor of two provides a safety margin for the generic identifiability conditions, and ensures that the global minimizer variety $\mathcal{M}^*$ is uniquely determined by the training data [Usevich et al., 2026].

Labels are then $Y = \mu_{r,d}(\theta^*) X^{\uparrow r^{h-1}}$, computed as the forward pass of the target network. The output layer is rescaled to enforce $\max_i \|y_i\|_2 = 1$, after which the labels are recomputed. This protocol guarantees an exact interpolating minimum (namely θ^* itself, after rescaling) with all label-energy ratios bounded by one.

D.4 Critical step-size computation

The critical step-size is computed in two stages. First, the data-free architectural component

$$S_{\text{arch}} \;:=\; 2\,\kappa\, r^{\rho/\kappa}\, \delta$$

is evaluated directly using the closed-form expressions in the notation of Section 2. Second, the data-geometry factor $\lambda_{\min}^{+}\big((X^\top X)^{\circ r^{h-1}}\big)$ is evaluated by the Gershgorin lower bound

$$\lambda_{\min}^{+}\big((X^\top X)^{\circ r^{h-1}}\big) \;\geq\; \min_i \|x_i\|_2^{2r^{h-1}} - (N-1)\,\mu^{r^{h-1}},$$

where $\mu := \max_{i \neq j} |\langle x_i, x_j \rangle|$ is the data coherence.

The data-geometry factor is computed using the arbitrary precision arithmetic of `BigFloat`.

Combining the two factors gives the GD critical step-size $\eta_{\text{crit}}^{\text{GD}} = 2/S$, where S is the bound of Theorem 1 evaluated on the normalized data. The optimizer-specific ceilings of Corollary 2 are then computed by direct application of the closed-form formulas, treating $\eta_{\text{crit}}^{\text{GD}}$ as the architectural reference value:

$$\eta_{\text{crit}}^{\text{Adam}} \;=\; \eta_{\text{crit}}^{\text{GD}} \cdot \epsilon \cdot \frac{1+\beta_1}{1-\beta_1}, \quad \eta_{\text{crit}}^{\text{EmaHB}} \;=\; \eta_{\text{crit}}^{\text{GD}} \cdot \frac{1+\beta_1}{1-\beta_1}, \quad \eta_{\text{crit}}^{\text{PrecondEmaHB}} \;=\; \eta_{\text{crit}}^{\text{GD}} \cdot \frac{1+\beta_1}{(1-\beta_1)\,\epsilon}.$$

D.5 Step-size sweep and convergence detection

For each configuration, the optimizer is run on the logarithmic grid

$$\eta \;\in\; \text{logspace}\big(\log(\eta_{\text{crit}}/10^8),\; \log(10^2 \eta_{\text{crit}});\; n_\eta\big),$$

with $n_\eta = 10$ points in the main experiments. The grid is swept from largest η to smallest η, with adaptive warm-starting between grid points: after each η completes, the mean final distance to $\mathcal{M}^*$ over the non-diverged seeds is used as the initialization radius for the next, smaller η. This initialization avoids the need for unmanageably large iteration counts at small η, where the optimizer would otherwise spend most of training traversing a fixed initialization radius.

A run is declared `:converged` if the distance to $\mathcal{M}^*$ falls below a specified tolerance on two consecutive distance checks separated by $\mathrm{stride} = \max(1, \mathrm{max_iters}/100)$ iterations; this two-check requirement guards against transient near-misses. A run is declared `:diverged` if the loss midpoint exceeds 10^6 times the initial loss or NaN/Inf appears. Otherwise the run terminates at `max_iters` and is declared `:max_iters_reached`.

The primary recorded quantity is `final_distance`, the distance to $\mathcal{M}^*$ at termination; `status` is its coarse-grained categorical label. The plot in Figure 1 of the main text is the mean of `final_distance` over $n_{\mathrm{seeds}} =$ between 20 and 100 initialization seeds (depending on architecture cost) at each η.

D.6 Distance to the global minimizer variety

The distance from a parameter θ to $\mathcal{M}^*$ is computed as the Euclidean distance from θ to its nearest point on $\mathcal{M}^*$, by solving the constrained optimization

$$\min_{\theta'} \frac{1}{2}\|\theta' - \theta\|^2 \quad \text{subject to} \quad f(\theta') := \mathrm{vec}\big(p_{\theta'}(X) - Y\big) = 0,$$

via Gauss-Newton iteration on the Karush-Kuhn-Tucker saddle-point system

$$\begin{pmatrix} 2I & J(\theta')^\top \\ J(\theta') & 0 \end{pmatrix} \begin{pmatrix} \delta\theta' \\ \delta\lambda \end{pmatrix} = \begin{pmatrix} -2(\theta' - \theta) - J(\theta')^\top \lambda \\ -f(\theta') \end{pmatrix},$$

with backtracking line search on the joint KKT residual and a small augmented-Lagrangian regularization on the (2, 2) block to handle rank-deficient Jacobians. The constraint Jacobian $J(\theta') = \partial f / \partial \theta'$ is computed by Zygote's reverse-mode automatic differentiation directly on the network; this avoids the symbolic polynomial expansion that becomes prohibitive at depth four and above.

The solver terminates when the KKT residual norm falls below a specified threshold, and reports `:high_residual` if the constraint norm exceeds the same threshold scaled by 10^4 at termination, in which case the distance is flagged as unreliable.

D.7 Tightness measurement: optimizing over the torus

The bound of Theorem 1 is on the *minimum* sharpness across interpolating minima, so a meaningful tightness comparison requires the empirical flattest-minimum sharpness, not just the sharpness at an arbitrary interpolating point. Under our experimental setup ($N \geq 2 \dim \mathcal{V}_{r,d}$, generic Gaussian inputs), the fiber $\mu_{r,d}^{-1}(\mu^*)$ is a single torus orbit under the scaling action $W_i \mapsto \alpha_i W_i$ with $\prod_i \alpha_i^{r^{h-i}} = 1$ (up to discrete permutation symmetries that do not affect sharpness). Finding the flattest minimum reduces to optimizing sharpness over this $(h-1)$-dimensional torus.

The procedure is as follows. Starting from the random target θ^*, we numerically minimize $\lambda_{\max}\big(\nabla^2 L(\theta(\alpha))\big)$ over the constrained orbit, using projected gradient descent on $(\alpha_1, \ldots, \alpha_{h-1})$ with α_h determined by the scaling constraint. The maximum eigenvalue of the Hessian is computed via Lanczos iteration on Hessian-vector products, evaluated by automatic differentiation on the loss; we use 20 iterations of Lanczos with a tolerance of 10^{-8}, which is more than sufficient for the dominant eigenvalue to converge given the well-separated leading spectrum of the polynomial-network Hessian at interpolating minima.

The minimizer $\theta^*_{\mathrm{flattest}}$ in the orbit is taken as the empirical estimate of the flattest minimum, and its sharpness $\lambda_{\max}(\nabla^2 L(\theta^*_{\mathrm{flattest}}))$ is the empirical quantity compared against the bound. The tightness ratio

$$\rho_{\mathrm{tight}}(r, h, d) := \frac{\lambda_{\max}(\nabla^2 L(\theta^*_{\mathrm{flattest}}))}{S(r, h, d, X, Y)}$$

is reported across the architecture sweep, where S is the bound from Theorem 1. A ratio of 1 means the bound is tight; a ratio bounded above as (r, h) vary means the depth-degree scaling is captured

exactly; a depth-dependent ratio would indicate a multiplicative tightness gap that grows with the architecture.

In the regime of moderate input dimension $d_0 \geq 4$ and depth $h \geq 3$ at activation degree $r \geq 3$, the empirical ratio ρ_{tight} stabilizes near a small constant independent of depth and degree, indicating that the bound captures the correct order of magnitude of the flattest-minimum sharpness. The full tightness scatter is reported in Figure 1 of the main text. The tightness gap closes as the depth-degree product grows, consistent with the r^{2h} scaling becoming the dominant contribution to the sharpness in the deep limit.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We took significant effort to formulate our claims reflecting our contribution and scope as clear as possible in the abstract and introduction

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss limitations of our approach in Sections 2, 3, and 4. Assumptions are clearly stated in Section 1 and 2.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: We always provide the full set of assumptions and complete proofs for all theoretical results. All proofs are provided in the appendices and referenced accordingly in the main paper.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: We detail all information needed to reproduce our experimental results in Appendix D.

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: All the details needed to reproduce our experiment are given in Appendix D.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We detail all relevant details (splits, hyperparameters, type of optimizer, etc.) to understand the results in Appendix D.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: For the illustrative experiment we present, all quantities have been computed analytically, meaning the results we have obtained are exact. Where an optimizer is used, we always use full batches and we document the seeds used to generate the data.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: he experiments is illustrative and small scale, with all plots generated on a standard desktop PC in about an hour.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We read the Code of Ethics and our research conforms with it in every respect.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: We are not aware of any societal impacts our work might have.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: We are not releasing any data or models, thus there is no risk for misuse.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: We do not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: We are not releasing any new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: We neither performed crowdsourcing experiments nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: We did not perform a study with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: No LLMs were used in the core method development of this research.

    Guidelines:

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