LINEARIZED ATTENTION CANNOT ENTER THE KERNEL REGIME AT ANY PRACTICAL WIDTH

Anonymous authors
Paper under double-blind review

ABSTRACT

Influence functions are now widely used to audit transformer training data, relying on the unverified assumption that attention reaches the neural tangent kernel (NTK) regime. Softmax attention admits no exact NTK, so this question is studied in this paper through linearized attention, the standard tractable proxy. This proxy is shown not to enter the kernel regime at any practical width. An exact correspondence is established between parameter-free linearized attention and a data-dependent Gram-induced kernel whose condition number is the cube of the input Gram's. With $\kappa_d(\mathbf{G})$ the effective condition number of the input Gram matrix, a width $m = \Omega(\kappa_d(\mathbf{G})^6 \log n)$ suffices for convergence, and a matching necessity bound makes the exclusion two-sided. For natural image datasets, with $\kappa_d(\mathbf{G}) \approx 9 \times 10^8$ (MNIST) and 4×10^7 (CIFAR-10), even the tight $\kappa_d(\mathbf{G})^3$ requirement exceeds 10^{22}, more than ten orders of magnitude beyond the $\sim 10^{12}$ parameters of the largest known architectures. *Influence malleability* is introduced to measure the consequence. Linearized attention exhibits 3–9$\times$ higher malleability than ReLU networks under adversarial data perturbation, with the gap set by dataset condition number and task. For trainable query, key, and value projections, an exact NTK block decomposition yields a precondition floor $\kappa^+(\mathbf{K}_\infty^{\mathrm{QKV}}) \geq \kappa_d(\mathbf{G})$. Non-convergence then persists at exponent two for every dataset studied. The cubic mechanism itself is erased under standard independent initialization and survives only under an explicit, checkpoint-measurable condition. This precondition is measured above the non-convergence threshold across more than twenty pretrained language, vision, and audio models, the setting where influence methods are deployed.

1 INTRODUCTION

The learning dynamics of attention remain largely uncharacterized, with analyses to date focused on architectural properties at initialization or on final performance, leaving uncharacterized how the flexibility that distinguishes attention emerges over training. The Neural Tangent Kernel (NTK) framework (Jacot et al., 2018; Lee et al., 2019) makes such dynamics tractable by showing that sufficiently wide networks train in a *lazy* regime with an approximately constant kernel, yet attention mechanisms have remained largely outside this framework.

This connection carries direct practical consequences. Modern influence methods (TRAK (Park et al., 2023), DataInf (Kwon et al., 2024), EK-FAC (Grosse et al., 2023)) share a foundational assumption: the model operates near the kernel regime, so the first-order NTK approximation holds throughout training. When applied to transformers, this assumption is taken for granted, and practitioners use influence scores to audit training data, detect poisoning, and select representative examples (Koh & Liang, 2017; Zhang & Zhang, 2022). The central question is therefore *when, if ever, does linearized attention enter the kernel regime, and what follows for model accountability if it does, or if it structurally cannot?* If convergence holds, influence methods on attention inherit their guarantees. If it fails at every practical width, those scores are ungrounded and exploitable.

Two-layer ReLU networks show the expected decrease in NTK distance $\|f_m - f_{\mathrm{NTK}}\|$ with width, while attention's NTK distance stays flat or grows (non-monotonic on MNIST, increasing on CIFAR-10), the signature of the *feature learning* regime (Chizat et al., 2019) (Figure 1).

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

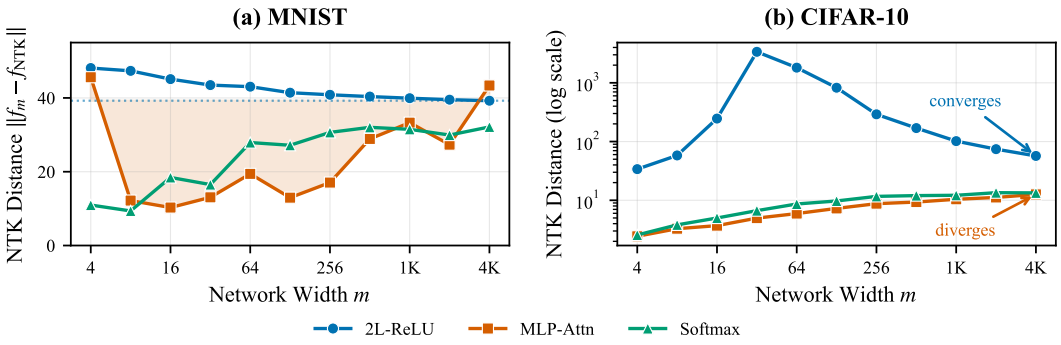


Figure 1: NTK distance $\|f_m - f_{\text{NTK}}\|$ across widths, with f_m the finite-width trained model and f_{NTK} the infinite-width NTK predictor. **2L-ReLU** (blue) converges toward the kernel regime, its distance decreasing as $m \to \infty$. **MLP-Attn** (orange), which uses *linearized* attention ($f^{\text{att}}(\mathbf{X}) = \mathbf{X}\mathbf{X}^T\mathbf{X}$), stays in the feature-learning regime, its distance non-monotonic on MNIST and rising on CIFAR-10, as Theorem 4.4 predicts (regimes in Remark 4.1). **Softmax** (green), parameter-free softmax attention at the deployed input scaling, is non-convergent on both datasets as well, with its distance measured against a static arc-cosine reference on the fixed softmax-transformed features, on the same footing as the 2L-ReLU and MLP-Attn curves (Appendix B.14). Each curve compares one model to its own infinite-width kernel.

To quantify the consequence, *influence malleability* is introduced, the capacity to alter reliance on training examples as their quality changes, and attention is found to exhibit 3–9× higher malleability than ReLU networks, a measurable signature of its sensitivity to training data.

1.1 Contributions

How far linearized attention sits from the kernel regime is shown to be governed exactly by the input conditioning $\kappa_d(\mathbf{G})$ (Figure 1). Theorem 4.2 gives the exact Gram-induced kernel, and Theorem 4.4 shows the attention map cubes the effective condition number, so that certifying the kernel regime requires a width $m = \Omega(\kappa_d(\mathbf{G})^6 \log n)$, infeasible for natural images where $\kappa_d(\mathbf{G}) \approx 9 \times 10^8$ (MNIST) and 4×10^7 (CIFAR-10) push even the tight $\kappa_d(\mathbf{G})^3$ requirement past 10^{22}, with a matching necessity bound (Theorem 4.5) making the exclusion two-sided and resolving the convergence case deferred by Tensor Programs (Yang, 2020) (Remark 7.4) for non-backpropagation-like architectures. *Influence malleability* is introduced as a quantifiable signature of this non-convergence. Linearized attention shows 3–9× higher influence flip rates than ReLU networks (6–9× multiclass, attenuated at low $\kappa_d(\mathbf{G})$), and the same data-dependent kernel reduces approximation error under spectral alignment (Proposition D.15) while creating adversarial vulnerability, a tradeoff with a common origin. For genuinely trainable query, key, and value projections, an exact NTK block decomposition (Theorem C.8) yields a precondition floor $\kappa^+(\mathbf{K}_\infty^{\text{QKV}}) \geq \kappa_d(\mathbf{G})$ (Theorem 4.9) and a phase transition between an aligned regime where the cube survives and independent initialization where it is erased (Theorem C.9). Non-convergence persists at exponent two, $H^\star = \Theta(\kappa_d(\mathbf{G})^2/\varepsilon^2)$ (Theorem 4.10), with the cube transferring only under a checkpoint-measurable alignment condition (Proposition 4.11). Finally, the conditioning diagnostic places every one of more than twenty pretrained language, vision, and audio models in the non-convergence regime ($\kappa_d(\mathbf{G}) > 10^2$, including Mamba, RWKV, DeepSeek-V3, and Kimi-K2, Appendix B.13), so the precondition holds where influence methods are deployed. Sections 5–6 supply the empirical, cross-model, and deployed-architecture evidence.

2 Related Work

The neural tangent kernel (Jacot et al., 2018; Lee et al., 2019) establishes that infinitely wide networks trained by gradient descent behave as kernel methods, with extensions to generalization (Arora et al., 2019), depth (Du et al., 2019), and arbitrary architectures (Yang, 2020). Chizat et al. (2019) separate this *lazy* regime from the *feature-learning* regime, and its known limits are architectural, with Wenger et al. (2023) tying the failure to networks far wider than deep and Nichani et al. (2023) to conditions under which networks leave it. The non-convergence identified here is instead *data-driven*,

indexed by the input conditioning $\kappa_d(\mathbf{G})$ in place of the width-to-depth ratio. For influence, Zhang & Zhang (2022) show NTK-based computation is more stable than inverse-Hessian methods and Loo et al. (2022) show the NTK evolves under training. On the architecture side, Ahn et al. (2024) show linearized attention reproduces Transformer optimization dynamics, Choromanski et al. (2021) give efficient Performer implementations, and Hron et al. (2020) treat multi-head attention as a Gaussian process as heads $\to \infty$, none of which identifies the cubic conditioning established here (Theorem 4.2). Yang (2020) (Remark 7.4) defers NTK convergence for non-backpropagation-like architectures, which this work resolves. From the softmax side, Sakai et al. (2025) show a single softmax attention layer has a non-Gaussian infinite-width limit, and Bae et al. (2022) show influence estimates are unreliable in deep nonlinear networks, consistent with Proposition B.1.

3 METHODOLOGY

Problem setup. Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ denote the training inputs (n examples of dimension d) with unit-norm rows and Gram matrix $\mathbf{G} = \mathbf{X}\mathbf{X}^\top$, and let m be the width of the MLP. The MLP-Attn predictor (Eq. 3) has empirical finite-width NTK $\mathbf{K}_m$ and infinite-width limit $\mathbf{K}_\infty$. Influence scores are computed from the regularized kernel system of Definition 3.3, following the NTK-based influence formulation of Zhang & Zhang (2022). The central question is quantitative, namely whether $\mathbf{K}_m$ approaches $\mathbf{K}_\infty$ closely enough for influence estimation, $\|\mathbf{K}_m - \mathbf{K}_\infty\| \lesssim \lambda$, at any width m attainable on current hardware. Throughout, the architecture is said to *enter the kernel regime* exactly when this criterion is met. Theorem 4.4 and Theorem 4.5 certify, from the sufficiency and necessity sides respectively, that this criterion cannot be met at any practical width for natural image data.

Influence via kernel ridge regression. The regularized kernel system above is the closed-form solution of a kernel ridge regression (KRR) problem. For kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$, ridge $\lambda > 0$, and labels $\mathbf{y} \in \mathbb{R}^n$, the *influence vector* $\boldsymbol{\mathcal{I}} := (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$ collects the fitted dual coefficients, and the test prediction is $f(\mathbf{x}_{\text{test}}) = \mathbf{k}(\mathbf{x}_{\text{test}})^\top \boldsymbol{\mathcal{I}}$ with $\mathbf{k}(\mathbf{x}_{\text{test}}) = [K(\mathbf{x}_{\text{test}}, \mathbf{x}_1), \ldots, K(\mathbf{x}_{\text{test}}, \mathbf{x}_n)]^\top$. This linearizes the influence-function framework of Koh & Liang (2017) around the NTK, follows the formulation validated by Zhang & Zhang (2022), and parallels TRAK (Park et al., 2023). Entry i of $\boldsymbol{\mathcal{I}}$ governs how strongly training point $\mathbf{x}_i$ pulls the prediction toward y_i, and leave-one-out influence (Definition 3.3) follows from a Sherman–Morrison update to $(\mathbf{K} + \lambda\mathbf{I})^{-1}$ without retraining. Here $\mathbf{K}$ is the empirical finite-width NTK $\mathbf{K}_m$, so $\boldsymbol{\mathcal{I}}$ is well-defined at every practical width whether or not $\mathbf{K}_m$ converges to $\mathbf{K}_\infty$. The ridge λ both stabilizes the solution and controls how strongly kernel or training-data perturbations propagate into $\boldsymbol{\mathcal{I}}$ (Theorem C.23). Appendix D develops this into a bias–variance–malleability analysis of $\boldsymbol{\mathcal{I}}$, whose generalization and bias are governed by the kernel eigenspectrum and its alignment with the target (Bordelon et al., 2020; Canatar et al., 2021).

3.1 LINEARIZED ATTENTION ARCHITECTURE

Standard scaled dot-product attention (Vaswani et al., 2017) maps the input $\mathbf{X} \in \mathbb{R}^{n\times d}$ to a query matrix $\mathbf{Q} = \mathbf{X}\mathbf{W}_Q$, a key matrix $\mathbf{K} = \mathbf{X}\mathbf{W}_K$, and a value matrix $\mathbf{V} = \mathbf{X}\mathbf{W}_V$ via learned projections $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$, and computes

$$\text{Attention}(\mathbf{X}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}. \tag{1}$$

Two simplifications isolate a parameter-free object that admits an exact kernel characterization and is the object of study throughout this paper: identity query, key, and value (QKV) projections ($\mathbf{W}_Q = \mathbf{W}_K = \mathbf{W}_V = \mathbf{I}$, so $\mathbf{Q} = \mathbf{K} = \mathbf{V} = \mathbf{X}$), and the first-order Taylor approximation of the softmax nonlinearity applied to the score matrix $\mathbf{A} = \mathbf{Q}\mathbf{K}^T/\sqrt{d}$, $\exp(A_{ij}) \approx 1 + A_{ij}$ (Appendix C). The term *linearized attention* refers to this identity-projection, first-order proxy throughout. Here "linearized" denotes only the Taylor relationship to softmax just described, not a trainable efficient-attention architecture that shares the name.

Definition 3.1 (Linearized Attention)**.** For input matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, the linearized attention mechanism is the identity-projection, first-order special case of scaled dot-product attention:

$$f^{\text{att}}(\mathbf{X}) = \mathbf{X}\mathbf{X}^T\mathbf{X} \tag{2}$$

Remark 3.2 (Scope of the Identity-Projection Reduction). The identity-projection map is the exactly solvable case of a larger family of trainable query, key, and value projections (Section 4.1) in which

the conditioning reaches the cube $\kappa_d(\mathbf{G})^3$, corresponding to aligned orthogonal initialization of that family. Results proved for it do not extend to the general member by default. Under either initialization scheme, the limiting kernel obeys the same precondition floor, $\kappa^+(\mathbf{K}) \geq \kappa_d(\mathbf{G})$ (Theorem 4.9). The cube is attained only under aligned orthogonal initialization, and standard independent initialization erases it exactly, replacing $\mathbf{G}^3$ by a scalar multiple of $\mathbf{G}$ (Theorem C.9).

This operation implements kernel-weighted feature aggregation, with $\mathbf{X} \in \mathbb{R}^{n \times d}$ the *entire* training set, so the transformation is computed once over all n examples, making the architecture transductive. In practice the rows of $f^{\text{att}}(\mathbf{X})$ are ℓ_2-normalized before the MLP, for comparable input magnitudes across architectures. The complete architecture adds a two-layer MLP,

$$f_{\text{MLP-Attn}}(\mathbf{X}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(f^{\text{att}}(\mathbf{X})\, \mathbf{w}_r) \tag{3}$$

where σ is ReLU, $\mathbf{w}_r \sim \mathcal{N}(0, \kappa^2 \mathbf{I})$ with $\kappa > 0$ the initialization scale, and the second-layer signs $a_r \in \{-1, +1\}$ are fixed. Only the first-layer weights $\boldsymbol{\theta} \coloneqq (\mathbf{w}_1, \ldots, \mathbf{w}_m)$ are trained, minimizing a loss regularized by the shrinkage penalty $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$ around the initialization $\boldsymbol{\theta}_0$ (Appendix B), the standard lazy-training regularizer that keeps the trained network near its NTK linearization and aligns the influence computation with the NTK. This is the standard two-layer NTK simplification, isolating the first-layer dynamics that define the kernel. A trainable a_r adds a second-layer term to the NTK but does not change the cubic conditioning of the attention features established in Theorem 4.4.

3.2 Influence Malleability

The data-dependent kernel structure of linearized attention (Section 3) suggests that attention should exhibit heightened *sensitivity* to training data. To quantify this, *influence malleability* is introduced, capturing how readily a model's reliance on specific training examples changes under perturbation. High malleability indicates high sensitivity, and low malleability indicates rigid, fixed data dependencies.

Definition 3.3 (Influence Function)**.** Following Koh & Liang (2017) and the NTK-based formulation of Zhang & Zhang (2022), the influence of removing training example $(\mathbf{x}_i, y_i)$ on a test prediction at $\mathbf{x}_{\text{test}}$ with true label y_{test} is:

$$I_i \coloneqq I(\mathbf{x}_i, \mathbf{x}_{\text{test}}) = \frac{1}{2}\left[(f^{\setminus i}(\mathbf{x}_{\text{test}}) - y_{\text{test}})^2 - (f(\mathbf{x}_{\text{test}}) - y_{\text{test}})^2\right] \tag{4}$$

where $f^{\setminus i}$ denotes the model trained without point i. Write I_i^{orig} for I_i under the original training set and I_i^{adv} after perturbation. Positive influence indicates removing the point increases loss (a helpful example), and negative influence indicates removing it decreases loss (a harmful example). Using the KRR view above, this is computed in closed form via the Sherman–Morrison update to $(\mathbf{K}_m + \lambda \mathbf{I})^{-1}$ without retraining, with $\mathbf{K}_m$ the empirical finite-width NTK (Park et al., 2023).

Definition 3.4 (Influence Malleability)**.** Given training set $\mathcal{S}$, perturbation budget $\varepsilon > 0$, and top-τ high-influence set $\mathcal{H}_\tau = \{i : I_i^{\text{orig}} > \text{quantile}_{1-\tau}\}$ (with $\tau = 0.1$, i.e., the top 10% by influence score), generate adversarial examples $\mathbf{x}_i^{\text{adv}}$ via Projected Gradient Descent (PGD, $\|\mathbf{x}_i^{\text{adv}} - \mathbf{x}_i\|_\infty \leq \varepsilon$) and compute perturbed scores I_i^{adv}. The **Influence Flip Rate** is:

$$\text{FlipRate} = \frac{|\{i \in \mathcal{H}_\tau : \text{sign}(I_i^{\text{orig}}) \neq \text{sign}(I_i^{\text{adv}})\}|}{|\mathcal{H}_\tau|} \tag{5}$$

The flip rate captures the capacity to *re-evaluate* which examples help versus hurt prediction, which is the signature of sensitivity to training data. PGD with $\varepsilon = 0.03$ (L_∞) is used throughout. As a complementary measure, **influence stability** is assessed via Spearman's rank correlation ρ between original and perturbed influence rankings, where lower ρ indicates greater malleability. Three data intervention strategies are evaluated: *Curated* (removing the top-τ influential examples from training), *Transformed* (replacing them with adversarial versions), and *Adversarial* (applying PGD perturbations to all training data).

Practical Implications for Influence Methods. The non-convergence result carries direct practical consequences for deploying TRAK (Park et al., 2023), DataInf (Kwon et al., 2024), and EK-FAC (Grosse et al., 2023) on attention architectures. Appendix B.12 gives a step-by-step diagnostic ($\kappa_d(\mathbf{G})$ via randomized SVD), the resulting data-auditing and active-learning adjustments, and the conditions under which the effect is attenuated.

4 THEORETICAL ANALYSIS

The mechanistic chain connecting linearized attention to influence malleability is:

lin. attn $\xrightarrow{\text{Thm 4.2}} \tilde{\mathbf{G}} = \mathbf{G}^3 \xrightarrow{\text{Thm 4.4}} \kappa_d(\tilde{\mathbf{G}}) = \kappa_d(\mathbf{G})^3 \xrightarrow{\text{Thm } C.23 \text{ (contrapositive)}}$ vacuous stability guarantee

For natural image data, cubic conditioning drives $\lambda_1(\mathbf{K}) \sim \lambda_1(\mathbf{G})^3 \gg 1$ against practical regularization $\lambda \sim 10^{-3}$, so the effective condition number $\kappa(\mathbf{K} + \lambda\mathbf{I}) \approx \lambda_1(\mathbf{K})/\lambda$ is enormous, and the stability scale $\epsilon/[\lambda(\lambda - \epsilon)]$ of Theorem C.23 is vacuous, so the influence vector is not certified stable, consistent with the instability measured in Section 5. The $\tilde{\mathbf{G}} = \mathbf{G}^3$ identity is a property of the cube-feature-map family (Lemma C.26). Attention's NTK non-convergence follows as a corollary under the linearization, with distinct fixed QKV projections addressed in Proposition C.4.

Remark 4.1 (Regimes of NTK convergence). Write $d_m := \|f_m - f_{\text{NTK}}\|_2$ for the NTK distance at width m. Three behaviors are distinguished. (i) *Convergence* ($d_m \to 0$): 2L-ReLU, at sufficient width $m_0 = O(n \log n)$. (ii) *Non-convergence* (d_m stays flat or grows, Figure 1): the regime into which Theorem 4.4 places MLP-Attn through a sufficient width beyond feasibility, with a matching lower bound on the certified width from Theorem 4.5, so no practical width certifies the kernel-regime criterion. (iii) *Divergence* ($d_m \to \infty$): not claimed, as MLP-Attn stays structured and bounded, far from its limit.

Theorem 4.2 (Data-Dependent Gram-Induced Kernel). *Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $\|\mathbf{x}_i\|_2 = 1$ for all i. The kernel induced by linearized attention $f^{att}(\mathbf{X}) = \mathbf{X}\mathbf{X}^T\mathbf{X}$ is:*

$$K_{\textit{LinAttn}}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k,\ell=1}^{n} (\mathbf{x}_i^T \mathbf{x}_k)(\mathbf{x}_k^T \mathbf{x}_\ell)(\mathbf{x}_\ell^T \mathbf{x}_j) \tag{6}$$

where k and ℓ are summation indices over all n training samples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathbf{X}$. This is a data-dependent kernel induced by the Gram matrix $\mathbf{X}\mathbf{X}^T$, exhibiting transitive similarity structure, where influence flows from $\mathbf{x}_i$ through intermediate points $\mathbf{x}_k, \mathbf{x}_\ell$ to $\mathbf{x}_j$.

Full proof in Appendix C.

For the MLP-Attn architecture $g(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(\mathbf{w}_r^T f^{\text{att}}(\mathbf{x}))$ with $\mathbf{w}_r \sim \mathcal{N}(0, \kappa^2 \mathbf{I})$ and fixed $a_r \in \{-1, +1\}$, the empirical NTK $\mathbf{K}_{m,\text{seq}}$ converges as $m \to \infty$ to $\mathbf{K}_{\infty,\text{seq}} = K_{\text{ReLU}}(f^{\text{att}}(\mathbf{X}), f^{\text{att}}(\mathbf{X}'))$, the standard ReLU NTK applied to the attention-transformed features (Appendix C).

The influence vector inverts the regularized kernel, so its stability is governed by the conditioning of $\mathbf{K} + \lambda\mathbf{I}$. A standard resolvent bound (Theorem C.23, Appendix C) gives $\kappa(\mathbf{K} + \lambda\mathbf{I}) \leq \lambda_1/\lambda + 1$ and, for a kernel perturbation $\|\Delta\mathbf{K}\|_2 \leq \epsilon < \lambda$, a propagated change $\|(\mathbf{K} + \lambda\mathbf{I})^{-1} - (\mathbf{K} + \Delta\mathbf{K} + \lambda\mathbf{I})^{-1}\|_2 \leq \epsilon/[\lambda(\lambda - \epsilon)]$. The influence vector is stable precisely when the kernel is well-conditioned.

Assumption 4.3 (μ-Incoherence). $\|\mathbf{x}_i\|_2 = 1$ for all i, and $\max_{i \neq j} |\mathbf{x}_i^T \mathbf{x}_j| \leq \mu/\sqrt{d}$ with $\mu = \Theta(\sqrt{\log n})$. For isotropic sub-Gaussian inputs a union bound over the $\binom{n}{2}$ pairs then holds with probability $\geq 1 - n^2 e^{-c\mu^2} \to 1$, while the incoherence level $\mu/\sqrt{d} = O(\sqrt{\log n/d})$ vanishes for $d = \Omega(\log n)$ (Tropp, 2012).

The linearized attention kernel exhibits data-dependent sensitivity to input perturbations that grows with $\lambda_1(\mathbf{G})$, unlike data-independent polynomial kernels, and its globally-coupled feature map causes the empirical NTK to evolve with the full correlation structure during training, the hallmark of feature learning (Proposition C.24, Remark C.25; Appendix C).

The following theorem formalizes why linearized attention cannot enter the kernel regime at any practical width.

Theorem 4.4 (Spectral Amplification and NTK Non-Convergence). *Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ with $\|\mathbf{x}_i\|_2 = 1$ and $r = \text{rank}(\mathbf{X}) \leq \min(n, d)$. Let $\mathbf{G} = \mathbf{X}\mathbf{X}^T$ have positive eigenvalues $\lambda_1 \geq \cdots \geq \lambda_r > 0$ and define the effective condition number $\kappa_d(\mathbf{G}) = \lambda_1/\lambda_r$. Assume the μ-incoherence condition (Assumption 4.3). Then the attention-transformed Gram matrix satisfies $\tilde{\mathbf{G}} = \mathbf{G}^3$, amplifying the effective condition number:*

$$\kappa_d(\tilde{\mathbf{G}}) = \kappa_d(\mathbf{G})^3 \tag{7}$$

Consequently, for the finite-width NTK of the MLP-Attn architecture to certify the relative approximation error $\|K_{m,seq} - K_{\infty,seq}\|_2/\lambda^+_{\min}(K_{\infty,seq}) \le \varepsilon$*, it suffices that the width satisfy* $m = \Omega(\kappa_d(\mathbf{G})^6 \log n/\varepsilon^2)$ *when* $n \le d$*. In the rank-deficient regime* $n > d$ *of the datasets studied here, the floor on* $\lambda^+_{\min}(\mathbf{K}_{\infty,seq})$ *requires an incoherence assumption* $\tilde{\mu} = \max_{i\neq j} |\hat{\mathbf{G}}_{ij}| < 1$ *on the attention-transformed features (with* $\hat{\mathbf{G}}$ *the row-normalized* $\mathbf{G}^3$*), under which* $\mathbf{K}_{\infty,seq}$ *is full rank and the honest sufficient width is* $m = \Omega\big(\kappa_d(\mathbf{G})^6 \lambda_r^6 \frac{\log n}{\log(1/\tilde{\mu})} \log n/\varepsilon^2\big)$ *(Appendix C). Either way the cubic conditioning renders the sufficient width physically infeasible for natural images, and the empirical non-convergence is reported in Figure 1. The analogous sufficient width for 2L-ReLU is* $m = \Omega(n \log n/\varepsilon^2)$*, without spectral amplification.*

Full proof in Appendix C. The analyzed object is transductive attention over the training set, so $\kappa_d(\mathbf{G})$ conditions the training-set Gram that governs influence-function reliability, while within-sequence attention obeys the same identity with a sequence Gram $\mathbf{G}_{\text{seq}}$ whose $\kappa_d(\mathbf{G}_{\text{seq}})$ is left open (Appendix A).

Theorem 4.5 (Necessity of Width for the Row-Normalized Architecture)**.** *Let* $\hat{\mathbf{G}} = \mathbf{D}^{-1}\mathbf{G}^3\mathbf{D}^{-1}$ *denote the effective Gram matrix of the row-normalized attention features used throughout the experiments (Appendix B), and let* $\mathbf{K}_{\infty,seq} = \mathbf{A} \circ \hat{\mathbf{G}}$ *be the corresponding infinite-width kernel, with* $\mathbf{A}$ *the arc-cosine kernel of the proof of Theorem 4.4. At every width* $m \ge 1$*, the relative spectral criterion of Theorem 4.4 can hold with probability exceeding* $15/16$ *only if*

$$m \;\ge\; \frac{1}{8\varepsilon^2 \lambda_{\min}(\mathbf{K}_{\infty,seq})^2}\,. \tag{8}$$

This anti-concentration lower bound complements the sufficiency direction of Theorem 4.4. Stated through the dataset's own $\lambda_{\min}(\mathbf{K}_{\infty,seq})$*, it applies to the real datasets directly, and its exponent reaches the worst case* $\kappa_d(\mathbf{G})^3$ *on a two-point coalescing family, matched there by an* $O(\kappa_d(\mathbf{G})^3/\varepsilon^2)$ *sufficiency. Row normalization removes the* $\kappa_d(\mathbf{G})^3$ *diagonal factor that inflates the unnormalized certificate of Theorem 4.4. Full proof in Appendix C.14.*

Remark 4.6 (Practical Width). Throughout, *practical width* denotes widths achievable on current hardware ($m \lesssim 10^5$; e.g., GPT-3: $m = 12{,}288$, ViT-Huge: $m = 1{,}280$). The tight required widths $m \gg 10^{26}$ (MNIST) and $m \gg 10^{22}$ (CIFAR-10) lie 21 and 17 orders of magnitude beyond this ceiling.

Remark 4.7 (Scope and Generalizations). Under standard layer-wise self-composition, k stacked linearized attention layers give $\tilde{\mathbf{G}} = \mathbf{G}^{3^k}$, doubly exponential in depth. Further scope (the milder frozen-reference-map recursion $\mathbf{G}^{2k+1}$, truncated attention, construction-specificity of the cubic identity, and the at-initialization restriction) is given in Appendix B.17.2.

The softmax operator does not admit an exact NTK, so a matrix-exponential surrogate is analyzed in its place (Proposition B.1, Appendix B.14). Under the normalization used here the faithful softmax feature Gram collapses to rank one, a failure mode distinct from the non-convergence proved for the linear proxy, while the exponential (FAVOR+) surrogate inherits the floor $\kappa_d(\mathbf{G})$ and regenerates the cube through its value pathway (Appendix B.16). Whether softmax at deployed scale, whose logits avoid this collapse, carries the mechanism is left open.

Distinct fixed query, key, and value projections preserve the cubic conditioning exactly in the orthogonal case, subject to a checkable condition that can fail under trained projections (Proposition C.4, Appendix C).

4.1 Trainable Query, Key, and Value Projections in the NTK

Every result so far treats the projections $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ as fixed, so attention supplies a static feature map and only the weights downstream of it enter the NTK. What remains is the setting in which attention itself learns, so that the query, key, and value parameters participate in the kernel along with everything else. The smallest model in which this question admits an exact answer is single-layer, multi-head linear attention with a linear output layer.

Definition 4.8 (Trainable Linear Attention)**.** Multi-head linear attention with per-head trainable projections $\mathbf{W}_Q^h, \mathbf{W}_K^h \in \mathbb{R}^{d\times p}$ and $\mathbf{w}_V^h \in \mathbb{R}^d$ (H heads), whose empirical NTK $\mathbf{K}_H(\theta)$ is taken over all trainable parameters; $\mathbf{K}_\infty^{\text{QKV}} := \lim_{H\to\infty} \mathbf{K}_H$ exists almost surely, with H playing the role

of the width m. Under *aligned orthogonal* initialization the model coincides at initialization with the parameter-free proxy $f^{\text{att}}(\mathbf{X}) = \mathbf{X}\mathbf{X}^\top\mathbf{X}$; under *independent Gaussian* initialization $\mathbf{W}_Q^h, \mathbf{W}_K^h$ are drawn independently, matching standard transformer initialization. Full construction in Appendix C.6.

The forward map is trilinear in $\mathbf{W}_Q^h, \mathbf{W}_K^h, \mathbf{w}_V^h$, so the Jacobian Gram admits an exact three-term block decomposition $\mathbf{K}_H = \mathbf{K}_V + \mathbf{K}_Q + \mathbf{K}_K$ into value, query, and key contributions (Theorem C.8, Appendix C.6). The cube can enter only through the value block, and only when the score matrix aligns with $\mathbf{G}$. The three blocks all factor through the same $\mathbf{X}$. At both analyzed initializations the limiting kernel's conditioning is then bounded below by that of $\mathbf{G}$ (Theorem 4.9). Write $\kappa^+(\mathbf{K}) \coloneqq \lambda_{\max}(\mathbf{K})/\lambda^+_{\min}(\mathbf{K})$ for the ratio of the largest to the smallest positive eigenvalue of a PSD matrix $\mathbf{K}$, so that $\kappa^+(\mathbf{G}) = \kappa_d(\mathbf{G})$.

Theorem 4.9 (Precondition Floor for Trainable Attention)**.** *Under either initialization scheme of Definition 4.8, the restriction of $\mathbf{K}_\infty^{\text{QKV}}$ to its range satisfies*

$$\kappa^+(\mathbf{K}_\infty^{\text{QKV}}) \;\geq\; \kappa_d(\mathbf{G}), \tag{9}$$

with equality under independent initialization and strict inequality, up to $\kappa_d(\mathbf{G})^3$, under aligned orthogonal initialization (Theorem C.9). The bound holds at both analyzed initializations for every initialization scale, the exponent on $\kappa_d(\mathbf{G})$ ranging from one (independent) to the cube (aligned orthogonal) but never below one. Which exponent a trained *checkpoint realizes is governed by the alignment diagnostic of Proposition 4.11. Proof in Appendix C.6.*

Where the kernel sits inside $[\kappa_d(\mathbf{G}), \kappa_d(\mathbf{G})^3]$ is set by a competition between the value block's cubic term and the query/key blocks' linear term. Under aligned orthogonal initialization this resolves exactly: $\mathbf{K}_\infty^{\text{QKV}} = \mathbf{G}^3 + \tau_0\mathbf{G}$ with $\tau_0 = 4\|\mathbf{G}\|_F^2/d^2$, sharing the eigenbasis of $\mathbf{G}$, so that

$$\kappa^+(\mathbf{K}_\infty^{\text{QKV}}) = \kappa_d(\mathbf{G}) \cdot \frac{\lambda_1^2 + \tau_0}{\lambda_r^2 + \tau_0} \;\in\; [\kappa_d(\mathbf{G}),\, \kappa_d(\mathbf{G})^3], \tag{10}$$

the cube ($\kappa^+ \geq \frac{1}{2}\kappa_d(\mathbf{G})^3$) reached only when $\kappa_d(\mathbf{G}) \leq d/2$. Under standard independent Gaussian initialization the limiting NTK is instead exactly $c_{\text{I}}\mathbf{G}$ for a positive constant c_{I}, so $\kappa^+ = \kappa_d(\mathbf{G})$ and the cube is erased at every order. Both forms are established exactly in Theorem C.9 (Appendix C.6) and collected in Table 8.

The threshold $\kappa_d(\mathbf{G}) \leq d/2$ separates a cubic regime from a lifted one, and every dataset studied (MNIST, CIFAR-10, DistilBERT; $\kappa_d(\mathbf{G}) \in [10^7, 10^9]$ against $d \in \{784, 3072, 768\}$) lies on the lifted side. The lift changes only the exponent governing how many heads certify the kernel regime, pinned down exactly by matching necessity and sufficiency arguments, without restoring that regime.

Theorem 4.10 (Non-Convergence at Exponent Two)**.** *In the lifted regime ($\tau_0 \geq \lambda_r^2$, in particular whenever $\kappa_d(\mathbf{G}) > d/2$), matching necessity and sufficiency give the certification threshold*

$$H^\star = \Theta\big(\kappa_d(\mathbf{G})^2/\varepsilon^2\big) \tag{11}$$

for the relative spectral criterion $\|\mathbf{K}_H - \mathbf{K}_\infty^{\text{QKV}}\|_2/\lambda^+_{\min}(\mathbf{K}_\infty^{\text{QKV}}) \leq \varepsilon$ under aligned orthogonal initialization, with explicit constants. For the conditioning values measured in this paper, $H^\star$ reaches 10^{15}–10^{18} for the image datasets and 10^{14} for the contextual embedding, against practical head counts of order 10^2 (a few hundred at most in the largest deployed transformers). Non-convergence therefore persists for trainable attention through exponent two, below the cube's three and the width exponent six. Proof in Appendix C.6.

Theorem 4.10 concerns the kernel at initialization. Training moves the parameters, and whether it moves the checkpoint toward or away from the value-path alignment that restores the cube is answered by a condition readable off any trained checkpoint.

Proposition 4.11 (Conditional Transfer at a Trained Checkpoint)**.** *For an arbitrary parameter point θ, let $\mathbf{A} = \frac{1}{s}\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{X}^\top$ be the attention score matrix, $\mathbf{u}_1, \mathbf{u}_r$ the eigenvectors of $\mathbf{G}$ for its largest and smallest positive eigenvalues, and write $a_1 \coloneqq \mathbf{u}_1^\top\mathbf{A}\mathbf{G}\mathbf{A}^\top\mathbf{u}_1$, $a_r \coloneqq \mathbf{u}_r^\top\mathbf{A}\mathbf{G}\mathbf{A}^\top\mathbf{u}_r$ (single head, sign dropped, summed over the H heads), and $R \coloneqq a_1/a_r$ the alignment ratio. If $\mathbf{A} = c\mathbf{G}$ and $\mathbf{W}_Q\mathbf{W}_Q^\top = \omega\mathbf{I}_d$ (the aligned family), the effective amplification exponent $p(\mathbf{K})$, defined by $\kappa^+(\mathbf{K}) = \kappa_d(\mathbf{G})^{p(\mathbf{K})}$, equals exactly*

$$p(\mathbf{K}) = 1 + \frac{\log(\kappa_d(\mathbf{G})^2 + t) - \log(1+t)}{\log \kappa_d(\mathbf{G})}, \qquad t \coloneqq \frac{\tau}{c^2\lambda_r^2}, \tag{12}$$

where τ is the linear-block coefficient of the per-head kernel $\mathbf{K} = c^2\mathbf{G}^3 + \tau\mathbf{G}$ (Theorem C.9). The exponent is strictly decreasing in t, with $p \geq 2 \iff t \leq \kappa_d(\mathbf{G})$ and $p = 3 \iff t = 0$, so the exact cube is a boundary value unattained at any checkpoint with nonzero value gradient. Within this aligned family, the checkable, checkpoint-measurable condition

$$R \;>\; \kappa_d(\mathbf{G})^2 \tag{13}$$

is necessary, and with a matched bound on the additive term sufficient, for $p(\mathbf{K}) \geq 2$. The whitening configuration $\mathbf{W}_Q\mathbf{W}_K^\top \propto \mathbf{M}^{-1}$ gives $R = \kappa_d(\mathbf{G}) < \kappa_d(\mathbf{G})^2$, so the condition correctly fails there. The ratio R stays directly measurable from $(\mathbf{W}_Q, \mathbf{W}_K, \mathbf{w}_V)$ and $\mathbf{G}$ at any trained checkpoint, giving an optimizer-agnostic diagnostic, while the exact exponent characterization above is specific to the aligned family. Proof in Appendix C.6.

The amplification exponent for each setting is summarized in Table 8 and plotted against measured conditioning in Figure 9 (both Appendix C.6).

Conditioning can also transfer without any training, through a random projection drawn once and never updated, such as He-initialized keys held fixed while the downstream network trains. In the over-complete regime (d_k large relative to $r = \mathrm{rank}(\mathbf{X})$) such a projection transfers the cubic amplification up to a factor $1 + O(\sqrt{r/d_k})$, but in the standard regime $d_k \leq r$ of ordinary multi-head attention no matching lower bound holds, and at $d_k = 1$ the projection is a rigorous whitening counterexample (Proposition C.14, Appendix C.7). Random initialization alone therefore does not certify the cube. Only training decides whether the condition of Proposition 4.11 is met.

From conditioning to influence malleability. The malleability measured in Section 5 follows from the same conditioning. Influence enters the kernel-ridge predictor through $(\mathbf{K} + \lambda\mathbf{I})^{-1}$ (Definition 3.3), whose sensitivity spectrum $\{1/(\lambda_i(\mathbf{K}) + \lambda)\}$ has ridgeless dynamic range $\kappa^+(\mathbf{K})$, so an ill-conditioned kernel concentrates influence on a few eigendirections and leaves it hyper-sensitive along the rest. Because every attention NTK here factors through $\mathbf{G}$, this range is bounded below by the data conditioning, $\kappa^+(\mathbf{K}_\infty^{\mathrm{QKV}}) \geq \kappa_d(\mathbf{G})$ (Theorem 4.9), for trainable attention exactly as for the proxy. Measured malleability therefore reflects the data conditioning.

5 Experiments

5.1 Experimental Setup

Two architectures are compared on MNIST and CIFAR-10 (Fashion-MNIST in the supplement): **2L-ReLU**, a two-layer ReLU network on raw inputs, and **MLP-Attn**, the same MLP preceded by linearized attention ($f^{\mathrm{att}}(\mathbf{X}) = \mathbf{X}\mathbf{X}^T\mathbf{X}$) in the transductive setup of Section 3. NTK distance is measured over widths $m \in \{4, 8, 16, \ldots, 4096\}$ (the ceiling for exact NTK computation), and influence malleability under FGSM, PGD, and MIM attacks at $\varepsilon = 0.03$ (L_∞) on the top-10% influential examples. Full architecture, initialization, dataset, training, and perturbation details are in Appendix B.

5.2 NTK Non-Convergence Across Practical Widths

2L-ReLU exhibits monotonically decreasing NTK distance, consistent with convergence to the kernel regime, while MLP-Attn shows qualitatively different behavior, non-monotonic on MNIST and monotonically increasing on CIFAR-10, consistent with the feature learning regime (Chizat et al., 2019) (Figure 1, with detailed measurements at each width in Table 3, Appendix B.4.3). Since $n > d$ for both datasets (MNIST: $n \approx 60{,}000$, $d = 784$; CIFAR-10: $n \approx 50{,}000$, $d = 3{,}072$), the effective condition number $\kappa_d(\mathbf{G})$ governs the spectral amplification (Remark C.27). Measured values $\kappa_d(\mathbf{G}) \approx 9 \times 10^8$ (MNIST, whose near-constant border pixels give it many near-zero Gram directions) and 4×10^7 (CIFAR-10) yield tight required widths $m \gg 10^{26}$ (MNIST) and $m \gg 10^{22}$ (CIFAR-10), far exceeding $m \leq 4096$ tested, in quantitative agreement with Theorem 4.4.

5.3 Influence Malleability Results

MLP-Attn shows 6–9$\times$ higher flip rates than 2L-ReLU in the 10-class setting, and in binary MNIST the advantage persists (3–4.9$\times$), while in binary CIFAR-10 the ratio narrows to $\approx 1\times$, a confounded comparison examined in Appendices B.2, B (Figure 2).

Table 1: Influence flip rates (%) across perturbation types ($\varepsilon = 0.03$). In the 10-class setting MLP-Attn shows 6–9× higher malleability, while the gap narrows to $\approx 1\times$ in binary CIFAR-10, a confounded setting analyzed in Appendix B.2. Binary setting follows Zhang & Zhang (2022).

Dataset	Setting	Model	FGSM	PGD	MIM
MNIST	10-class	2L-ReLU	4.1	3.3	3.4
		MLP-Attn	34.6	28.9	21.9
	Binary	2L-ReLU	8.4	8.4	8.6
		MLP-Attn	25.9	41.0	40.5
CIFAR-10	10-class	2L-ReLU	3.3	3.1	3.2
		MLP-Attn	26.4	19.1	20.5
	Binary	2L-ReLU	15.2	15.5	15.3
		MLP-Attn	14.3	14.0	14.8

Adversarial training. Under PGD adversarial training (PGD-AT, $\varepsilon = 0.03$), malleability rises sharply for both architectures (2L-ReLU rises from 3.3% to 43.4% on MNIST), yet MLP-Attn reaches comparable malleability under standard training alone, confirming that attention structurally induces the sensitivity that adversarial training forces on ReLU networks, since K_{LinAttn} sensitivity propagates globally through $\mathbf{G} = \mathbf{X}\mathbf{X}^T$ (Proposition C.24) whereas 2L-ReLU sensitivity is dataset-independent (Table 9, Figure 4, Appendices F, B).

6 Discussion

Dual implications of malleability. The data-dependent kernel K_{LinAttn} reduces approximation error under a checkable spectral condition (Proposition D.15), while the same sensitivity creates a structural malleability gap $O(\lambda_1(\mathbf{G})^{5/2}/\sqrt{n})$ (Proposition D.19, Appendix D), which is two-sided on top-aligned labels (Proposition D.5). The effect is bidirectional, since the *Trans* intervention improves accuracy over *Curated* while enabling adversarial manipulation of influence weights (Tables 10, 9). This matches the mean-field feature-learning placement of MLP-plus-attention Transformers by Kim & Suzuki (2024), for which Theorem 4.4 supplies the spectral mechanism.

Evidence beyond toy datasets. The diagnostic and malleability pipeline extend past MNIST and CIFAR-10. Across more than twenty pretrained language, vision, and audio models (Table 4, Appendix B.13), the effective condition number exceeds the non-convergence threshold in every case, reaching $\kappa_d(\mathbf{G}) \approx 1.0 \times 10^7$ for DistilBERT, where malleability on frozen embeddings tracks the condition number as on pixels (AG-News, a 61.3% flip rate versus 0.6% at a whitened $\kappa_d(\mathbf{G}) = 183$, Appendix B.17.3).

7 Limitations

The central results use the fixed parameter-free proxy $f^{\text{att}}(\mathbf{X}) = \mathbf{X}\mathbf{X}^T\mathbf{X}$, though Section 4.1 extends the floor and exponent-two persistence to a trainable-QKV head (Theorems 4.9, 4.10), with a nonlinear MLP showing the same non-convergence empirically (Figure 1, Appendix B.15) and softmax and exponential attention collapsing to rank-1 kernels (Appendix B.14). The width bounds hold at initialization and the $n > d$ floor's incoherence runs to $\tilde{\mu} \approx 0.9999$ here, so the thresholds are conditional while the infeasibility is empirical. The approximation benefit (Proposition D.15) can reverse on an $(n-d)$-dimensional cone of targets, and exact NTK computation restricts experiments to $m \leq 4096$ while the diagnostic extends to any scale (Appendix A).

8 Conclusion

Linearized attention cannot enter the kernel regime at any practical width, since cubic conditioning forces $m = \Omega(\kappa_d(\mathbf{G})^6 \log n)$ for the proxy and trainable QKV projections stay excluded at exponent two above the floor $\kappa^+(\mathbf{K}) \geq \kappa_d(\mathbf{G})$, the cube surviving only under a checkable alignment condition (Theorems 4.4, 4.9, 4.10, Proposition 4.11). The same conditioning drives a 3–9× malleability gap with a computable $O(nd \log n)$ diagnostic at any scale (Appendix B.12).

Ethics Statement

This paper presents theoretical work whose goal is to advance the field of Machine Learning, specifically the reliability of influence-function methods on attention architectures. Its central finding, that influence scores on such models can be highly malleable under small input perturbations, has a dual-use dimension, since the same sensitivity that undermines auditing could inform data-poisoning or influence-manipulation attacks. The intent here is defensive, since the paper provides a computable diagnostic $\kappa_d(\mathbf{G})$ that flags when influence-based auditing is unreliable, so that practitioners do not act on ungrounded scores. All experiments use existing public benchmark datasets, and no human subjects or private data are involved.

Reproducibility Statement

All theoretical claims are stated with explicit assumptions and complete proofs in Appendix C and Appendix D. The experimental setup, datasets, preprocessing, architectures, and hyperparameters are given in Appendix B, and the effective-conditioning and NTK-distance measurements are reproducible from the configuration and scripts provided in the supplementary materials. The conditioning diagnostic $\kappa_d(\mathbf{G})$ is computable in $O(nd \log n)$ from any input matrix, as described in Appendix B.12.

AI Use Statement

Generative AI tools were used to assist with manuscript editing (improving the clarity and consistency of the prose), with supporting code for the experiments and figures, and with checking derivations and surveying related work. The central theoretical results and their proofs, and all reported measurements, were produced and verified by the authors. All AI-assisted text, code, and derivation checks were reviewed by the authors, who take responsibility for the final content of this work, including its claims and artifacts.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Minhak Song, Chulhee Yun, Ali Jadbabaie, and Suvrit Sra. Linear attention is (maybe) all you need (to understand transformer optimization). In *International Conference on Learning Representations*, 2024.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Ruslan Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Juhan Bae, Nathan Ng, Alston Lo, Marzyeh Ghassemi, and Roger Baker Grosse. If influence functions are the answer, then what is the question? In *Advances in Neural Information Processing Systems*, volume 35, pp. 17953–17967, 2022.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pp. 1024–1034. PMLR, 2020.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):2914, 2021.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

Kenneth R. Davidson and Stanislaw J. Szarek. Local operator theory, random matrices and Banach spaces. In *Handbook of the Geometry of Banach Spaces*, volume 1, pp. 317–366. North-Holland, 2001.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685. PMLR, 2019.

Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM Journal on Matrix Analysis and Applications*, 9(4):543–560, 1988.

Roger Grosse, Juhan Bae, Cem Anil, Nelson Elhage, Alex Tamkin, Amirhossein Tajdini, Benoit Steiner, Dustin Li, Esin Durmus, Ethan Perez, Evan Hubinger, Kamilė Lukošiūtė, Karina Nguyen, Nicholas Joseph, Sam McCandlish, Jared Kaplan, and Samuel R Bowman. Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*, 2023.

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

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011.

Roger A Horn and Charles R Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1994.

Jiri Hron, Yasaman Bahri, Jascha Sohl-Dickstein, and Roman Novak. Infinite attention: NNGP and NTK for deep attention. In *International Conference on Machine Learning*, pp. 4376–4386. PMLR, 2020.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5156–5165. PMLR, 2020.

Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pp. 1885–1894. PMLR, 2017.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. DataInf: Efficiently estimating data influence in LoRA-tuned LLMs and diffusion models. In *International Conference on Learning Representations*, 2024. arXiv preprint arXiv:2310.00902.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Evolution of neural tangent kernels under benign and adversarial training. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Elizbar A Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964.

Eshaan Nichani, Alex Damian, and Jason D Lee. Provable guarantees for nonlinear feature learning in three-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 27074–27113. PMLR, 2023.

Mark Rudelson and Roman Vershynin. The Littlewood–Offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

Mana Sakai, Ryo Karakida, and Masaaki Imaizumi. Infinite-width limit of a single attention layer: Analysis via Tensor Programs. In *Advances in Neural Information Processing Systems*, 2025. arXiv preprint arXiv:2506.00846.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

A. van der Sluis. Condition numbers and equilibration of matrices. *Numerische Mathematik*, 14(1): 14–23, 1969.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Geoffrey S Watson. Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 26(4):359–372, 1964.

Jonathan Wenger, Felix Dangel, and Agustinus Kristiadi. On the disconnect between theory and practice of neural networks: Limits of the NTK perspective. *arXiv preprint arXiv:2310.00137*, 2023.

Greg Yang. Tensor programs II: Neural tangent kernel for any architecture. *arXiv preprint arXiv:2006.14548*, 2020.

Rui Zhang and Shihua Zhang. Rethinking influence functions of neural networks in the over-parameterized regime. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8): 9082–9090, 2022. doi: 10.1609/aaai.v36i8.20893.

SUPPLEMENTARY MATERIAL

This supplementary provides extended proofs, detailed experimental methodology, additional results on Fashion-MNIST, and formal malleability theory supporting the main paper's findings.

Notation: Bold-faced letters represent vectors and matrices, with $[\mathbf{A}]_{ij}$ denoting the (i, j)-th entry of matrix $\mathbf{A}$. Throughout, $\mathbf{1} \in \mathbb{R}^n$ denotes the vector of all ones and $\mathrm{vec}(\cdot)$ denotes vectorization. $\boldsymbol{\mathcal{I}} \coloneqq (\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{y} \in \mathbb{R}^n$ denotes the *influence vector* whose i-th entry is the leave-one-out influence weight for training point i.

A EXTENDED LIMITATIONS

This section expands the limitations summarized in Section 7.

Parameter-free proxy and its trainable extension. The central result (Theorem 4.4) uses the parameter-free map $f^{\text{att}}(\mathbf{X}) = \mathbf{X}\mathbf{X}^T\mathbf{X}$, whose attention component is fixed and does not participate in the NTK dynamics. Section 4.1 analyzes the genuinely trainable setting, in which query, key, and value parameters enter the NTK of a single-layer linear-attention head with a linear output layer, and there the precondition floor and exponent-two persistence (Theorems 4.9–4.10) hold under both initializations analyzed, while the cubic mechanism transfers only under the checkable alignment condition of Proposition 4.11, a per-checkpoint question for a given optimizer. A trained head that adds a nonlinear MLP on top of the attention output (Equation 3) is shown empirically (Figure 1, Appendix B.15), while the theory here assumes a linear output layer.

Transductive formulation and the within-sequence case. The analyzed object is transductive attention over the training set, so $f^{\text{att}}(\mathbf{X}) = \mathbf{X}\mathbf{X}^\top\mathbf{X}$ acts on all n examples at once and $\kappa_d(\mathbf{G})$ is a dataset-level condition number. That is the right object here, because influence functions, and the methods built on them (TRAK, DataInf, EK-FAC), are training-set-level quantities whose reliability is governed by the dataset Gram. The precondition also holds on deployed models. The effective condition number exceeds the non-convergence threshold on the token-embedding matrices of more than twenty pretrained models and on a contextual DistilBERT representation at $\kappa_d \approx 10^7$ (Appendix B.13, B.17.3), the representations influence methods run on in practice.

The same mechanism governs within-sequence attention. For a sequence Gram $\mathbf{G}_{\text{seq}} = \mathbf{X}_{\text{seq}}\mathbf{X}_{\text{seq}}^\top \in \mathbb{R}^{L\times L}$ the eigenvalue identity $\kappa_d(\mathbf{G}_{\text{seq}}^3) = \kappa_d(\mathbf{G}_{\text{seq}})^3$ holds unchanged, the width requirement becomes $m = \Omega(\kappa_d(\mathbf{G}_{\text{seq}})^6 \log L)$, and the precondition floor $\kappa^+ \geq \kappa_d(\mathbf{G}_{\text{seq}})$ still applies. The one quantity the theory leaves open is the value of $\kappa_d(\mathbf{G}_{\text{seq}})$. That value depends on the tokens in the sequence, so it must be measured for a given model and dataset. The tokens are drawn from embedding spaces that already carry κ_d up to 10^7–10^9, and repetitive, padded, or near-uniform sequences raise it further. The obstruction lifts only for sequences whose tokens are well spread relative to the model dimension.

Linearization and softmax. The cubic conditioning result uses the first-order (identity-projection) proxy. Appendix B.14 shows softmax and exponential attention instead collapse to rank-1 kernels ($\cos(\phi_i, \phi_j) = 1.000$), a distinct failure mode consistent with Bae et al. (2022) and Sakai et al. (2025). The consequences for influence methods deployed on softmax transformers operating in this collapse regime are not characterized here.

Initialization-time scope. The width bound is derived at initialization. Figure 1 and Table 3 report post-training behavior, which corroborates the direction but lies outside the theorem's coverage. A post-training lower bound is left open.

Conditional benefit. The approximation-error reduction (Proposition D.15) holds under the single checkable spectral condition (S) on the target function and regularization. The inequality provably reverses on an $(n-d)$-dimensional cone of targets whenever $n > \mathrm{rank}(\mathbf{X})$ (the regime studied here), and can reverse elsewhere as well (e.g., low-κ_d subsets).

Incoherence assumption. The μ-incoherence condition (Assumption 4.3) behind the concentration step is stated for isotropic sub-Gaussian inputs and is not verified for natural images, so the specific thresholds ($m \gg 10^{26}$, 10^{22}) carry this conditional status while the qualitative infeasibility is supported empirically across every effective-conditioning cutoff. The $n > d$ floor rests on the

separate transformed-feature incoherence $\tilde{\mu} = \max_{i \neq j} |\hat{\mathbf{G}}_{ij}| < 1$ on the row-normalized $\mathbf{G}^3$, which is measured at $\tilde{\mu} \approx 0.99998$ (MNIST) and 0.99995 (CIFAR-10). These sit just below one, so $\mathbf{K}_{\infty,\text{seq}}$ is full rank and the resulting small $\log(1/\tilde{\mu})$ only enlarges the already-infeasible $n > d$ width, consistent with the measured non-convergence in Figure 1.

Scale and datasets. The controlled setting ($m \leq 4096$, MNIST/CIFAR-10) is what enables exact NTK computation. The cross-model diagnostic (Section 6, Appendix B.13) extends the precondition to deployed embeddings via randomized SVD in $O(nd \log n)$ (Halko et al., 2011), without gradient computation. A characterization of the contextualized representations produced by deeper trained layers remains open.

B Experimental Setup and Implementation Details

B.1 Architecture Specifications

The experimental framework compares two primary architectures to isolate the effect of linearized attention:

1. **Baseline (2L-ReLU):** Two-layer ReLU network with direct input processing
2. **Attention-Enhanced (MLP-Attn):** Sequential architecture with linearized attention preprocessing

Both architectures employ identical MLP components with the following specifications:

- **Hidden layer width:** $m = 1024$ neurons
- **First layer initialization:** $w_r \sim \mathcal{N}(0, \kappa^2 I_d)$ with $\kappa = 0.01$
- **Second layer weights:** $a_r \in \{-1, 1\}$
- **Activation function:** ReLU
- **Output scaling:** $1/\sqrt{m}$ normalization

The linearized attention mechanism transforms inputs via $f^{\text{att}}(\mathbf{X}) = \mathbf{X}\mathbf{X}^T\mathbf{X}$ before MLP processing, with model capacity held fixed across architectures so that observed differences are attributable to attention effects. Here $\mathbf{X} \in \mathbb{R}^{n \times d}$ denotes the *entire training set*, and the attention transformation is computed once over all training data, not per mini-batch, making the architecture transductive. This design ensures the data-dependent kernel sensitivity (Proposition C.24) reflects the full dataset correlation structure.

Scale normalization: To isolate structural properties of the architectures from input scale effects, all input data is normalized to unit ℓ_2 norm ($\|\mathbf{x}_i\|_2 = 1$) as assumed in Theorem 4.2. Additionally, the attention-transformed features $f^{\text{att}}(\mathbf{X})$ are row-normalized before being passed to the MLP, ensuring comparable input magnitudes across architectures.

B.2 Binary CIFAR-10 Spectral Argument

The binary CIFAR-10 subset (cars vs. airplanes) yields an $\approx 1\times$ malleability ratio, matching 2L-ReLU and well below the 6–9$\times$ gap seen in 10-class CIFAR-10, a difference traceable to the spectral mechanism described below.

Clarification on $\kappa_d(\mathbf{G})$ computation. The Gram matrix $\mathbf{G} = \mathbf{X}\mathbf{X}^T$ depends only on the input features $\mathbf{X}$, not on labels. *Binary CIFAR-10* here means the binary-class input *subset*: $\mathbf{X}$ consists only of cars and airplanes images (different from the full 10-class $\mathbf{X}$). The condition number $\kappa_d(\mathbf{G})$ is computed on this binary-subset Gram matrix, not the full-dataset matrix. The distinction matters. Changing from 10-class to binary relabeling on the same input set would leave $\kappa_d(\mathbf{G})$ unchanged, while selecting a 2-class *input subset* changes $\mathbf{X}$ and hence $\mathbf{G}$.

Cars and airplanes are visually similar, so their images concentrate in a lower-dimensional structure with a longer near-zero eigenvalue tail. Because $\kappa_d(\mathbf{G}) = \lambda_{\max}/\lambda^+_{\min}$ is set by that tail and not by the spread of the top eigenvalues, the binary subset is in fact *more* ill-conditioned than the full dataset: at matched sample count $\kappa_d(\mathbf{G}_{\text{binary}}) \approx 5.8 \times 10^7$, above the 4.4×10^7 of 10-class CIFAR-10 (and

8.5×10^7 at the binary set's own $n = 2000$). The narrowing of the malleability ratio to $\approx 1\times$ is therefore not a lower-κ_d effect. Table 1 shows it is driven by the 2L-ReLU flip rate rising from 3.1% to 15.5% (PGD) while MLP-Attn barely moves (19.1% to 14.0%), and the binary setting changes both the inputs (cars and airplanes only) and the objective (MSE in place of cross-entropy), each of which raises 2L-ReLU's sensitivity. The binary comparison thus does not isolate κ_d. The controlled single-variable test of the conditioning effect is the whitening experiment of Appendix B.17.3, where the same representation at $\kappa_d \approx 10^7$ against a whitened $\kappa_d = 183$ moves the flip rate from 61.3% to 0.6%.

B.3 Dataset Configuration and Preprocessing

The experimental evaluation employs three standard computer vision datasets with specific preprocessing protocols, following established benchmarking practices (Krizhevsky & Hinton, 2009):

B.3.1 Dataset Specifications

Table 2: Dataset specifications. Dim. = input dimension, Norm. = (mean, std) normalization parameters. Standard training set sizes are used.

Dataset	Dim.	Classes	Norm.
MNIST	784	10	(.13, .31)
Fashion-MNIST	784	10	(.29, .35)
CIFAR-10	3072	10	(.5, .5, .5)

B.3.2 Class Selection and Data Sampling

The experimental design considers two classification scenarios:

- **Binary Classification:** Select representative class pairs (e.g., MNIST digits 3 vs 8, CIFAR-10 cars vs planes)
- **Multi-class Classification:** All ten (10) classes.

Main results (Tables 1, 3, and 9, Figure 1) use the **10-class setting**. Table 1 (main paper) reports both 10-class and binary classification results, with the binary setting following Zhang & Zhang (2022). Epoch-by-epoch training progression across all datasets and intervention types is provided in Table 10 (Appendix F).

B.4 Training Configuration

B.4.1 Optimization Parameters

- **Optimizer:** Adam with learning rate $\eta = 10^{-3}$
- **Batch size:** 128 samples
- **Training epochs:** 500
- **Regularization:** $\lambda = 10^{-3}$ for kernel ridge regression
- **Mixed precision:** enabled

B.4.2 Loss Function and Regularization

The training objective incorporates both task loss and parameter regularization:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \ell(f(\mathbf{x}_i; \boldsymbol{\theta}), y_i) + \frac{\lambda}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$$

where:

- $\ell(\cdot, \cdot)$ is cross-entropy loss for multi-class or MSE loss for binary classification
- $\boldsymbol{\theta}_0$ represents initial parameters
- $\lambda = 10^{-3}$ provides regularization strength

B.4.3 NTK Distance Measurement

The distance $\|f_m - f_{\text{NTK}}\|_2$ is measured across widths $m \in \{4, 8, 16, \ldots, 4096\}$, the ceiling at which the NTK can still be computed exactly. Since $\kappa_d(\mathbf{G})^6 \gg m$ for all $m \lesssim 10^5$, the conclusion of Theorem 4.4 holds at any practical width (§7). Nyström approximation is avoided throughout because its estimation error would conflate with the convergence signal being measured. The initialization scale is $\mathbf{w}_r \sim \mathcal{N}(0, \kappa^2 \mathbf{I}_d)$ with $\kappa = 0.01$, and the width requirement $m = \Omega(\kappa_d(\mathbf{G})^6 \log n)$ holds for this and the standard initialization families (He, Xavier, LeCun) without structural change.

Table 3: NTK distance $\|f_m - f_{\text{NTK}}\|$ across network widths. 2L-ReLU exhibits monotonically decreasing distance, consistent with convergence to the kernel regime. MLP-Attn shows qualitatively different behavior, non-monotonic across all widths on MNIST (Figure 1) and monotonically increasing on CIFAR-10.

Model	m**=16**	m**=1024**	m**=4096**
MNIST			
2L-ReLU	45.1	39.9	39.2
MLP-Attn	10.3	33.3	43.4
CIFAR-10			
2L-ReLU	246.2	101.7	56.9
MLP-Attn	3.7	10.4	12.6

B.4.4 Adversarial Perturbation Methods

Influence malleability is evaluated under three attacks of increasing strength: FGSM (Fast Gradient Sign Method, a single-step attack), PGD (Projected Gradient Descent, step size $\alpha = 0.007$, 40 steps), and MIM (Momentum Iterative Method, PGD augmented with a momentum term). All three are run at $\varepsilon = 0.03$ in L_∞ norm, applied to the top-10% influential examples ($\tau = 0.1$), matching the perturbation budget and threshold used throughout Section 5. The perturbation is applied in the standardized input space (the mean/variance-normalized pixels the network receives) before the model's internal ℓ_2 normalization, so $\varepsilon = 0.03$ is a small perturbation of the raw input and not of a unit-norm feature vector.

B.5 Evaluation Metrics and Analysis

B.5.1 Influence Function Computation

Following the NTK-based approach of Zhang & Zhang (2022) and the foundational influence function methodology of Koh & Liang (2017), influence is computed using the *empirical finite-width NTK* $\mathbf{K}_m$ at each width m. For each trained model, the NTK is computed as $[\mathbf{K}_m]_{ij} = \langle \nabla_{\boldsymbol{\theta}} f(\mathbf{x}_i), \nabla_{\boldsymbol{\theta}} f(\mathbf{x}_j) \rangle$ using the actual model gradients. This approach ensures that (1) influence comparisons between architectures are fair (both evaluated at the same width), and (2) the influence patterns reflect the actual model dynamics. The divergence of MLP-Attn from its infinite-width limit is consistent with this methodology, since the empirical NTK captures how each architecture actually represents training data relationships at practical widths.

B.5.2 Malleability Metrics

The analysis employs several quantitative measures:

1. **Influence Flip Rate:**

$$\text{Flip Rate} = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \mathbf{1}[\text{sign}(I_i^{\text{orig}}) \neq \text{sign}(I_i^{\text{adv}})]$$

where $\mathcal{H}$ denotes the set of high-influence training examples (top 10% positive influence).

2. **Rank Correlation (Spearman's ρ):**

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between ranks of influence scores before and after perturbation.

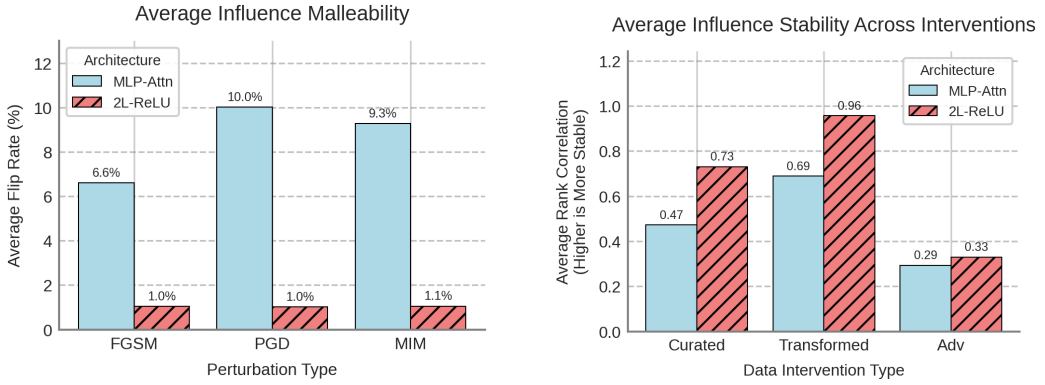
3. **Top-K Stability:**

$$\text{Stability}_K = \frac{|\mathcal{T}_K^{\text{orig}} \cap \mathcal{T}_K^{\text{adv}}|}{K}$$

where $\mathcal{T}_K^{\text{orig}}$ and $\mathcal{T}_K^{\text{adv}}$ are the sets of top-K influential examples before and after perturbation.

B.6 Empirical Analysis of Influence Stability and Malleability

The **2L-ReLU** network and the attention-enhanced **MLP-Attn** architecture are compared on two quantities, the stability of influence rankings under data interventions and the malleability of influence signs under adversarial attacks, using the metrics defined above.



(a) Influence flip rate by perturbation type.

(b) Influence rank correlation by intervention strategy.

Figure 2: Analysis of influence dynamics. (a) MLP-Attn exhibits consistently higher influence malleability (flip rate) across all perturbation types, reflecting its operation in the feature learning regime. (b) Rank correlation analysis reveals that 2L-ReLU maintains rigid influence rankings while MLP-Attn shows lower correlation, indicating continuous re-evaluation of data dependencies. The *Transformed* intervention (replacing influential examples with adversarial versions) produces the most structured adaptation pattern.

B.6.1 Influence Stability Across Data Interventions

Influence stability is assessed using **Rank Correlation (Spearman's ρ)**. A higher correlation indicates a more rigid influence structure, while a lower correlation suggests a more flexible, adaptive re-ranking of training point importance. Two distinct stability patterns emerge.

- For the curated and transformed interventions, stability for both architectures is constant throughout training across both classification settings (binary and 10-class), so the relative importance of training examples is a fixed structural property in these settings. In nearly all settings the 2L-ReLU architecture holds a significantly higher Spearman's ρ, a more rigid influence structure than the MLP-Attn model.
- The adversarial ('Adv') intervention induces dynamic stability profiles that evolve over training. For the more complex Fashion-MNIST and CIFAR-10 datasets, both architectures generally show a decline in stability. On the simpler MNIST dataset, the 2L-ReLU model recovers and increases its influence stability over time, a behavior the attention-enhanced model does not show.

B.6.2 Influence Malleability Under Adversarial Attacks

Influence malleability is evaluated using the **Influence Flip Rate**, which measures the percentage of high-influence examples that reverse their helpful/harmful status after being perturbed.

Influence malleability differs by attack strength.

- The stronger PGD attack reveals more consistent trends. On Fashion-MNIST, the MLP-Attn model is more malleable than 2L-ReLU. On the more complex CIFAR-10 dataset, the results are mixed, with relative malleability depending on the number of classes.
- The weaker FGSM attack produces more erratic results than PGD. On Fashion-MNIST, flip rates for both models fluctuate across training, while on CIFAR-10 the binary and 10-class contrast is consistent with the spectral argument (Appendix B.2).

B.6.3 Synthesis and Insights

These results support the central thesis of NTK non-convergence and its dual implications. The **2L-ReLU** architecture, which converges toward its infinite-width NTK limit, shows *high stability* and *low malleability*, with a strong and rigid dependence on training data consistent with kernel-regime behavior. The **MLP-Attn** architecture, which stays in the feature-learning regime, shows *lower stability* and *higher malleability*.

This trade-off carries dual implications. The architecture continuously re-evaluates data dependencies, which allows flexible data relationships and at the same time makes it more susceptible to adversarial manipulation of training data, as quantified by the elevated Influence Flip Rate.

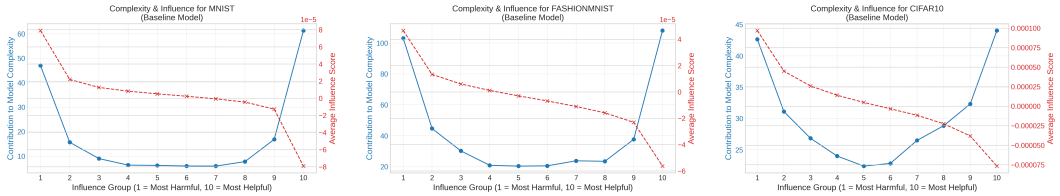
B.7 Reproducibility and Validation

Reproducibility rests on two elements.

1. Fixed random seeds (42) across all random-number generators, with deterministic operations enabled.
2. Identical preprocessing pipelines for all datasets, with fixed normalization parameters and feature scaling.

B.8 Model Complexity Analysis

The relationship between influence scores and model-complexity contributions is shown in Figure 3. The training data is partitioned into 10 groups by influence score, ordered from most harmful (group 1) to most helpful (group 10).



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

Figure 3: Contribution to model complexity (blue) and average influence score (red) for MLP-Attn across three datasets. The U-shaped complexity curve indicates that the most influential points (both harmful and helpful) contribute most to model complexity, consistent with findings of Zhang & Zhang (2022).

Across all three datasets, influence malleability concentrates on the training examples that matter most for model complexity. This connects the malleability metric to the RKHS complexity of the learned function, since high-malleability architectures like MLP-Attn are most sensitive precisely at the training points that contribute most to model capacity.

B.9 Numerical Stability Verification

Numerical stability is verified at each experimental configuration through four checks.

1. Positive definiteness, $\lambda_{\min}(\mathbf{K}) > -\tau$ with $\tau = 10^{-12}$.
2. Conditioning, $\kappa(\mathbf{K} + \lambda\mathbf{I}) < 10^{12}$.
3. Inversion accuracy, $\|(\mathbf{K} + \lambda\mathbf{I})^{-1}(\mathbf{K} + \lambda\mathbf{I}) - \mathbf{I}\|_F < \tau$.
4. Symmetry, $\|\mathbf{K} - \mathbf{K}^T\|_F < \tau$.

All configurations passed these checks.

B.10 Influential Training Examples

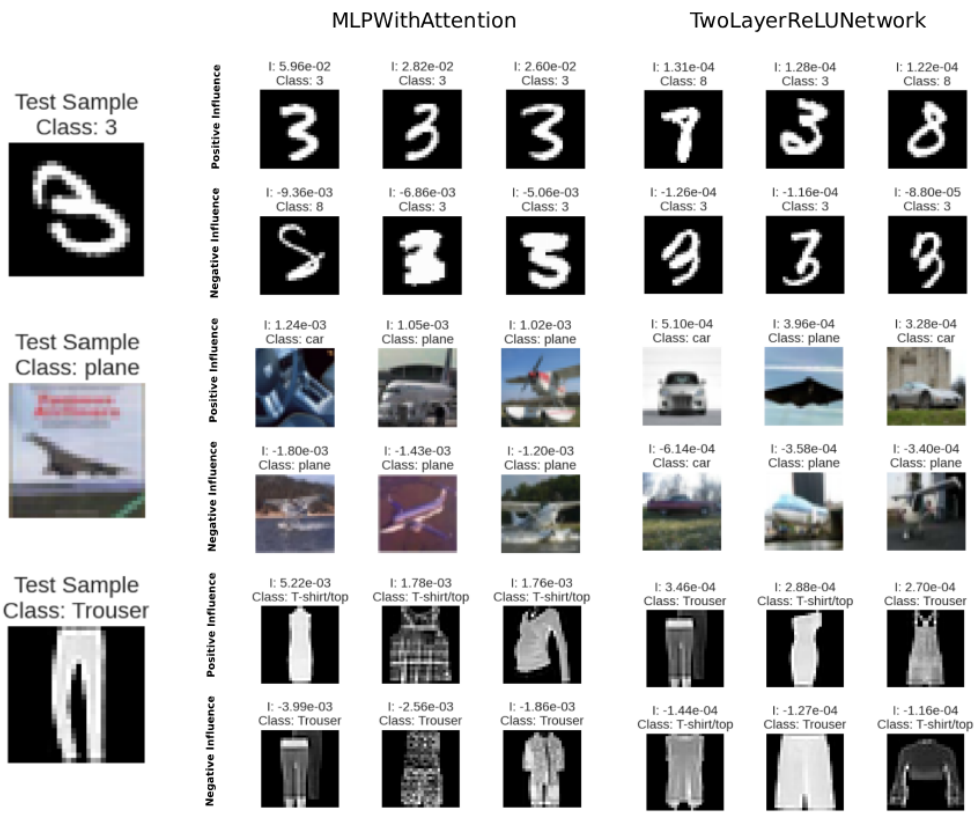


Figure 4: Top influential training examples for a representative test digit. Positive influencers (top rows) increase test loss when removed, and negative influencers (bottom rows) decrease it. The key distinction is quantitative, with MLP-Attn flip rate 28.9% vs. 3.3% for 2L-ReLU under PGD (Table 1), reflecting the Gram-induced kernel's data-dependent structure.

B.11 Loss Landscape Analysis

The 3D loss landscape for both architectures at width $m = 1024$ on MNIST is shown in Figure 5, visualized along two random directions in parameter space.

The landscape geometry reflects the dual implications. The broader minimum of MLP-Attn corresponds to the feature-learning regime (NTK non-convergence), where the model continuously adapts its representation. This geometric flexibility underlies both the higher influence malleability and the sensitivity to training-data perturbations.

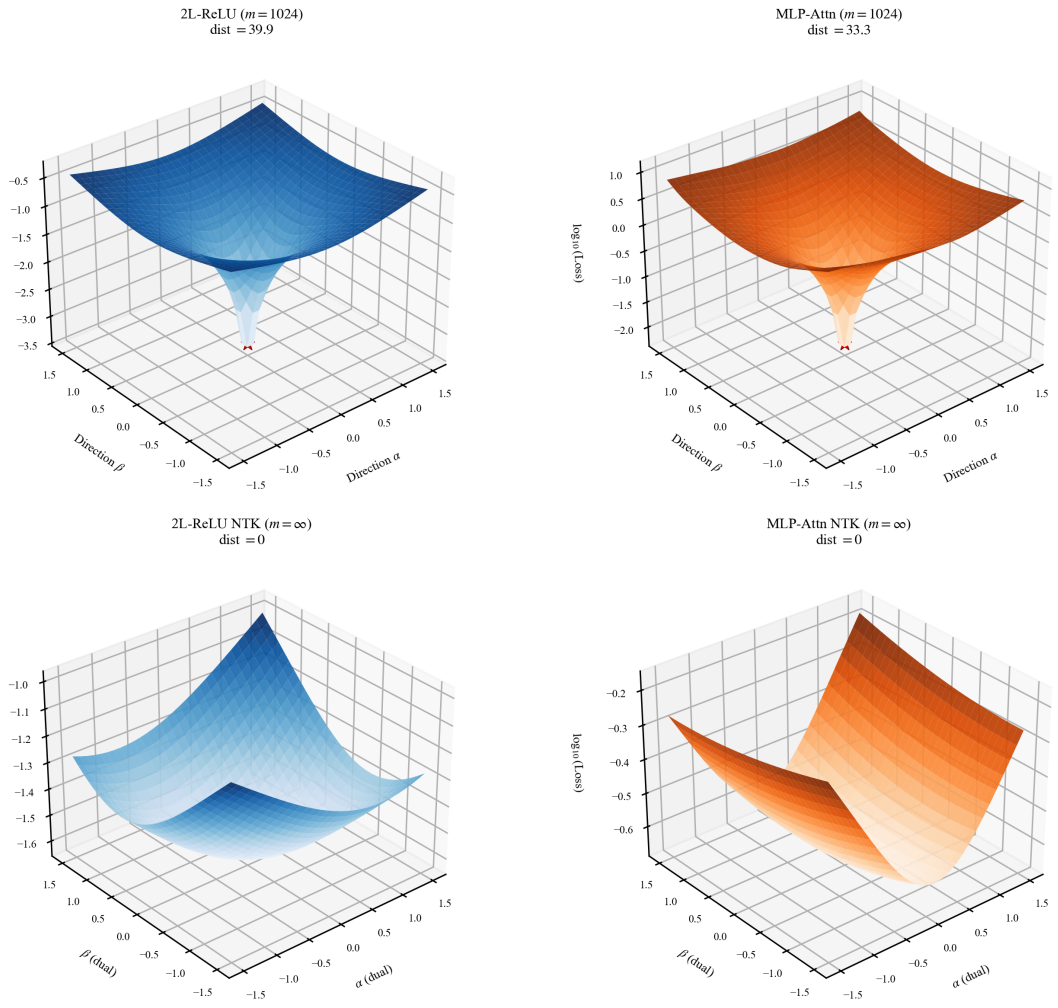


Figure 5: Loss landscape comparison (MNIST) at finite width $m = 1024$ (left two panels) and infinite-width NTK limit (right two panels). **2L-ReLU** (blue) converges toward its NTK landscape: both finite and infinite-width surfaces share similar geometry. **MLP-Attn** (orange) shows a qualitatively different finite-width landscape (sharp, deep minimum) compared to its NTK limit (broad, shallow basin), visualizing the NTK non-convergence from a loss geometry perspective. NTK distances: 2L-ReLU $= 39.93$, MLP-Attn $= 33.30$.

B.12 Practitioner Guidance for Influence Methods on Attention Architectures

The following steps operationalize the non-convergence result for practitioners deploying TRAK (Park et al., 2023), DataInf (Kwon et al., 2024), or EK-FAC (Grosse et al., 2023) on attention-based models.

Diagnostic. Compute $\kappa_d(\mathbf{G})$ via randomized SVD (Halko et al., 2011) in $O(nd \log n)$ time. This requires no gradient computation and no training. Values exceeding 10^2 place the dataset in the non-convergence regime: MNIST yields $\kappa_d \approx 9 \times 10^8$, CIFAR-10 yields $\kappa_d \approx 4 \times 10^7$. In both cases $m = \Omega(\kappa_d^6 \log n)$ is astronomically large, and even the tight row-normalized requirement κ_d^3 exceeds 10^{26} (MNIST) and 10^{22} (CIFAR-10), so no practical width satisfies the kernel regime assumption required by these tools' theoretical guarantees. Storing 10^{26} neurons at single precision would require $\sim 4 \times 10^8$ exabytes, fourteen orders of magnitude beyond the largest known architectures ($\sim 10^{12}$ parameters), with MNIST the more ill-conditioned of the two because its constant-background pixels leave many near-zero Gram directions.

Data auditing and poisoning detection. The 6–9× higher flip rates imply that influence scores for attention are less stable than for ReLU networks. Three adjustments help. Lower the confidence threshold for labeling examples *harmful* or *helpful*, ensemble influence scores across training checkpoints, and give borderline high-influence examples added scrutiny before removal.

Active learning. Influence-based active learning pipelines (BADGE, coreset methods) assume stable rankings. For attention at practical widths, rankings are structurally unstable as a consequence of Theorem 4.4. Re-rank after each training phase, since only a freshly computed ranking stays valid.

Attenuation. In binary classification with homogeneous classes (e.g., cars vs. airplanes), inter-class spectral spread is lower, yielding a smaller $\kappa_d(\mathbf{G})$ and reducing the cubic amplification. The diagnostic above quantifies this directly and is dataset-specific.

Adversarial training equivalence. Adversarial training raises 2L-ReLU malleability from 3.3% to 43.4%, matching attention's structural level. Practitioners combining attention with adversarial training face compounded instability and should validate influence scores across multiple perturbation budgets $\epsilon \in \{0.1, 0.2, 0.3, 0.5\}$.

B.13 Conditioning Across Language, Vision, and Audio Models

To assess whether the non-convergence regime identified for natural-image datasets extends to deployed architectures, the effective condition number diagnostic was applied to the embedding matrices of a range of publicly available models spanning three modalities. For each model an embedding matrix $\mathbf{E}$ was row-normalized to unit ℓ_2 norm, matching the preprocessing used throughout, and $\kappa_d(\mathbf{G}) = \lambda_1/\lambda_r$ of the Gram matrix $\mathbf{G} = \mathbf{E}\mathbf{E}^\top$ was computed, with λ_r the smallest positive eigenvalue. For a language model, $\mathbf{E}$ is the vocabulary-by-dimension token-embedding table. For a vision or audio model it is the patch- or convolution-projection weight reshaped to output-by-input, and the spectrum length is the smaller of the two dimensions. Eigenvalues were obtained as squared singular values via exact SVD, or, for the largest vocabularies, via exact eigendecomposition of the small $d \times d$ matrix $\mathbf{E}^\top\mathbf{E}$ (whose spectrum equals the nonzero spectrum of $\mathbf{G}$), and verified against the randomized SVD of Halko et al. (2011), with the estimates agreeing to the reported precision.

The obstruction reaches the state-space and linear-attention family (Mamba, RWKV) to which the analysis conjectures it should generalize (Appendix E) and spans a third modality, audio (via Whisper's decoder embedding and its raw audio convolutions), in addition to vision (CLIP, ViT, and a ResNet convolution stem). The two square vision patch-projections (marked †) are near-rank-deficient, inflating κ_d to 10^{10}–10^{11}, with the cleaner non-degenerate vision value given by CLIP's 2.3×10^4. Language, vision, and audio embedding spaces therefore fall within the same non-convergence regime as the natural-image datasets studied here, so the cubic conditioning obstruction extends beyond vision data and reaches the settings in which influence methods such as TRAK (Park et al., 2023), DataInf (Kwon et al., 2024), and EK-FAC (Grosse et al., 2023) are deployed. This measurement is restricted to the static input embedding. It establishes that the precondition for the obstruction, a large effective condition number, holds broadly, while the contextualized representations produced by deeper trained layers are left for later work.

B.14 Extended NTK Validation: Softmax and Exponential Attention

Two additional attention variants are evaluated to place the non-convergence of linearized attention in context, **softmax attention** and **exponential attention**. Both use the same transductive setup as the main experiment, with the reference set fixed to the full training split before the width is varied.

Softmax attention applies row-wise normalization to the score matrix:

$$\phi_{\text{sm}}(\mathbf{x}) = \text{row-}\ell_2\text{-norm}\left(\text{softmax}\left(\frac{\hat{\mathbf{x}}^\top\hat{\mathbf{X}}}{\sqrt{d}}\right)\hat{\mathbf{X}}\right), \tag{14}$$

where $\hat{\mathbf{x}} = \mathbf{x}/\|\mathbf{x}\|_2$, $\hat{\mathbf{X}}$ is the row-normalized training matrix, and row-ℓ_2-norm$(\cdot)$ scales each output row to unit ℓ_2 norm (a single row for the per-point map, applied row-wise when acting on the full matrix $\mathbf{X}$).

Table 4: Effective condition number $\kappa_d(\mathbf{G})$ of row-normalized embedding Gram matrices across language, state-space/linear-attention, vision, and audio models. The language-model group includes an encoder model (RoBERTa-base) alongside two frontier mixture-of-experts models (DeepSeek-V3, Kimi-K2), measured via their token-embedding shard. The DistilBERT row marked ‡ is a contextual-representation datapoint (AG-News texts mean-pooled through the frozen model), showing the obstruction reaches contextual representations as well as static token tables, while all other rows are static token or patch/convolution embeddings. Every value exceeds the threshold ($\kappa_d \approx 10^2$) that places a dataset in the non-convergence regime, and several reach the 10^7–10^9 range of the natural-image datasets studied here. Values marked † are inflated by near-degenerate directions in the learned projection.

Model	d	$\kappa_d(\mathbf{G})$
Language models (token embeddings)		
GPT-2	768	1.7×10^7
distilGPT-2	768	8.2×10^6
GPT-2-large	1280	5.3×10^3
BERT-base	768	9.4×10^3
DistilBERT (contextual)‡	768	1.0×10^7
RoBERTa-base	768	2.4×10^3
T5-small	512	1.0×10^5
Switch-base-8	768	2.3×10^4
Qwen2.5-0.5B	896	2.4×10^3
OLMo-1B	2048	1.2×10^3
Llama-3.2-1B	2048	3.2×10^3
TinyLlama-1.1B	2048	1.4×10^2
GPT-OSS-20B	2880	6.4×10^9
DeepSeek-V3	7168	4.1×10^2
Kimi-K2	7168	7.4×10^2
CLIP (text)	512	1.6×10^2
SmolVLM (text)	576	6.0×10^3
State-space / linear-attention models		
Mamba-130m	768	1.4×10^6
RWKV-4-169m	768	3.4×10^2
Vision (patch / convolution projections)		
CLIP (vision)	768	2.3×10^4
ViT-B/16	768	4.2×10^{10} †
SmolVLM (vision)	768	2.8×10^{11} †
ResNet-50 (conv stem)	147	3.7×10^3
Audio (Whisper)		
Whisper (decoder embedding)	384	7.8×10^4
Whisper (audio convolutions)	240–1152	1.4×10^3–3.4×10^5

Exponential attention omits the normalizing denominator, retaining only the unnormalized exponential weights:

$$\phi_{\text{exp}}(\mathbf{x}) = \text{row-}\ell_2\text{-norm}\left(\exp\left(\frac{\hat{\mathbf{x}}^\top \hat{\mathbf{X}}}{\sqrt{d}} - c(\mathbf{x})\right)\hat{\mathbf{X}}\right), \tag{15}$$

where $c(\mathbf{x}) = \max_j \hat{\mathbf{x}}^\top \hat{\mathbf{x}}_j / \sqrt{d}$ is a per-row stability shift that cancels exactly under ℓ_2-normalization. Unlike softmax, the unnormalized exponential amplifies score differences, so the aggregation is not a simple weighted mean. This corresponds to the Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964) with exponential kernel and no bandwidth normalization.

Figure 6 shows the NTK distance $1 - \cos(K_m, K_\infty)$ across widths. Both softmax and exponential attention exhibit near-zero distance at all widths, which at first glance appears to indicate rapid convergence. Figure 7 reveals this to be a measurement artifact: the feature maps ϕ_{sm} and ϕ_{exp} collapse to a single direction for all inputs (mean pairwise cosine similarity $= 1.000$ on both MNIST

and CIFAR-10). When all $\phi(\mathbf{x}_i)$ are identical, both K_m and K_∞ degenerate to rank-1 matrices and their cosine similarity is trivially 1, independent of width.

The collapse mechanism for softmax attention is well understood: at standard temperature $T = \sqrt{d}$, all scores $\hat{\mathbf{x}}^\top \hat{\mathbf{x}}_j / \sqrt{d}$ fall in a narrow range for high-dimensional inputs ($d = 784$ for MNIST, $d = 3072$ for CIFAR-10), making the softmax distribution nearly uniform. The aggregated feature $\phi_{\text{sm}}(\mathbf{x}) \approx \ell_2\text{-norm}(\bar{\mathbf{x}})$ for all $\mathbf{x}$, where $\bar{\mathbf{x}}$ is the training set mean. This collapse belongs to the unit-ℓ_2 input convention adopted here for comparability across architectures, since a deployed softmax layer receives features whose components have $O(1)$ variance (norm $\approx\sqrt{d}$ after LayerNorm), for which the same $1/\sqrt{d}$ scaling leaves the logits at $O(1)$ spread and the attention stays selective. A scaling ablation confirms this, with the mean pairwise $\cos(\phi_i, \phi_j)$ falling from ≈ 1 at the unit-norm operating point to 0.70 (MNIST) and 0.35 (CIFAR-10) once the logits reach $O(1)$ scale, so characterizing softmax attention at deployed scale is left open. Exponential attention avoids this via unnormalized weights, but the large-d concentration of inner products similarly causes exponential weights to concentrate near uniform, yielding the same collapse.

In contrast, linearized attention and 2L-ReLU maintain diverse feature maps (mean pairwise cosine similarity 0.754 and 0.318 on MNIST; 0.066 and 0.044 on CIFAR-10), confirming that the non-trivial NTK distances measured for these architectures reflect genuine structural differences between K_m and K_∞, not a measurement artifact.

The softmax curve in Figure 1. The main-text softmax curve is a separate measurement from the $1 - \cos(K_m, K_\infty)$ collapse analysis above, built to sit on the same footing as the 2L-ReLU and MLP-Attn curves. It uses the function-space distance $\|f_m - f_{\text{NTK}}\|/\sqrt{n_{\text{test}}}$ with a parameter-free softmax attention block feeding the same MLP, under the training protocol shared with those two models. The scores use the standardized features at the deployed scaling, $\mathrm{softmax}(\mathbf{X}\mathbf{X}^\top/\sqrt{d})$ with $O(1)$-variance components in place of unit-ℓ_2 rows, so the logits carry $O(1)$ spread and the rank-one collapse above does not occur. The reference f_{NTK} is the arc-cosine kernel on the *fixed* softmax-transformed features, a static infinite-width target of the same kind used for the other two models, in place of a reference read off a trained checkpoint. The softmax distance is then non-monotonic on MNIST, with a shallow minimum near $m = 8$ and a rise to ≈ 30 at $m = 4096$, and monotone increasing on CIFAR-10 to ≈ 12, tracking MLP-Attn and staying an order of magnitude above the converging 2L-ReLU model. Softmax attention at deployed scale therefore stays in the feature-learning regime. A small learning rate keeps the finite-width networks near their linearization throughout the measurement.

The two figures measure non-convergence in different ways. The NTK distance $1 - \cos(K_m, K_\infty)$ used in Figure 6 is *scale-invariant*, treating K_m and K_∞ as unit vectors of their entries and measuring only the angle between them, the eigenvector alignment. The function-space distance $\|f_m - f_{\text{NTK}}\|$ used in Figure 1 is *magnitude-sensitive*, capturing both eigenvector misalignment and eigenvalue amplification. For linearized attention, Theorem 4.4 establishes that the effective condition number scales as $\kappa_d(G)^3$, so eigenvalues are cubically amplified relative to the reference kernel. At large widths, the *eigenvectors* of K_m may align with those of K_∞ (low cosine distance), while the *eigenvalues* remain severely amplified (high function-space distance). The two figures therefore agree, with Figure 6 isolating structural misalignment and Figure 1 capturing total deviation including spectral amplification. The failure lies in the spectral scaling imposed by the cubic Gram structure, while the directional geometry of the kernel stays aligned. The softmax curves in the two figures sit at different input scalings, and using both is deliberate. Figure 6 uses the unit-ℓ_2 normalization shared by all four architectures there, which places softmax on the same footing as the others but drives it into the collapse of Appendix B.14, so its $1 - \cos$ distance is vacuous. Figure 1 then measures softmax alone at the deployed scaling, where the logits carry $O(1)$ spread and the collapse is absent. Its function-space distance still grows with width, so the non-convergence holds both under the shared normalization and at the realistic operating point. The other two architectures need no such check, since they carry no temperature that a normalization choice can over-cool.

Proposition B.1 (Spectral Amplification of the Matrix-Exponential Surrogate)**.** *Let* $\mathbf{G} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ *with positive eigenvalues* $\lambda_1 \geq \cdots \geq \lambda_r > 0$*, where* $r = \mathrm{rank}(\mathbf{G})$ *(for rank-deficient image data* λ_r *is the smallest positive eigenvalue, not* λ_n*). Consider the matrix-exponential surrogate of unnormalized attention with identity projections,* $f^{\text{unnorm}}(\mathbf{X}) = \exp(\mathbf{G}/\sqrt{d})\,\mathbf{X}$*, where* $\exp(\cdot)$ *denotes the* matrix *exponential. This is a tractable proxy analyzed for its spectral behavior. Entrywise softmax is a different operator whose leading term is a rank-one all-ones component, and it collapses*

to a rank-1 kernel at practical temperatures. For the surrogate, the induced kernel matrix is:

$$\tilde{\mathbf{G}}^{\mathrm{unnorm}} = \mathbf{U}\,\mathrm{diag}\Big(\lambda_i\, e^{2\lambda_i/\sqrt{d}}\Big)\,\mathbf{U}^\top \tag{16}$$

with condition number:

$$\kappa_d\Big(\tilde{\mathbf{G}}^{\mathrm{unnorm}}\Big) = \kappa_d(\mathbf{G}) \cdot \exp\left(\frac{2(\lambda_1 - \lambda_r)}{\sqrt{d}}\right) \tag{17}$$

This is strictly larger than the linearized bound $\kappa_d(\mathbf{G})^3$ whenever $\lambda_1 - \lambda_r > \sqrt{d}\,\log \kappa_d(\mathbf{G})$, a condition satisfied for MNIST and CIFAR-10 at all tested widths. Proof in Appendix C.

Remark B.2 (Exponential Softmax Amplification). Linearized attention yields polynomial amplification $\kappa(\mathbf{G})^3$; unnormalized softmax yields $\kappa(\mathbf{G}) \cdot \exp(2(\lambda_1 - \lambda_n)/\sqrt{d})$, exponential in the spectral gap, and strictly larger than $\kappa(\mathbf{G})^3$ whenever $\lambda_1 - \lambda_n > \sqrt{d}\,\log \kappa(\mathbf{G})$. The linearized model corresponds to the first-order Taylor truncation $\exp(\mathbf{G}/\sqrt{d}) \approx \mathbf{I} + \mathbf{G}/\sqrt{d}$; higher-order terms only increase spectral amplification. Normalized softmax does not preserve this structure at the standard temperature $T = \sqrt{d}$, where for high-dimensional inputs it instead collapses to a rank-one kernel (Remark B.3), consistent with Bae et al. (2022) and Sakai et al. (2025).

Remark B.3 (Reconciling Proposition B.1 with Empirical Collapse). Proposition B.1 analyzes *unnormalized* softmax $f^{\mathrm{unnorm}}(\mathbf{X}) = \exp(\mathbf{G}/\sqrt{d})\,\mathbf{X}$, yielding a kernel with large but finite condition number. The empirical results above use *row-normalized* softmax (standard attention), where each row of the attention matrix is divided by its sum. Row-wise normalization concentrates attention weights: at temperature $T = \sqrt{d}$, high-dimensional inputs cause all scores to cluster near uniform, making $\phi_{\mathrm{sm}}(\mathbf{x}) \approx \ell_2$-norm$(\bar{\mathbf{x}})$ for all $\mathbf{x}$ and collapsing the kernel to rank 1. The two analyses describe different regimes. Unnormalized softmax produces a large-κ regime (Proposition B.1), while normalized softmax undergoes feature collapse to $\kappa = \infty$ (rank-1 kernel, $\lambda_2 = 0$), a qualitatively distinct failure mode. Both escape the NTK kernel regime, since genuine convergence requires the limiting kernel K_∞ to be positive definite, and a rank-1 kernel carries no information about input geometry and supports no nontrivial prediction. The near-zero NTK distance observed for softmax and exponential attention is a measurement artifact, since both K_m and K_∞ degenerate to the same rank-1 matrix independently of width, making cosine similarity trivially unity regardless of whether any convergence has occurred.

B.15 Trained Softmax Attention with Learnable Projections

The theorems above analyze linearized attention, and the collapse analysis uses parameter-free softmax at identity projections. To probe whether the obstruction persists once the attention projections are genuinely trained, a single-head softmax attention block with learnable query, key, and value maps ($\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$, initialized away from any whitening configuration) feeding a ReLU output layer is trained end-to-end on MNIST, alongside a two-layer ReLU model and the parameter-free proxy $f(\mathbf{X}) = \mathbf{X}\mathbf{X}^\top\mathbf{X}$, all under mean-squared-error regression on one-hot targets (widths 256–2048, 1000 training points).

The attended-feature Gram of the trained head is ill-conditioned at $\kappa_d \approx 9 \times 10^{11}$ at every width, the same order as the parameter-free proxy (9.8×10^{11}) and about $330\times$ the input conditioning, so the trained projections leave the conditioning amplified (Figure 8, right, with values in Table 5). The function-space NTK distance of the trained head stays near 0.45 across the width range, about twice that of the two-layer ReLU model, whose distance decreases monotonically toward its kernel (Figure 8, left). Once the projections are trained jointly with the output layer the head has no closed-form infinite-width kernel, so its distance is measured against a conditional reference, the arc-cosine kernel evaluated on the head's own learned attention output. The influence flip rate is not reported at this width and sample scale. The same trainable-QKV softmax head also appears in Figure 1, showing the same feature-learning behavior across all widths on both MNIST and CIFAR-10 under cross-entropy. The measurement here adds the conditioning diagnostic and a check under mean-squared-error loss.

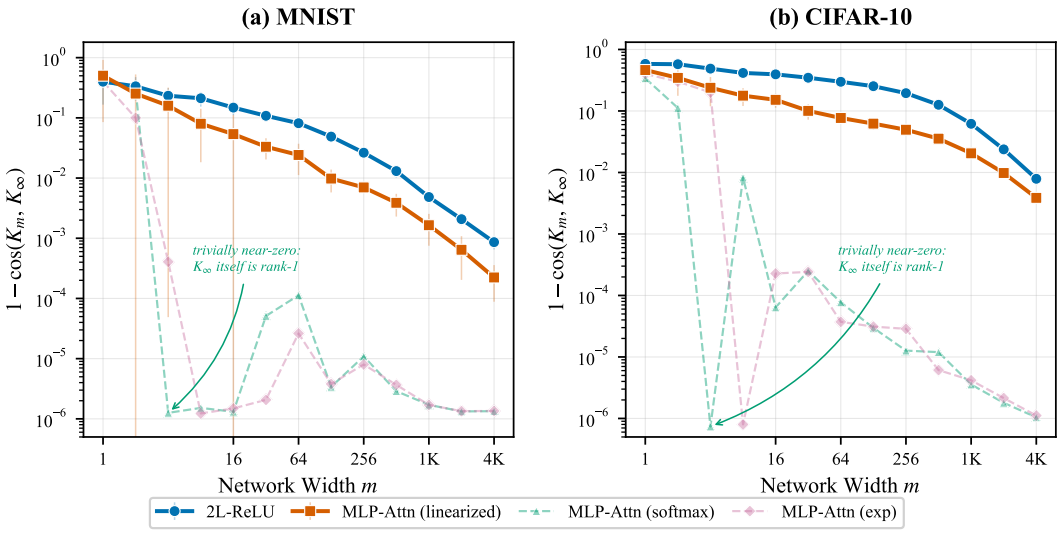


Figure 6: NTK distance $1 - \cos(K_m, K_\infty)$ versus network width m for four attention variants (mean $\pm$ std over 10 seeds). *This metric is scale-invariant*, measuring angular alignment (eigenvector structure) between K_m and K_∞ while remaining blind to eigenvalue magnitudes. Compare with Figure 1, which uses the magnitude-sensitive function-space distance $\|f_m - f_{\mathrm{NTK}}\|$ and shows non-monotonic or increasing divergence for linearized attention. The two are consistent, since at large widths the eigenvectors of K_m align with those of K_∞ (low cosine distance) while the eigenvalues remain cubically amplified by $\kappa_d(G)^3$ (high function-space distance, cf. Theorem 4.4). **2L-ReLU and linearized attention** decrease gradually with width, indicating genuine structural alignment, and the widths extend to $m = 4096$, covering and exceeding the hidden dimensions of standard transformers (BERT: 768, ViT-L: 1024, GPT-2-XL: 1600). **Softmax and exponential attention** show near-zero distance at all widths, but this is vacuous, since their feature maps collapse to a single direction (Figure 7, mean pairwise cosim $= 1.000$), so K_∞ itself is rank-1. Both K_m and K_∞ are identical degenerate matrices at every width, making cosine similarity trivially unity regardless of convergence. The oscillations of the dashed lines near the numerical floor are floating-point noise in this degenerate regime.

Table 5: Values behind Figure 8. Function-space NTK distance $\|f_m - f_{\mathrm{NTK}}\|$ and the effective condition number κ_d of each architecture's attended-feature Gram, on MNIST under mean-squared-error regression ($n = 1000$), across width m. The input conditioning is $\kappa_d(\mathbf{G}) = 2.7 \times 10^9$. The two-layer ReLU distance decreases with width while both attention variants stay flat, and both attention variants hold κ_d near 10^{12}, about $330\times$ the input, at every width.

	$\|f_m - f_{\mathrm{NTK}}\|$			κ_d of attended-feature Gram		
m	2L-ReLU	MLP-Attn	Softmax-QKV	2L-ReLU	MLP-Attn	Softmax-QKV
256	0.27	0.64	0.46	2.7×10^9	9.8×10^{11}	8.7×10^{11}
512	0.24	0.63	0.45	2.7×10^9	9.8×10^{11}	9.6×10^{11}
1024	0.22	0.61	0.42	2.7×10^9	9.8×10^{11}	6.1×10^{11}
2048	0.20	0.60	0.45	2.7×10^9	9.8×10^{11}	9.7×10^{11}

B.16 The Entrywise-Exponential Softmax Kernel

The matrix-exponential surrogate of Proposition B.1 is a coarse spectral proxy, since its expansion $\exp(\mathbf{G}/\sqrt{d}) = \mathbf{I} + \mathbf{G}/\sqrt{d} + \cdots$ begins at the identity, so it never collapses and instead amplifies the spectrum. The faithful feature Gram of the softmax numerator is the *entrywise* exponential of the score matrix,

$$[\mathbf{K}_{\exp}]_{ij} := \exp(G_{ij}/\sqrt{d}), \qquad \mathbf{K}_{\exp} = \mathbf{J} + \frac{\mathbf{G}}{\sqrt{d}} + \frac{\mathbf{G}^{\circ 2}}{2d} + \cdots, \tag{18}$$

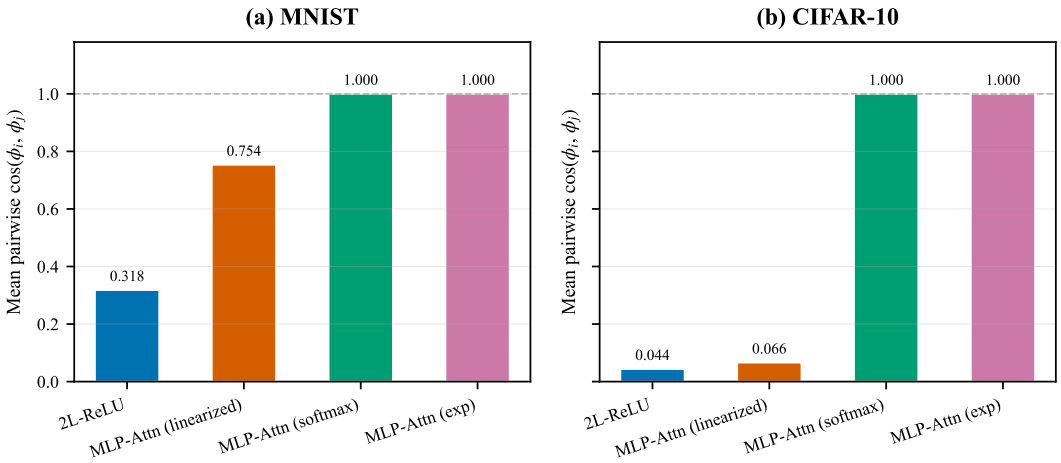


Figure 7: Feature map diversity for each attention variant, measured as mean off-diagonal pairwise cosine similarity of $\phi(\mathbf{x}_i)$ embeddings. Values near 1 indicate collapse to a single direction (degenerate). Softmax and exponential attention achieve $\cos = 1.000$ on both datasets, confirming that their near-zero NTK distances in Figure 6 are artifacts of rank-1 kernel collapse, not genuine convergence. Linearized attention and 2L-ReLU maintain diverse embeddings, validating the non-convergence result of Theorem 4.4.

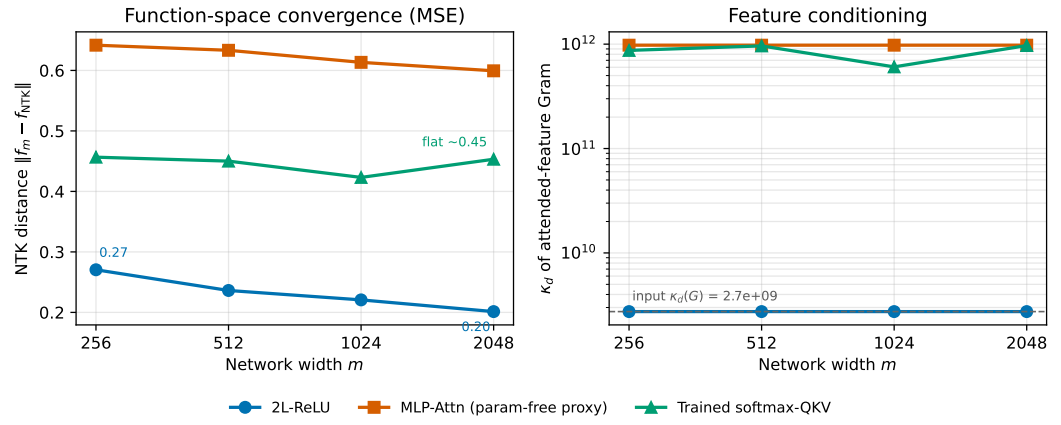


Figure 8: Trained softmax attention with learnable query, key, and value projections, on MNIST under mean-squared-error regression, against a two-layer ReLU model and the parameter-free proxy. Left: function-space NTK distance $\|f_m - f_{\mathrm{NTK}}\|$ versus width, where the two-layer ReLU distance decreases while the trained softmax head stays flat. Right: effective condition number κ_d of each architecture's attended-feature Gram, where the trained softmax head and the parameter-free proxy both amplify the input conditioning by about $330\times$ while the two-layer ReLU leaves it unchanged. The trained head's distance is measured against the arc-cosine kernel applied to its own learned attention output, since no closed-form infinite-width kernel exists once the projections are trained.

whose leading term is the rank-one all-ones matrix $\mathbf{J} = \mathbf{1}\mathbf{1}^\top$ ($\mathbf{G}^{\circ k}$ is the k-th Hadamard power). By the exact positive-random-feature factorization of the exponential kernel, $\mathbf{K}_{\exp}$ is the infinite-feature FAVOR+ Gram (Choromanski et al., 2021) of the rows of $\mathbf{X}$, so it is a genuine feature-space Gram with $\mathbf{K}_{\exp} \succeq 0$. Unit-norm rows and temperature $\tau = \sqrt{d}$ are assumed throughout. Three properties with explicit constants (full proofs in the supplementary material) make the empirically measured collapse a theorem and connect it to the cubic conditioning law.

Proposition B.4 (Rank-one collapse at rate $d^{-1/2}$)**.** $\|\mathbf{K}_{\exp} - \mathbf{J}\|_2 \le (e-1)\|\mathbf{G}\|_2/\sqrt{d}$, *so for* $d \ge 3$ *the spectral gap obeys* $\lambda_2(\mathbf{K}_{\exp})/\lambda_1(\mathbf{K}_{\exp}) \le (e-1)/(\sqrt{d}-(e-1)) = O(d^{-1/2})$ *and*

the top eigenvector aligns with $\mathbf{1}/\sqrt{n}$ *up to angle* $\sin\theta \le (e-1)\|\mathbf{G}\|_2/(n\sqrt{d}-(e-1)\|\mathbf{G}\|_2)$. *After row normalization the collapse is* $\mathbf{G}$*-independent,* $\|\mathbf{D}^{-1}\mathbf{K}_{\exp}-\mathbf{J}/n\|_2 \le e^{2/\sqrt{d}}-1$ *with* $\mathbf{D}=\mathrm{diag}(\mathbf{K}_{\exp}\mathbf{1})$. *This is the analytic form of the measured* $\cos = 1.000$ *(Figure 7).*

Proposition B.5 (The centered component inherits the precondition floor $\kappa_d(\mathbf{G})$)**.** *With* $\mathbf{P}=\mathbf{I}-\mathbf{J}/n$ *the centering projector,* $\mathbf{P}\mathbf{K}_{\exp}\mathbf{P}=\mathbf{P}\mathbf{G}\mathbf{P}/\sqrt{d}+\mathbf{R}$ *with* $\|\mathbf{R}\|_2 \le (e-2)\|\mathbf{G}\|_2/d$. *Cauchy interlacing gives* $\kappa^+(\mathbf{P}\mathbf{G}\mathbf{P}) \ge \lambda_2(\mathbf{G})/\lambda_{r-1}(\mathbf{G})$, *which matches* $\kappa_d(\mathbf{G})$ *up to the two extreme eigenvalues and, in the rank-deficient regime* $n > d$, *can exceed it. Once* $\sqrt{d} \ge 2(e-2)\|\mathbf{G}\|_2/\lambda^+_{\min}(\mathbf{P}\mathbf{G}\mathbf{P})$, *this transfers to* $\mathbf{K}_{\exp}$ *up to a factor* 3. *The informative part of the softmax kernel therefore sits above a precondition floor of order* $\kappa_d(\mathbf{G})$ *that Theorem 4.9 establishes for trainable attention, while the degenerate mean direction dominates it by a factor at least* $(n/\|\mathbf{G}\|_2)\sqrt{d}/3$.

Proposition B.6 (Cubic conditioning in the centered value pathway)**.** *Let* $\mathbf{A}=\mathbf{D}^{-1}\mathbf{K}_{\exp}$ *be the row-softmax attention matrix and* $\mathbf{A}\mathbf{G}\mathbf{A}^\top$ *the value-pathway kernel (the softmax analogue of* $\mathbf{A}\mathbf{G}\mathbf{A}^\top=\mathbf{G}^3$ *at* $\mathbf{A}=\mathbf{G}$*). For* $d \ge 4$,

$$\left\|\mathbf{P}\mathbf{A}\mathbf{G}\mathbf{A}^\top\mathbf{P}-(\mathbf{P}\mathbf{G}\mathbf{P})^3/(n^2d)\right\|_2 \le \frac{2c_1\|\mathbf{G}\|_2^2}{n\,d^{3/2}}+\frac{c_1^2\|\mathbf{G}\|_2}{d^2}, \qquad c_1 \le 8.04, \tag{19}$$

so on the part of its spectrum above this error scale the softmax value-pathway kernel obeys the cubic conditioning law $\kappa \sim \kappa^+(\mathbf{P}\mathbf{G}\mathbf{P})^3$ *of Theorem 4.4, with the centered Gram* $\mathbf{P}\mathbf{G}\mathbf{P}$ *as its base and amplitude suppressed by* n^2d.

These three properties place softmax within the same framework as the paper's other results. At temperature $\sqrt{d}$ softmax attention collapses to a rank-one kernel at leading order (Proposition B.4), so it too stays outside a well-conditioned kernel regime, while its non-degenerate component inherits the input conditioning $\kappa_d(\mathbf{G})$ (Proposition B.5) and reproduces the cubic law through the value pathway (Proposition B.6). These are initialization-time feature-kernel statements at identity projections. A trained-softmax NTK is left for later work. The conditioning of the softmax kernel is governed by the temperature through the log-sum-exp partition, since softmax is the gradient of LSE and the temperature sets its smoothing scale $\mathrm{LSE}_\tau(\mathbf{z})=\tau\,\mathrm{LSE}(\mathbf{z}/\tau)$. As $\tau \to 0$ the partition sharpens to the hard maximum, so $\mathbf{A}\to\mathbf{I}$ and the kernel recovers $\mathbf{G}$ with $\kappa_d(\mathbf{G})$, while at $\tau=\sqrt{d}$ it over-smooths toward the uniform average and produces the rank-one collapse above. No temperature choice therefore conditions the kernel below the $\kappa_d(\mathbf{G})$ floor on generic data, and a tuned interior $\tau^\star$ can even perfectly condition the kernel on adversarially chosen data (for anti-correlated $\mathbf{x}_1^\top\mathbf{x}_2=-c$ the residual eigenvalue is $\tanh^2((1+c)/2\tau)$, equalized with the collapse direction at a single $\tau^\star$), a measure-zero exception that does not restore convergence generically. The rank-one collapse is a loss of injectivity of the softmax kernel, the log-sum-exp functions of the scores becoming linearly dependent as they concentrate at $\tau=\sqrt{d}$.

B.17 Transfer to Real Architectures: Precondition versus Mechanism

Theorem 4.4 is proved for the parameter-free, single-layer linearized attention map at initialization. This subsection determines how far the cubic mechanism itself extends once LayerNorm, depth, and trainable query/key/value (QKV) projections are introduced, separating two claims that are easily conflated. The *precondition* for the obstruction, $\kappa_d(\mathbf{G}) \gg 1$, is measured directly and transfers broadly (Appendix B.13). The *mechanism*, $\tilde{\mathbf{G}}=\mathbf{G}^3$, is proved only for the parameter-free construction and transfers to real components only under explicit, checkable conditions.

B.17.1 A Conditioning Floor Under LayerNorm

Proposition B.7 (LayerNorm Is a Diagonal Congruence of the Centered Gram, with a Row-Rescaling-Invariant Floor)**.** *Let* $P \coloneqq \mathbf{I}_d-\frac{1}{d}\mathbf{1}_d\mathbf{1}_d^T$ *denote the feature-centering projector, and consider LayerNorm with unit gain, zero bias, and no numerical-stability offset, so that* $\mathrm{LN}(\mathbf{x})=\sqrt{d}\,P\mathbf{x}/\|P\mathbf{x}\|_2$ *whenever* $P\mathbf{x}\neq\mathbf{0}$. *Let* $\mathbf{G}_P \coloneqq \mathbf{X}P\mathbf{X}^T$ *denote the Gram matrix of the feature-centered rows (well-defined whenever* $P\mathbf{x}_i\neq\mathbf{0}$ *for all* i*), and let* $\mathbf{D}\coloneqq\mathrm{diag}(\mathbf{G}_P)$. *Then the Gram matrix of the row-wise LayerNormed features is*

$$\mathbf{G}_{\mathrm{LN}}=d\,\mathbf{D}^{-1/2}\,\mathbf{G}_P\,\mathbf{D}^{-1/2}, \tag{20}$$

a positive diagonal congruence of $\mathbf{G}_P$. Define the row-rescaling-invariant floor

$$\kappa_{\mathrm{opt}}(\mathbf{G}_P) \;:=\; \inf_{\substack{\mathbf{D}'\succ 0\\ \mathbf{D}' \text{ diagonal}}} \kappa_d(\mathbf{D}'\mathbf{G}_P\mathbf{D}'). \tag{21}$$

Then: (i) Floor. *$\kappa_d(\mathbf{G}_{\mathrm{LN}}) \geq \kappa_{\mathrm{opt}}(\mathbf{G}_P)$: LayerNorm, or any per-token rescaling, can never push the effective condition number below $\kappa_{\mathrm{opt}}(\mathbf{G}_P)$, a geometric invariant of the centered token directions. (ii)* Near-optimality for *$n \leq d-1$. If $\mathbf{G}_P$ is positive definite, the unit-diagonal scaling realized by LayerNorm is within a factor of n of optimal (van der Sluis, 1969): $\kappa_d(\mathbf{G}_{\mathrm{LN}}) \leq n\,\kappa_{\mathrm{opt}}(\mathbf{G}_P)$. This factor-$n$ bound is not known to extend to the rank-deficient regime $n > d$ studied throughout this paper, and closing that gap is left open. (iii)* Composition. *Since κ_d is scale-invariant, Theorem 4.4 applies verbatim to the LayerNormed features, giving $\kappa_d(\mathbf{G}_{\mathrm{LN}}^3) = \kappa_d(\mathbf{G}_{\mathrm{LN}})^3$: LayerNorm changes only the base of the cube, never the exponent.*

Proof. **Congruence.** For each row, $P\mathbf{x}_i = \mathbf{x}_i - \frac{1}{d}(\mathbf{x}_i^T\mathbf{1}_d)\mathbf{1}_d$ and $\mathrm{LN}(\mathbf{x}_i) = \sqrt{d}\,P\mathbf{x}_i/\|P\mathbf{x}_i\|_2$. Using $P^2 = P$, $[\mathbf{G}_P]_{ii} = \mathbf{x}_i^T P\mathbf{x}_i = \mathbf{x}_i^T P^T P\mathbf{x}_i = \|P\mathbf{x}_i\|_2^2$, so row i of the LayerNormed feature matrix $\hat{\mathbf{X}}$ is $\sqrt{d}\,[\mathbf{D}]_{ii}^{-1/2}\,(P\mathbf{x}_i)^T$, i.e. $\hat{\mathbf{X}} = \sqrt{d}\,\mathbf{D}^{-1/2}\mathbf{X}P$. Hence $\mathbf{G}_{\mathrm{LN}} = \hat{\mathbf{X}}\hat{\mathbf{X}}^T = d\,\mathbf{D}^{-1/2}\mathbf{X}PP^T\mathbf{X}^T\mathbf{D}^{-1/2} = d\,\mathbf{D}^{-1/2}\mathbf{G}_P\mathbf{D}^{-1/2}$ (using $P^T = P$, $P^2 = P$), a positive diagonal congruence of $\mathbf{G}_P$.

(i). $\mathbf{D}^{-1/2}$ is one member of the diagonal family over which κ_{opt} is minimized, and κ_d is invariant to the overall scalar factor d (a ratio of eigenvalues), so $\kappa_d(\mathbf{G}_{\mathrm{LN}}) = \kappa_d(\mathbf{D}^{-1/2}\mathbf{G}_P\mathbf{D}^{-1/2}) \geq \inf_{\mathbf{D}'\succ 0}\kappa_d(\mathbf{D}'\mathbf{G}_P\mathbf{D}') = \kappa_{\mathrm{opt}}(\mathbf{G}_P)$.

(ii). For positive definite $\mathbf{G}_P$ (equivalently $n \leq d-1$, since $\mathrm{rank}(\mathbf{G}_P) \leq \min(n, d-1)$), the classical equilibration result of van der Sluis (1969) states that scaling a symmetric positive definite matrix to unit diagonal achieves a condition number within a factor of its order of the condition number under the optimal diagonal scaling. $\mathbf{D}^{-1/2}\mathbf{G}_P\mathbf{D}^{-1/2}$ has unit diagonal by construction and is exactly $\mathbf{G}_{\mathrm{LN}}/d$, which has the same κ_d as $\mathbf{G}_{\mathrm{LN}}$; the bound follows with order n, the size of $\mathbf{G}_P$.

(iii). Immediate from Theorem 4.4 applied with $\mathbf{G} = \mathbf{G}_{\mathrm{LN}}$, since the effective condition number depends only on the ratio λ_1/λ_r of the transformed Gram matrix and is unaffected by which positive semidefinite matrix supplies it. □

Remark B.8 (What the Floor Does Not Say). LayerNorm can genuinely *reduce* κ_d when ill-conditioning is driven purely by row-norm heterogeneity: for rows $\{\mathbf{e}_1, \varepsilon\mathbf{e}_2\}$, $\kappa_d = \varepsilon^{-2}$ before normalization and $\kappa_d = 1$ after, so no monotonicity holds in either direction in general. LayerNorm whitens *rows*, not the *spectrum*: a Gram matrix with unit diagonal can be arbitrarily ill-conditioned when centered token directions nearly coincide, and Proposition B.7(i) shows this directional ill-conditioning is untouchable by LayerNorm or any per-token rescaling. Whether LayerNorm rescues convergence in a given deployment is therefore conditional on the measured post-LayerNorm conditioning, not a structural guarantee: it holds only if $\kappa_d(\mathbf{G}_{\mathrm{LN}})$, lower-bounded by $\kappa_{\mathrm{opt}}(\mathbf{G}_P)$, is itself moderate. This condition fails for every representation measured in this paper: the row-normalized embeddings of Appendix B.13 already reflect a per-token rescaling and still span $\kappa_d \in [1.4 \times 10^2, 2.8 \times 10^{11}]$, and the contextual DistilBERT representation, which passes through the model's own internal LayerNorms, still measures $\kappa_d \approx 1.0 \times 10^7$ (Appendix B.17.3). In every case examined, the ill-conditioning is directional, above the floor LayerNorm could remove.

B.17.2 Transfer to Deployed Architectures

The mechanism's reach into deployed models turns on four architectural features, softmax nonlinearity, LayerNorm, depth, and trainable QKV projections. Each is addressed in turn, reusing the scoped statements established elsewhere in this paper.

Softmax. Row-wise softmax is not a spectral function of $\mathbf{G}$, since the attention matrix depends on $\mathbf{G}$ entrywise, not through its eigenbasis, so the mechanism of Theorem 4.4 does not apply to it directly. Appendix B.14 already establishes, empirically, that row-normalized softmax collapses to a rank-1 kernel (mean pairwise cosine similarity 1.000 on both MNIST and CIFAR-10, Figure 7), consistent with a mean-collapse mechanism in which, at temperature $T = \sqrt{d}$, every score falls within $O(1/\sqrt{d})$ of the row mean for high-dimensional inputs, so the attention weights flatten to

within a factor $e^{2/\sqrt{d}}$ of uniform and the aggregated feature approaches the training-set mean for every row, driving the kernel toward rank one. This is a failure mode distinct from cubic amplification, a collapse floor that grows with d and the norm of the data mean, not a κ_d^p amplification law, and it exhibits no amplification at the low-temperature endpoint ($T \to 0$ recovers $\mathbf{A} \to \mathbf{I}$ and no distortion of $\mathbf{G}$). The cubic mechanism accordingly does not transfer to softmax attention at leading order, where the rank-one collapse dominates. It reappears only in the centered value pathway, once the degenerate mean direction is projected out (Proposition B.6, Appendix B.16), consistent with the scoping of Remark B.3.

LayerNorm. By Proposition B.7 and Remark B.8, the mechanism composes through LayerNorm without a change in exponent, but LayerNorm can never push conditioning below the geometric floor $\kappa_{\text{opt}}(\mathbf{G}_P)$, and every representation measured in this paper sits above it.

Stacking. Under standard layer-wise self-composition, where each layer applies the attention map to its own current features ($\mathbf{X}_{l+1} = \mathbf{X}_l \mathbf{X}_l^T \mathbf{X}_l$), the Gram recursion is $\mathbf{G}_{l+1} = \mathbf{G}_l^3$, giving $\mathbf{G}_k = \mathbf{G}^{3^k}$ and $\kappa_d(\mathbf{G}_k) = \kappa_d(\mathbf{G})^{3^k}$, so conditioning compounds doubly exponentially in depth and does not saturate. Remark 4.7 already distinguishes this convention from the frozen-reference-map recursion ($\mathbf{X}_{l+1} = \mathbf{X}_0 \mathbf{X}_0^T \mathbf{X}_l$), which instead gives the milder $\mathbf{G}^{2k+1}$ (singly exponential in k) and coincides with the former only at $k = 1$ (e.g., κ^9 versus κ^5 at $k = 2$), diverging for $k \geq 2$. That distinction is reused here unchanged. Interleaved LayerNorm breaks the closed spectral form, since the diagonal congruence of Proposition B.7 does not commute with the spectral cube. What survives is a per-layer sandwich of the same Ostrowski congruence type used in Remark C.6, with a measurable per-layer diagonal spread. Whether real trained stacks compound toward the full bound or self-regulate through this mechanism is open. Separately, truncated attention (top-r singular components) caps the condition number at λ_1/λ_r, suggesting low-rank regularization as a mechanism for restoring convergence. Theorem 4.4 bounds NTK deviation *at initialization* only and does not directly predict the post-training trajectory, and the cubic identity $\tilde{\mathbf{G}} = \mathbf{G}^3$ is construction-specific, requiring query, key, and value to share the same $\mathbf{X}$ (the SVD collapse $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ in $\mathbf{X}\mathbf{X}^T\mathbf{X} = \mathbf{U}\mathbf{\Sigma}^3\mathbf{V}^T$). Independent projections (Proposition C.4) generalize via the same argument on the projected SVD.

Trained QKV. Section 4.1 replaces the qualitative sandwich of Remark C.6 with an exact treatment of the genuinely trainable NTK. Two parts of that treatment hold under both initializations analyzed: the precondition floor $\kappa^+(\mathbf{K}_\infty^{\text{QKV}}) \geq \kappa_d(\mathbf{G})$ (Theorem 4.9), and the exponent-two persistence result $H^\star = \Theta(\kappa_d(\mathbf{G})^2/\varepsilon^2)$ (Theorem 4.10), which shows non-convergence survives trainability even when the cube does not. The cubic mechanism itself remains conditional: it requires the checkpoint to satisfy the directional alignment ratio $R > \kappa_d(\mathbf{G})^2$ (Proposition 4.11), an optimizer-agnostic diagnostic computable from $(\mathbf{W}_Q, \mathbf{W}_K, \mathbf{w}_V)$ and $\mathbf{G}$. Whether SGD, AdamW, Muon, or Shampoo drive trained projections toward or away from this region is read off directly by this cheap per-checkpoint diagnostic.

Table 6 summarizes the calibrated status of each.

Table 6: Calibrated status of the four architectural features and the precondition and consequence measurements. A result is marked proved when it holds unconditionally, conditional when it holds given an explicit checkable gate, empirical when it is a direct measurement, and open when it is not established here.

Aspect	Statement	Condition	Status
Softmax	Rank-one collapse at leading order, cube in the centered value pathway (Rmk. B.3, Prop. B.6)	—	Proved (collapse & centered cube)
LayerNorm	$\kappa_d(\mathbf{G}_{\mathrm{LN}}) \geq \kappa_{\mathrm{opt}}(\mathbf{G}_P)$, unchanged exponent (Prop. B.7)	$\kappa_d(\mathbf{G}_{\mathrm{LN}}) \gg 1$ measured	Proved floor, conditional rescue
Stacking	$\mathbf{G}_k = \mathbf{G}^{3^k}$, doubly exponential (Rmk. 4.7)	—	Proved (bare), interleaved norm. open
Trained QKV	Floor $\kappa^+(\mathbf{K}_\infty^{\mathrm{QKV}}) \geq \kappa_d(\mathbf{G})$, exponent-2 persistence (Thms. 4.9, 4.10); cube conditional on $R > \kappa_d(\mathbf{G})^2$ (Prop. 4.11)	$R = a_1/a_r$ on checkpoint	Proved floor & exponent-2, conditional cube
Precondition	$\kappa_d(\mathbf{G}) \gg 1$ across 20+ static and 1 contextual embedding (App. B.13, B.17.3)	—	Empirical
Consequence	Flip rate tracks κ_d on real embeddings (App. B.17.3)	—	Empirical
Training dynamics	The mechanism along the full optimization trajectory	—	Open

The precondition and the mechanism transfer differently. The precondition is the ill-conditioning of the input Gram, measured directly across the more than twenty deployed models of Appendix B.13, including the contextual DistilBERT representation of Appendix B.17.3 at $\kappa_d \approx 10^7$. For trainable attention this conditioning cannot be reduced below $\kappa_d(\mathbf{G})$ (Theorem 4.9), and that floor alone forces non-convergence at exponent two (Theorem 4.10). The cubic mechanism is more delicate, transferring only under conditions that can be checked on a trained checkpoint, namely that the kernel is not collapsed by a softmax, that the post-normalization conditioning remains large, and that the directional alignment ratio exceeds $\kappa_d(\mathbf{G})^2$. A whitening configuration instead yields $R = \kappa_d(\mathbf{G})$ and escapes the obstruction. None of these conditions is currently proven to hold for a trained transformer.

B.17.3 Malleability on Real Contextual Embeddings

The transferability study of Appendix B.13 measures the precondition on static embedding tables. This subsection reports a direct run of the non-convergence and malleability pipeline on a *contextual* representation: frozen DistilBERT embeddings (mean-pooled last hidden state, ℓ_2-normalized rows, $d = 768$) of AG-News text, 1,600 training and 400 test examples across four classes, using the same parameter-free linearized-attention proxy and hyperparameters as the main experiments (Adam $\eta = 10^{-3}$, $\lambda = 10^{-3}$, initialization scale $\kappa = 0.01$, PGD $\varepsilon = 0.03$ in L_∞).

The contextual embedding Gram matrix measures $\kappa_d(\mathbf{G}) \approx 1.0 \times 10^7$ ($n = 1{,}600$, $d = 768$), matching the value already reported for the same representation in Table 4 (‡ row): the precondition holds for a deep, trained, softmax, LayerNormed, residual representation, and for static token tables alike.

Influence malleability under PGD was measured at two conditioning levels: the raw embeddings ($\kappa_d = 1.0 \times 10^7$) and the same embeddings whitened to a moderate $\kappa_d = 183$, a matched negative control on which a well-behaved network is expected. Table 7 reports the influence flip rate at the top-10% influence threshold.

Table 7: Influence flip rate (top-10%, PGD $\varepsilon = 0.03$) on frozen DistilBERT AG-News embeddings, at two levels of input conditioning. The whitened row is a matched negative control at moderate κ_d; the raw row is the unmodified contextual embedding.

Input	κ_d **(input)**	**2L-ReLU flip rate**	**MLP-Attn flip rate**
Whitened (moderate)	183	0.0%	0.6%
Raw DistilBERT (high)	1.0×10^7	41.9%	61.3%

At moderate conditioning the influence vector is stable for both architectures, while at the high conditioning of the raw contextual representation it is highly malleable for both, and more so for attention. Influence malleability therefore tracks the effective condition number as Theorem 4.4 predicts, and real contextual embeddings sit in the malleable regime, the condition-number-gated consequence the mechanism implies, on the representation where influence methods such as TRAK, DataInf, and EK-FAC would actually be deployed. The architecture gap narrows to $1.46\times$ at high conditioning (from the 6–9$\times$ multiclass gap on raw pixels, Table 1): when the input itself is already severely ill-conditioned, the incremental effect of attention is secondary to the precondition.

Two limits apply to this experiment. First, the finite-width NTK-distance contrast used in Figure 1 does not cleanly reproduce on these embedding-derived inputs, since even at the matched moderate conditioning the 2L-ReLU distance does not converge monotonically as it does on raw pixels, so the clean NTK-distance contrast remains the raw-pixel result of Figure 1. Second, the experiment remains at the level of the parameter-free proxy analyzed via the NTK at initialization, leaving a direct study on a trained softmax transformer open.

C Theoretical Proofs and Derivations

C.1 Linearized Attention Approximation

Theorem C.1 (Taylor Approximation of Softmax Attention)**.** *The linearized attention mechanism $f^{att}(\mathbf{X}) = \mathbf{X}\mathbf{X}^T\mathbf{X}$ provides a first-order Taylor approximation to standard softmax attention with bounded approximation error.*

Proof. Consider standard softmax attention (Vaswani et al., 2017): $\text{Attention}(\mathbf{X}) = \text{softmax}\left(\frac{\mathbf{X}\mathbf{X}^T}{\sqrt{d}}\right)\mathbf{X}$

For row i, the softmax output is $[\text{softmax}(\mathbf{A})]_{ij} = \exp(A_{ij})/\sum_k \exp(A_{ik})$ where $\mathbf{A} = \mathbf{X}\mathbf{X}^T/\sqrt{d}$. For small entries $|A_{ij}| \ll 1$, using $\exp(x) \approx 1 + x + O(x^2)$:

$$\begin{aligned}[\text{softmax}(\mathbf{A})]_{ij} &= \frac{1 + A_{ij} + O(A_{ij}^2)}{n + \sum_k A_{ik} + O(\|\mathbf{A}\|^2)} \\ &= \frac{1}{n}\left(1 + A_{ij} - \bar{A}_i + O(\|\mathbf{A}\|^2)\right) \end{aligned} \tag{22}$$

where $\bar{A}_i = \frac{1}{n}\sum_k A_{ik}$ is the row mean. The attention output for row i is:

$$[\text{Attention}(\mathbf{X})]_i = \frac{1}{n}\sum_j (1 + A_{ij} - \bar{A}_i)\mathbf{x}_j + O(\|\mathbf{A}\|^2) \tag{23}$$

The uniform term $\frac{1}{n}\sum_j \mathbf{x}_j$ contributes only the data mean. The data-dependent term is $\frac{1}{n}\sum_j A_{ij}\mathbf{x}_j = \frac{1}{n\sqrt{d}}[\mathbf{X}\mathbf{X}^T\mathbf{X}]_i$. Absorbing constants and dropping the mean-centering term (which does not affect relative similarities), the linearized attention $f^{\text{att}}(\mathbf{X}) = \mathbf{X}\mathbf{X}^T\mathbf{X}$ captures the essential data-dependent structure with approximation error $O(\|\mathbf{X}\mathbf{X}^T\|^2/d)$. □

C.2 Exact Gram-Induced Kernel Characterization

Theorem C.2 (Parameter-Free Gram-Induced Structure)**.** *The parameter-free linearized attention kernel admits the following expansion:*

$$K_{\textit{LinAttn}}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k,\ell=1}^{n} (\mathbf{x}_i^T\mathbf{x}_k)(\mathbf{x}_k^T\mathbf{x}_\ell)(\mathbf{x}_\ell^T\mathbf{x}_j)$$

This establishes $K_{\textit{LinAttn}}$ as a parameter-free, data-dependent Gram-induced kernel with transitive interaction structure through training data.

Proof. The parameter-free linearized attention transformation is: $[\mathbf{X}\mathbf{X}^T\mathbf{X}]_i = \sum_{k=1}^n (\mathbf{x}_i^T\mathbf{x}_k)\mathbf{x}_k$

The kernel is the inner product of transformed features:

$$\begin{aligned} K_{\text{LinAttn}}(\mathbf{x}_i, \mathbf{x}_j) &= \langle [\mathbf{X}\mathbf{X}^T\mathbf{X}]_i, [\mathbf{X}\mathbf{X}^T\mathbf{X}]_j \rangle \\ &= \left\langle \sum_{k=1}^n (\mathbf{x}_i^T\mathbf{x}_k)\mathbf{x}_k, \sum_{\ell=1}^n (\mathbf{x}_j^T\mathbf{x}_\ell)\mathbf{x}_\ell \right\rangle \\ &= \sum_{k,\ell=1}^n (\mathbf{x}_i^T\mathbf{x}_k)(\mathbf{x}_j^T\mathbf{x}_\ell)(\mathbf{x}_k^T\mathbf{x}_\ell) \\ &= \sum_{k,\ell=1}^n (\mathbf{x}_i^T\mathbf{x}_k)(\mathbf{x}_k^T\mathbf{x}_\ell)(\mathbf{x}_\ell^T\mathbf{x}_j) \end{aligned} \tag{24}$$

Each term $(\mathbf{x}_i^T\mathbf{x}_k)(\mathbf{x}_k^T\mathbf{x}_\ell)(\mathbf{x}_\ell^T\mathbf{x}_j)$ represents a fourth-order interaction involving four vectors $(\mathbf{x}_i, \mathbf{x}_k, \mathbf{x}_\ell, \mathbf{x}_j)$ through three pairwise inner products. The summation over training data indices (k, ℓ) creates a data-dependent fourth-order polynomial structure without trainable parameters. □

Corollary C.3 (Parameter-Free Fourth-Order Properties)**.** *The kernel K_{LinAttn} defined by $\sum_{k,\ell}(\mathbf{x}_i^T\mathbf{x}_k)(\mathbf{x}_k^T\mathbf{x}_\ell)(\mathbf{x}_\ell^T\mathbf{x}_j)$ exhibits fourth-order polynomial interactions (each term involves four vectors $\mathbf{x}_i, \mathbf{x}_k, \mathbf{x}_\ell, \mathbf{x}_j$ at degree 4), requires zero learnable parameters (since $f^{att}(\mathbf{X}) = \mathbf{X}\mathbf{X}^T\mathbf{X}$ contains no weight matrices), and depends on the full training set $\mathbf{X}$ through the summation indices k, ℓ. In particular, influence from $\mathbf{x}_i$ to $\mathbf{x}_j$ propagates transitively via intermediate points: $\mathbf{x}_i \to \mathbf{x}_k$ (via $\mathbf{x}_i^T\mathbf{x}_k$), $\mathbf{x}_k \to \mathbf{x}_\ell$ (via $\mathbf{x}_k^T\mathbf{x}_\ell$), $\mathbf{x}_\ell \to \mathbf{x}_j$ (via $\mathbf{x}_\ell^T\mathbf{x}_j$).*

C.3 Proof of Softmax Spectral Amplification (Proposition B.1)

Proof. Let $\mathbf{G} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ be the eigendecomposition with $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. Since $\mathbf{G}$ is symmetric, $\exp(\mathbf{G}/\sqrt{d}) = \mathbf{U}\exp(\mathbf{\Lambda}/\sqrt{d})\mathbf{U}^\top$ where $\exp(\mathbf{\Lambda}/\sqrt{d}) = \operatorname{diag}(e^{\lambda_i/\sqrt{d}})$. The transformed feature matrix is:

$$\tilde{\mathbf{X}} = \exp(\mathbf{G}/\sqrt{d})\,\mathbf{X} = \mathbf{U}\operatorname{diag}(e^{\lambda_i/\sqrt{d}})\,\mathbf{U}^\top\mathbf{X}.$$

The induced kernel matrix is $\tilde{\mathbf{G}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$. Using $\mathbf{G} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, so $\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U} = \mathbf{U}^\top\mathbf{G}\mathbf{U} = \mathbf{\Lambda}$:

$$\tilde{\mathbf{G}} = \mathbf{U}\,e^{\mathbf{\Lambda}/\sqrt{d}}\,(\mathbf{U}^\top\mathbf{X}\mathbf{X}^\top\mathbf{U})\,e^{\mathbf{\Lambda}/\sqrt{d}}\,\mathbf{U}^\top = \mathbf{U}\operatorname{diag}\Big(\lambda_i\,e^{2\lambda_i/\sqrt{d}}\Big)\,\mathbf{U}^\top.$$

The eigenvalues of $\tilde{\mathbf{G}}$ are $\mu_i = \lambda_i e^{2\lambda_i/\sqrt{d}}$. Since $\mu_i/\mu_j = (\lambda_i/\lambda_j)\exp(2(\lambda_i - \lambda_j)/\sqrt{d})$ is increasing in i (for $\lambda_i \geq \lambda_j$), the condition number is:

$$\kappa_d(\tilde{\mathbf{G}}) = \frac{\mu_1}{\mu_r} = \frac{\lambda_1}{\lambda_r}\cdot\exp\left(\frac{2(\lambda_1 - \lambda_r)}{\sqrt{d}}\right) = \kappa_d(\mathbf{G})\cdot\exp\left(\frac{2(\lambda_1 - \lambda_r)}{\sqrt{d}}\right),$$

where $\mu_r = \lambda_r e^{2\lambda_r/\sqrt{d}}$ is the smallest positive eigenvalue of $\tilde{\mathbf{G}}$ (the map $\lambda \mapsto \lambda e^{2\lambda/\sqrt{d}}$ is increasing on $\lambda > 0$, so it preserves the ordering of the positive spectrum). □

C.4 Fixed QKV Projections: Orthogonal Case and Caveats

Proposition C.4 (Fixed QKV (Query, Key, Value) Projections: Orthogonal Case)**.** *Let $\mathbf{G}_W := \mathbf{X}\mathbf{W}_K\mathbf{W}_K^\top\mathbf{X}^\top$ denote the projected Gram matrix (distinct from the NTK matrix $\mathbf{K}$). For orthogonal key projections with $\mathbf{W}_K^\top\mathbf{W}_K = \frac{d}{2}\mathbf{I}$, the effective condition number is preserved exactly, $\kappa_d(\mathbf{G}_W) = \kappa_d(\mathbf{G})$, so the cubic conditioning and the sufficient width $m = \Omega(\kappa_d(\mathbf{G})^6\log n)$ of Theorem 4.4 transfer without change. For general full-rank $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$, the induced kernel retains data-dependence, transitive similarity, and fourth-order structure (Proposition C.7 below). The exact cubic identity requires the query, key, and value factorizations to share a common basis, so distinct projections are covered structurally but not with the same constant.*

Proof. The projected Gram is $\mathbf{G}_W = (\mathbf{X}\mathbf{W}_K)(\mathbf{X}\mathbf{W}_K)^\top = \mathbf{X}\,\mathbf{W}_K\mathbf{W}_K^\top\,\mathbf{X}^\top = \frac{d}{2}\mathbf{G}$. A positive scalar multiple of $\mathbf{G}$ has identical eigenvalue ratios, so $\kappa_d(\mathbf{G}_W) = \kappa_d(\mathbf{G})$, and the cube map $\mathbf{G} \mapsto \mathbf{G}^3$ acts on the same spectrum, transferring the cubic conditioning and the sufficient width of Theorem 4.4 exactly. The general full-rank case is proved below (Proposition C.7). □

Remark C.5 (Fixed He-Projected Feature Gram, and Contrast with the Trainable NTK). The projection $\mathbf{W}_K$ here is drawn once, at He scale ($\mathbf{W}_K \sim \mathcal{N}(0, 2/d)$, $\mathbb{E}[\mathbf{G}_W] = 2\mathbf{G}$), and held fixed, so attention still supplies only a static feature map, and $\kappa_d(\mathbf{G}_W)$ depends on the particular draw. The distribution-level bound left open in earlier analysis is resolved by Proposition C.14 (Appendix C.7). Whether a fixed random draw transfers the cubic conditioning depends on a single dimension count, over-complete keys against the standard multi-head width. The He scaling itself sets only the overall magnitude of $\mathbf{G}_W$, leaving its conditioning unchanged. This fixed-projection question is distinct from the *trainable* query/key/value NTK analyzed in Section 4.1, where $\mathbf{W}_Q, \mathbf{W}_K$ are trained parameters and independent Gaussian initialization erases the cube exactly (Theorem C.9), replacing $\mathbf{G}^3$ by a scalar multiple of $\mathbf{G}$.

Remark C.6 (A Checkable Condition for Trained Projections). The extension to trainable QKV projections holds only under an explicit, checkable condition on the learned projections. Let $\mathbf{M} = \mathbf{X}^T\mathbf{X}$ and $\mathbf{C} := \mathbf{W}_Q\mathbf{W}_K^T\mathbf{M}\mathbf{W}_V/\sqrt{d}$, so the full distinct-QKV layer collapses to $f(\mathbf{X}) = \mathbf{X}\mathbf{C}$. An Ostrowski congruence argument gives the two-sided sandwich $\kappa_d(\mathbf{G}_{\text{out}}) \in \kappa_d(\mathbf{G})\cdot[\kappa(\mathbf{C}^T\mathbf{C})^{-1}, \kappa(\mathbf{C}^T\mathbf{C})]$,

generalizing the orthogonal-only statement of Proposition C.4 to any full-rank projection with a measurable degradation factor $\kappa(\mathbf{C}^T\mathbf{C})$. The upper end recovers $\kappa_d(\mathbf{G})^3(\kappa(\mathbf{W}_Q)\kappa(\mathbf{W}_K)\kappa(\mathbf{W}_V))^2$. No matching unconditional lower bound holds, however. A whitening configuration $\mathbf{W}_Q\mathbf{W}_K^T \propto \mathbf{M}^{-1}$ gives $\mathbf{C} \propto \mathbf{I}$ and erases the cube exactly, $\kappa_d(\mathbf{G}_{\text{out}}) = \kappa_d(\mathbf{G})$, so trained attention can in principle learn projections that avoid the obstruction. The mechanism for trainable QKV is thus conditional on the learned projections not converging to such a whitening configuration, a checkable condition on the trained checkpoint, measured directly as $\kappa(\mathbf{C}^T\mathbf{C})$ or equivalently $\kappa_d(\mathbf{G}_{\text{out}})$. The genuinely trainable NTK in Section 4.1 makes this exact, replacing the qualitative sandwich above with a closed-form amplification exponent and an optimizer-agnostic reachability criterion (Proposition 4.11).

C.5 Proof of QKV Equivalence

Proposition C.7 (Kernel Equivalence Under Linear Projections)**.** *For full-rank projection matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$, the generalized linearized attention $f^{att}_{\mathbf{W}}(\mathbf{X}) = (\mathbf{X}\mathbf{W}_Q)(\mathbf{X}\mathbf{W}_K)^T(\mathbf{X}\mathbf{W}_V)$ induces a kernel with the same structural properties as $K_{LinAttn}$: data-dependence, transitive similarity propagation, and fourth-order polynomial interactions.*

Proof of Proposition C.7. The generalized linearized attention with projection matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$ is:

$$f^{\text{att}}_{\mathbf{W}}(\mathbf{X}) = (\mathbf{X}\mathbf{W}_Q)(\mathbf{X}\mathbf{W}_K)^T(\mathbf{X}\mathbf{W}_V) \tag{25}$$

For row i, this computes:

$$[f^{\text{att}}_{\mathbf{W}}(\mathbf{X})]_i = \sum_{k=1}^{n}(\mathbf{x}_i^T\mathbf{W}_Q\mathbf{W}_K^T\mathbf{x}_k)(\mathbf{x}_k^T\mathbf{W}_V) \tag{26}$$

where the attention weight is $(\mathbf{x}_i^T\mathbf{W}_Q)(\mathbf{W}_K^T\mathbf{x}_k) = \mathbf{x}_i^T\mathbf{M}_{QK}\mathbf{x}_k$ with $\mathbf{M}_{QK} = \mathbf{W}_Q\mathbf{W}_K^T$.

The induced kernel is the inner product of transformed features:

$$\begin{aligned} K_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_j) &= \langle [f^{\text{att}}_{\mathbf{W}}(\mathbf{X})]_i, [f^{\text{att}}_{\mathbf{W}}(\mathbf{X})]_j \rangle \\ &= \sum_{k,\ell}(\mathbf{x}_i^T\mathbf{M}_{QK}\mathbf{x}_k)(\mathbf{x}_j^T\mathbf{M}_{QK}\mathbf{x}_\ell)(\mathbf{x}_k^T\mathbf{M}_{VV}\mathbf{x}_\ell) \end{aligned} \tag{27}$$

where $\mathbf{M}_{VV} = \mathbf{W}_V\mathbf{W}_V^T$.

Setting $\mathbf{W}_Q = \mathbf{W}_K = \mathbf{W}_V = \mathbf{I}$ gives $\mathbf{M}_{QK} = \mathbf{M}_{VV} = \mathbf{I}$, recovering exactly:

$$K_{\text{LinAttn}}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k,\ell}(\mathbf{x}_i^T\mathbf{x}_k)(\mathbf{x}_j^T\mathbf{x}_\ell)(\mathbf{x}_k^T\mathbf{x}_\ell) \tag{28}$$

For general full-rank projections, the kernel retains the same structural properties: (i) data-dependence through the (transformed) Gram matrix, (ii) transitive similarity propagation via intermediate summation indices, and (iii) fourth-order polynomial interactions. The projections merely rotate the feature space without altering these structural characteristics. □

C.6 Trainable Query, Key, and Value Projections: Exact NTK and Phase Transition

Throughout this subsection, $\mathbf{G} = \mathbf{X}\mathbf{X}^\top$ has eigendecomposition $\mathbf{G} = \sum_{i=1}^{r}\lambda_i\mathbf{u}_i\mathbf{u}_i^\top$ with $\lambda_1 \geq \cdots \geq \lambda_r > 0$, and $\mathbf{M} = \mathbf{X}^\top\mathbf{X}$ shares the nonzero spectrum $\{\lambda_i\}$. All statements concern the model of Definition 4.8. The head index h is dropped in single-head computations and restored by summing over h where indicated.

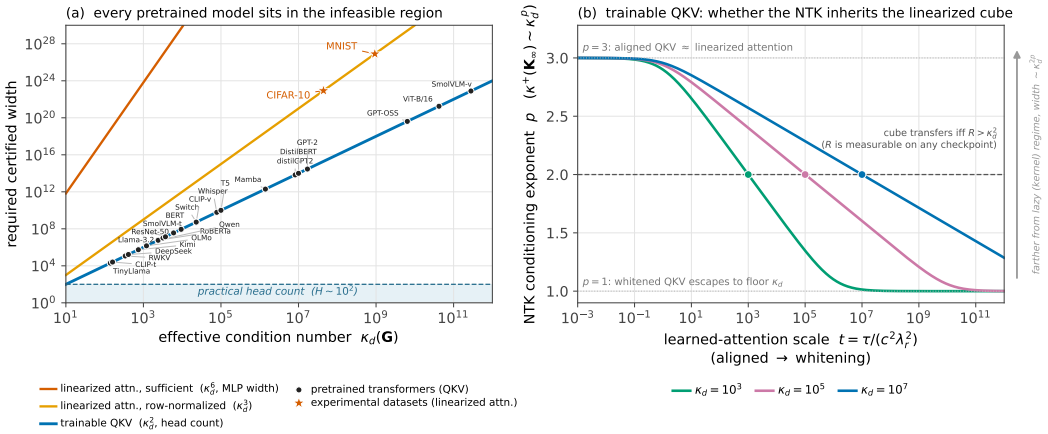


Figure 9: (a) Required certified width plotted against the effective condition number $\kappa_d(\mathbf{G})$, log-log. Parameter-free linearized attention requires MLP width $\Omega(\kappa_d^6 \log n)$ (sufficient) or κ_d^3 (row-normalized, tight); trainable QKV attention requires head count $\Theta(\kappa_d^2/\varepsilon^2)$. MNIST and CIFAR-10 (red stars, linearized-attention experiments) and more than twenty pretrained transformers (black dots, trainable QKV) are placed at their measured $\kappa_d(\mathbf{G})$, and all lie far above the practical head count (of order 10^2), so no practical width certifies the kernel regime. (b) For trainable QKV, the limiting-NTK conditioning satisfies $\kappa^+(\mathbf{K}_\infty^{\mathrm{QKV}}) \sim \kappa_d(\mathbf{G})^p$, with the exponent p set by the learned-attention scale $t = \tau/(c^2\lambda_r^2)$ (horizontal axis, from aligned to whitening). When the learned projection stays aligned ($p \to 3$), the NTK reproduces the linearized-attention cube $\kappa_d(\mathbf{G})^3$; when it whitens ($p \to 1$), the conditioning falls to the data-Gram floor $\kappa_d(\mathbf{G})$. Certifying the lazy (kernel) regime requires width of order $\kappa_d(\mathbf{G})^{2p}$, so a larger p places the regime farther out of reach. The cubic mechanism transfers only under the explicit, checkpoint-measurable condition $R = a_1/a_r > \kappa_d(\mathbf{G})^2$ (equivalently $p \geq 2$), a diagnostic computable directly from the trained parameters of any checkpoint.

Table 8: Condition-number amplification exponent by attention setting. The cube ($\kappa_d(\mathbf{G})^3$) and the $\kappa_d(\mathbf{G})^6$ width certificate are properties of the parameter-free proxy alone, while every trainable-attention threshold carries exponent two unless the checkable alignment condition of Proposition 4.11 is met.

Setting	**Kernel**	**Conditioning**	**Width / head threshold**
Parameter-free linearized attention (proxy, exact)	$\mathbf{G}^3$	$\kappa_d(\mathbf{G})^3$	$\Omega(\kappa_d(\mathbf{G})^6 \log n)$ sufficient; $\kappa_d(\mathbf{G})^3$ tight (row-normalized)
Trainable QKV, independent init	$c_{\mathrm{I}}\mathbf{G}$	$\kappa_d(\mathbf{G})$	$H^\star = \Theta(\kappa_d(\mathbf{G})^2/\varepsilon^2)$ (Cor. C.13)
Trainable QKV, aligned & checkable $R > \kappa_d(\mathbf{G})^2$	$\mathbf{G}^3 + \tau\mathbf{G}$	$\kappa_d(\mathbf{G})^p$, $p \in [2, 3)$	$H = \Theta(\kappa_d(\mathbf{G})^{2p}/\varepsilon^2)$ at aligned endpoints $p = 1, 3$ (Cor. C.11, Cor. C.12)
Trainable QKV, whitening ($R = \kappa_d(\mathbf{G})$)	$c\mathbf{G}$	$\kappa_d(\mathbf{G})$	escapes

The construction of Definition 4.8 fixes a head count $H \geq 1$, a key/query width $p \geq 1$, a score scale $s > 0$, and fixed signs $a_h \in \{-1, +1\}$. Per head h, the trainable parameters are $\mathbf{W}_Q^h, \mathbf{W}_K^h \in \mathbb{R}^{d\times p}$ and $\mathbf{w}_V^h \in \mathbb{R}^d$, and the model is

$$f(\mathbf{X};\theta) = \frac{1}{\sqrt{H}} \sum_{h=1}^{H} a_h \, \mathbf{A}_h \mathbf{X} \mathbf{w}_V^h \in \mathbb{R}^n, \qquad \mathbf{A}_h = \frac{1}{s}(\mathbf{X}\mathbf{W}_Q^h)(\mathbf{X}\mathbf{W}_K^h)^\top \in \mathbb{R}^{n\times n}. \tag{29}$$

The empirical NTK $\mathbf{K}_H(\theta) = J(\theta)J(\theta)^\top$ is taken over all trainable parameters $\{\mathbf{W}_Q^h, \mathbf{W}_K^h, \mathbf{w}_V^h\}_{h=1}^H$, with the signs a_h fixed, mirroring the fixed second-layer signs a_r of Equation 3. Write $\mathbf{K}_\infty^{\mathrm{QKV}} := \lim_{H\to\infty} \mathbf{K}_H$, which exists almost surely by the strong law of large numbers over i.i.d. heads, with H playing the role of the width m. Two initialization schemes recur below.

Under *aligned orthogonal* initialization ($p = d$, $\mathbf{W}_Q^h = \mathbf{W}_K^h = \sqrt{d/2}\,\mathbf{O}_h$ for arbitrary orthogonal $\mathbf{O}_h$, $s = d/2$, value variance $1/d$), $\mathbf{A}_h = \mathbf{G}$ exactly for every head, so the model coincides at initialization with the parameter-free proxy $f^{\text{att}}(\mathbf{X}) = \mathbf{X}\mathbf{X}^\top\mathbf{X}$. Under *independent Gaussian* initialization, $\mathbf{W}_Q^h$ and $\mathbf{W}_K^h$ are drawn independently, matching standard transformer initialization.

The forward map is trilinear in the three parameter groups $\mathbf{W}_Q^h, \mathbf{W}_K^h, \mathbf{w}_V^h$, so differentiating with respect to one group while holding the others fixed produces, for the query and key groups, a Jacobian that is rank one in the row index. The value group instead produces a full Gram-type Jacobian. Summing the resulting outer products over the three parameter groups and over heads gives an exact three-term decomposition of the NTK, with no approximation at any width.

Theorem C.8 (Exact Block Decomposition of the Trainable NTK)**.** *For any parameter values, the empirical NTK of Definition 4.8 decomposes exactly as* $\mathbf{K}_H = \mathbf{K}_V + \mathbf{K}_Q + \mathbf{K}_K$*, with*

$$\mathbf{K}_V = \frac{1}{H}\sum_{h=1}^{H} \mathbf{A}_h \mathbf{G} \mathbf{A}_h^\top, \quad \mathbf{K}_Q = \frac{1}{H}\sum_{h=1}^{H} \|\mathbf{g}_h\|_2^2\, \mathbf{G}, \quad \mathbf{K}_K = \frac{1}{H}\sum_{h=1}^{H} \frac{\|\mathbf{h}_h\|_2^2}{s^2}\, \mathbf{X}\mathbf{W}_Q^h (\mathbf{W}_Q^h)^\top \mathbf{X}^\top, \tag{30}$$

where $\mathbf{g}_h := \frac{1}{s}(\mathbf{W}_K^h)^\top \mathbf{M}\mathbf{w}_V^h$, $\mathbf{h}_h := \mathbf{M}\mathbf{w}_V^h$, *and* $\mathbf{M} = \mathbf{X}^\top\mathbf{X}$. *The query and key blocks are, for every parameter value, exact scalar multiples of* $\mathbf{G}$*; the value block equals* $c^2\mathbf{G}^3$ *whenever the score matrix is Gram-aligned,* $\mathbf{A}_h = c\mathbf{G}$*, and equals* $\mathbf{G}^3$ *exactly under aligned orthogonal initialization. All three blocks are positive semidefinite with range contained in* $\mathrm{range}(\mathbf{G})$.

Proof of Theorem C.8. Since the parameter vector θ is partitioned into disjoint per-head, per-group blocks, the Jacobian Gram $\mathbf{K}_H = JJ^\top$ splits as a sum of per-block Grams. The fixed signs satisfy $a_h^2 = 1$ and the head prefactor $1/\sqrt{H}$ contributes $1/H$ to each block, so it suffices to compute the three per-head blocks and average over h. Fix a head and drop the index. The i-th output coordinate is

$$f_i = \frac{1}{s}\sum_{k=1}^{n} \langle \mathbf{W}_Q^\top \mathbf{x}_i, \mathbf{W}_K^\top \mathbf{x}_k \rangle \langle \mathbf{x}_k, \mathbf{w}_V \rangle.$$

Value path. $\partial f_i/\partial \mathbf{w}_V = \sum_k \mathbf{A}_{ik}\mathbf{x}_k = [\mathbf{A}\mathbf{X}]_{i,:}$, so the block Gram is $(\mathbf{A}\mathbf{X})(\mathbf{A}\mathbf{X})^\top = \mathbf{A}\mathbf{X}\mathbf{X}^\top\mathbf{A}^\top = \mathbf{A}\mathbf{G}\mathbf{A}^\top$.

Query path. For entry (a, b) of $\mathbf{W}_Q$, $\partial\langle \mathbf{W}_Q^\top \mathbf{x}_i, \mathbf{W}_K^\top \mathbf{x}_k\rangle/\partial[\mathbf{W}_Q]_{ab} = [\mathbf{x}_i]_a [\mathbf{W}_K^\top \mathbf{x}_k]_b$, so

$$\frac{\partial f_i}{\partial \mathbf{W}_Q} = \frac{1}{s}\mathbf{x}_i \Big(\sum_k \langle \mathbf{x}_k, \mathbf{w}_V \rangle \mathbf{W}_K^\top \mathbf{x}_k \Big)^\top = \mathbf{x}_i \mathbf{g}^\top, \qquad \mathbf{g} = \frac{1}{s}\mathbf{W}_K^\top \mathbf{X}^\top \mathbf{X}\mathbf{w}_V = \frac{1}{s}\mathbf{W}_K^\top \mathbf{M}\mathbf{w}_V,$$

a rank-one matrix whose second factor does not depend on i. Hence $[\mathbf{K}_Q]_{ij} = \langle \mathbf{x}_i\mathbf{g}^\top, \mathbf{x}_j\mathbf{g}^\top\rangle_F = \langle \mathbf{x}_i, \mathbf{x}_j\rangle \|\mathbf{g}\|_2^2 = \|\mathbf{g}\|_2^2 [\mathbf{G}]_{ij}$: an exact scalar multiple of $\mathbf{G}$ for every parameter value, since no assumption on $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{w}_V$ beyond their dimensions was used.

Key path. Symmetrically, $\partial f_i/\partial \mathbf{W}_K = \frac{1}{s}\mathbf{h}(\mathbf{W}_Q^\top \mathbf{x}_i)^\top$ with $\mathbf{h} = \mathbf{M}\mathbf{w}_V$ independent of i, giving $[\mathbf{K}_K]_{ij} = \frac{1}{s^2}\|\mathbf{h}\|_2^2\, \mathbf{x}_i^\top \mathbf{W}_Q \mathbf{W}_Q^\top \mathbf{x}_j$, an exact scalar multiple of $\mathbf{X}\mathbf{W}_Q\mathbf{W}_Q^\top\mathbf{X}^\top$.

Positive semidefiniteness and the range inclusion are immediate: each block is the Gram matrix of a set of vectors lying in the row space of $\mathbf{X}$ (for $\mathbf{K}_V$, $\mathbf{A} = \frac{1}{s}\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{X}^\top$ has column space in $\mathrm{range}(\mathbf{X}\mathbf{X}^\top) = \mathrm{range}(\mathbf{G})$). Averaging the three per-head blocks over h gives the stated $\mathbf{K}_H$. For the cubic seed, if $\mathbf{A} = c\mathbf{G}$ then $\mathbf{A}\mathbf{G}\mathbf{A}^\top = c^2\mathbf{G}^3$ by symmetry of $\mathbf{G}$; under aligned orthogonal initialization, $\mathbf{A} = \frac{2}{d}(\mathbf{X}\mathbf{W}_Q)(\mathbf{X}\mathbf{W}_K)^\top = \frac{2}{d}\mathbf{X}(\frac{d}{2}\mathbf{I})\mathbf{X}^\top = \mathbf{G}$ exactly (using $\mathbf{W}_Q = \mathbf{W}_K = \sqrt{d/2}\,\mathbf{O}$ and $\mathbf{O}\mathbf{O}^\top = \mathbf{I}$), so $c = 1$ and the value block equals $\mathbf{G}^3$. □

Proof of Theorem 4.9. The bound follows as a special case of Theorem C.9: under independent initialization, equality $\kappa^+(\mathbf{K}_\infty^{\mathrm{QKV},(\mathrm{I})}) = \kappa_d(\mathbf{G})$ holds exactly, and under aligned orthogonal initialization $\kappa^+(\mathbf{K}_\infty^{\mathrm{QKV}}) = \kappa_d(\mathbf{G})\cdot(\lambda_1^2+\tau_0)/(\lambda_r^2+\tau_0) \geq \kappa_d(\mathbf{G})$ for every $\tau_0 \geq 0$, since $\lambda_1 \geq \lambda_r$ makes the ratio $(\lambda_1^2+\tau_0)/(\lambda_r^2+\tau_0) \geq 1$. Both cases share the same structural reason. At each analyzed initialization $\mathbf{K}_\infty^{\mathrm{QKV}}$ is a polynomial in $\mathbf{G}$ with nonnegative coefficients and no constant term, namely $c_{\mathrm{I}}\mathbf{G}$ under independent initialization and $\mathbf{G}^3 + \tau_0\mathbf{G}$ under aligned orthogonal initialization (Theorem C.8),

so it shares the eigenbasis of $\mathbf{G}$ with eigenvalues $q(\lambda_i)$ for a polynomial $q(\lambda) = \sum_{k\geq 1} c_k \lambda^k$ with $c_k \geq 0$. Since $q(\lambda)/\lambda$ is then nondecreasing, $\kappa^+ = q(\lambda_1)/q(\lambda_r) \geq \lambda_1/\lambda_r = \kappa_d(\mathbf{G})$. Range containment in $\mathrm{range}(\mathbf{G})$ alone would not floor this ratio, since in a sum of positive semidefinite blocks a well-conditioned block can dominate an ill-conditioned one, so the floor rests on the shared eigenbasis at the two initializations rather than on the range inclusion. □

Theorem C.9 (Spectral Phase Transition)**.** *Under aligned orthogonal initialization, the limiting NTK is exactly $\mathbf{K}_\infty^{\mathrm{QKV}} = \mathbf{G}^3 + \tau_0 \mathbf{G}$ with $\tau_0 = 4\|\mathbf{G}\|_F^2/d^2$, so that $\mathbf{K}_\infty^{\mathrm{QKV}}$, $\mathbf{G}^3$, and $\mathbf{G}$ share the eigenbasis $\{\mathbf{u}_i\}$ of $\mathbf{G}$, with eigenvalues $\lambda_i^3 + \tau_0\lambda_i$. Along the top direction $\mathbf{u}_1$ the cubic term dominates whenever $\lambda_1^2 > \tau_0$; along the bottom direction $\mathbf{u}_r$, which sets $\lambda_{\min}^+$, the linear term dominates once $\tau_0 > \lambda_r^2$. Consequently*

$$\kappa^+(\mathbf{K}_\infty^{\mathrm{QKV}}) = \kappa_d(\mathbf{G}) \cdot \frac{\lambda_1^2 + \tau_0}{\lambda_r^2 + \tau_0} \;\in\; [\kappa_d(\mathbf{G}),\, \kappa_d(\mathbf{G})^3], \tag{31}$$

and the cubic regime, $\kappa^+(\mathbf{K}_\infty^{\mathrm{QKV}}) \geq \frac{1}{2}\kappa_d(\mathbf{G})^3$, occurs only when $\tau_0 \leq \lambda_r^2$, equivalently $\kappa_d(\mathbf{G}) \leq d/2$. Under independent Gaussian initialization, for any $s, p, \sigma_q, \sigma_k, \sigma_v > 0$, the limiting NTK is instead exactly proportional to $\mathbf{G}$,

$$\mathbf{K}_\infty^{\mathrm{QKV,(I)}} = c_{\mathrm{I}}\mathbf{G}, \qquad c_{\mathrm{I}} = \frac{p\,\mathrm{tr}(\mathbf{M}^2)}{s^2}\big(\sigma_q^2\sigma_k^2 + \sigma_k^2\sigma_v^2 + \sigma_q^2\sigma_v^2\big), \tag{32}$$

so $\kappa^+(\mathbf{K}_\infty^{\mathrm{QKV,(I)}}) = \kappa_d(\mathbf{G})$ exactly and the cube is erased at every order.

Proof of Theorem C.9. Under aligned orthogonal initialization, Theorem C.8 gives $\mathbf{K}_V = \mathbf{G}^3$ deterministically for every head. For the query block, using $s = d/2$, $\mathbf{W}_K = \sqrt{d/2}\,\mathbf{O}$, and orthogonal invariance of the Euclidean norm,

$$\|\mathbf{g}\|_2^2 = \frac{1}{s^2}\|\mathbf{W}_K^\top \mathbf{M}\mathbf{w}_V\|_2^2 = \frac{4}{d^2}\cdot\frac{d}{2}\|\mathbf{O}^\top\mathbf{M}\mathbf{w}_V\|_2^2 = \frac{2}{d}\|\mathbf{M}\mathbf{w}_V\|_2^2.$$

For the key block, $\mathbf{X}\mathbf{W}_Q\mathbf{W}_Q^\top\mathbf{X}^\top = \frac{d}{2}\mathbf{G}$, so $\mathbf{K}_K^{(h)} = \frac{\|\mathbf{M}\mathbf{w}_V^h\|_2^2}{(d/2)^2}\cdot\frac{d}{2}\mathbf{G} = \frac{2}{d}\|\mathbf{M}\mathbf{w}_V^h\|_2^2\,\mathbf{G}$. Adding the query and key coefficients gives $\frac{4}{d}\|\mathbf{M}\mathbf{w}_V^h\|_2^2$ per head. The variables $\tau_h \coloneqq \frac{4}{d}\|\mathbf{M}\mathbf{w}_V^h\|_2^2$ are i.i.d. with mean $\frac{4}{d}\sigma_v^2\mathrm{tr}(\mathbf{M}^2) = \frac{4\,\mathrm{tr}(\mathbf{M}^2)}{d^2}$ (using $\sigma_v^2 = 1/d$), and the strong law of large numbers gives $\bar{\tau}_H \to \tau_0 \coloneqq \frac{4\,\mathrm{tr}(\mathbf{M}^2)}{d^2} = \frac{4\|\mathbf{G}\|_F^2}{d^2}$ almost surely, since $\mathrm{tr}(\mathbf{M}^2) = \sum_i \lambda_i^2 = \|\mathbf{G}\|_F^2$. Both $\mathbf{G}^3$ and $\tau_0\mathbf{G}$ are polynomials in $\mathbf{G}$, so they share its eigenbasis and their eigenvalues add: $\mu_i = \lambda_i^3 + \tau_0\lambda_i$. The ratio $\mu_1/\mu_r = \kappa_d(\mathbf{G})\cdot(\lambda_1^2+\tau_0)/(\lambda_r^2+\tau_0)$ is monotone decreasing in τ_0 (the derivative of its logarithm with respect to τ_0 is $1/(\lambda_1^2+\tau_0) - 1/(\lambda_r^2+\tau_0) < 0$ since $\lambda_1 > \lambda_r$), with limiting values $\kappa_d(\mathbf{G})^3$ at $\tau_0 = 0$ and $\kappa_d(\mathbf{G})$ as $\tau_0 \to \infty$, giving $\kappa^+(\mathbf{K}_\infty^{\mathrm{QKV}}) \in [\kappa_d(\mathbf{G}), \kappa_d(\mathbf{G})^3]$ for every $\tau_0 \geq 0$. If $\tau_0 \leq \lambda_r^2$ then $\lambda_r^2 + \tau_0 \leq 2\lambda_r^2$ and $\lambda_1^2 + \tau_0 \geq \lambda_1^2$, so $\kappa^+ \geq \kappa_d(\mathbf{G})\cdot\lambda_1^2/(2\lambda_r^2) = \frac{1}{2}\kappa_d(\mathbf{G})^3$, the cubic regime. Since $\tau_0 \geq 4\lambda_1^2/d^2$ always (Corollary C.10), this regime requires $\kappa_d(\mathbf{G}) \leq d/2$.

Under independent Gaussian initialization, write $\mathbf{K}_\infty^{\mathrm{QKV,(I)}} = \mathbb{E}[\mathbf{K}_V^{(1)}] + \mathbb{E}[\mathbf{K}_Q^{(1)}] + \mathbb{E}[\mathbf{K}_K^{(1)}]$ (heads are i.i.d., one head suffices). Let $\mathbf{q}_a, \mathbf{k}_a \in \mathbb{R}^d$ $(a = 1, \ldots, p)$ be the columns of $\mathbf{W}_Q, \mathbf{W}_K$, so $\mathbf{W}_Q\mathbf{W}_K^\top = \sum_a \mathbf{q}_a\mathbf{k}_a^\top$ with $\mathbf{q}_a \sim \mathcal{N}(0, \sigma_q^2\mathbf{I})$, $\mathbf{k}_a \sim \mathcal{N}(0, \sigma_k^2\mathbf{I})$ mutually independent. For the value block, with $\mathbf{B} \coloneqq \mathbf{M}^2$,

$$\mathbb{E}[\mathbf{A}\mathbf{G}\mathbf{A}^\top] = \frac{1}{s^2}\mathbf{X}\,\mathbb{E}\Big[\sum_{a,b}\mathbf{q}_a(\mathbf{k}_a^\top\mathbf{B}\mathbf{k}_b)\mathbf{q}_b^\top\Big]\,\mathbf{X}^\top;$$

cross terms $a \neq b$ vanish since $\mathbb{E}[\mathbf{q}_a\mathbf{q}_b^\top] = 0$ by independence, and diagonal terms give $\mathbb{E}[\mathbf{k}_a^\top\mathbf{B}\mathbf{k}_a]\mathbb{E}[\mathbf{q}_a\mathbf{q}_a^\top] = \sigma_k^2\mathrm{tr}(\mathbf{B})\sigma_q^2\mathbf{I}$. Summing over a and conjugating by $\mathbf{X}$ gives $\mathbb{E}[\mathbf{K}_V^{(1)}] = \frac{p\sigma_q^2\sigma_k^2\mathrm{tr}(\mathbf{M}^2)}{s^2}\mathbf{G}$. For the query block, independence of $\mathbf{W}_K$ and $\mathbf{w}_V$ gives $\mathbb{E}\|\mathbf{g}\|_2^2 = \frac{1}{s^2}\mathbb{E}_{\mathbf{w}_V}[\mathbf{w}_V^\top\mathbf{M}(p\sigma_k^2\mathbf{I})\mathbf{M}\mathbf{w}_V] = \frac{p\sigma_k^2\sigma_v^2\mathrm{tr}(\mathbf{M}^2)}{s^2}$, contributing this scalar times $\mathbf{G}$. For the key block, independence of $\mathbf{w}_V$ and $\mathbf{W}_Q$ gives $\mathbb{E}[\mathbf{K}_K^{(1)}] = \frac{\sigma_v^2\mathrm{tr}(\mathbf{M}^2)\,p\sigma_q^2}{s^2}\mathbf{G}$. Adding the three coefficients gives c_{I} as stated. Proportionality to $\mathbf{G}$ preserves every eigenvalue ratio exactly, so $\kappa^+(\mathbf{K}_\infty^{\mathrm{QKV,(I)}}) = \kappa_d(\mathbf{G})$. □

Corollary C.10 (The cubic regime forces $\kappa_d(\mathbf{G}) \le d/2$)**.** *Under aligned orthogonal initialization, $\tau_0 = \frac{4}{d^2}\sum_i \lambda_i^2 \ge \frac{4\lambda_1^2}{d^2}$, so the cubic-regime condition $\tau_0 \le \lambda_r^2$ forces $\kappa_d(\mathbf{G}) \le d/2$. Every dataset in the NTK and malleability study (MNIST, CIFAR-10, and the AG-News/DistilBERT embedding) satisfies $\kappa_d(\mathbf{G}) > d/2$, so, contrapositively, $\lambda^+_{\min}(\mathbf{K}^{\mathrm{QKV}}_\infty) = \lambda_r^3 + \tau_0\lambda_r > 2\lambda_r^3$, strictly above the cubic floor.*

Proof. Immediate from $\sum_i \lambda_i^2 \ge \lambda_1^2$ and Theorem C.9. For the last display, $\kappa_d(\mathbf{G}) > d/2$ gives $\tau_0 \ge 4\lambda_1^2/d^2 > \lambda_r^2$, so $\tau_0\lambda_r > \lambda_r^3$. □

Proof of Theorem 4.10. Under aligned orthogonal initialization the deviation $\mathbf{K}_H - \mathbf{K}^{\mathrm{QKV}}_\infty$ is exactly $(\bar\tau_H - \tau_0)\mathbf{G}$, so $\|\mathbf{K}_H - \mathbf{K}^{\mathrm{QKV}}_\infty\|_2 = |\bar\tau_H - \tau_0|\lambda_1$: the entire deviation lies in $\mathrm{range}(\mathbf{G})$, and no direction can escape the criterion.

Necessity. Let $\sigma_\tau^2 := \mathrm{Var}(\frac{4}{d}\|\mathbf{M}\mathbf{w}_V\|_2^2) = \frac{32}{d^4}\mathrm{tr}(\mathbf{M}^4) \ge \frac{32\lambda_1^4}{d^4}$ (a Gaussian quadratic-form variance identity, $\mathrm{Var}(\mathbf{w}^\top\mathbf{P}\mathbf{w}) = 2\sigma_v^4\|\mathbf{P}\|_F^2$ with $\mathbf{P} = \mathbf{M}^2$, $\sigma_v^2 = 1/d$). Let $\xi_h := \tau_h - \tau_0$, i.i.d. centered, $S = \sum_{h=1}^H \xi_h$, so $\bar\tau_H - \tau_0 = S/H$. Each ξ_h is a Gaussian chaos of order two, so by hypercontractivity $\beta := \mathbb{E}\xi^4/(\mathbb{E}\xi^2)^2 \le 81$, giving $\mathbb{E}S^2 = H\sigma_\tau^2$ and $\mathbb{E}S^4 \le (\beta+3)H^2\sigma_\tau^4 \le 84H^2\sigma_\tau^4$. The Paley–Zygmund inequality applied to $W = S^2$ gives $\Pr[W \ge \frac{1}{2}\mathbb{E}W] \ge \frac{1}{4\cdot 84} = \frac{1}{336}$, on which event $|S| \ge \sigma_\tau\sqrt{H/2}$, hence $\|\mathbf{K}_H - \mathbf{K}^{\mathrm{QKV}}_\infty\|_2 \ge \lambda_1\sigma_\tau/\sqrt{2H}$. If the relative spectral criterion held with probability exceeding $1 - \frac{1}{336}$, this event would intersect the criterion's success event, forcing $\varepsilon\lambda^+_{\min}(\mathbf{K}^{\mathrm{QKV}}_\infty) \ge \lambda_1\sigma_\tau/\sqrt{2H}$, i.e.

$$H \ge \frac{\sigma_\tau^2\lambda_1^2}{2\varepsilon^2\left(\lambda_r^3 + \tau_0\lambda_r\right)^2}.$$

In the lifted regime ($\tau_0 \ge \lambda_r^2$), $\lambda_r^3 + \tau_0\lambda_r \le 2\tau_0\lambda_r$, and $\sum_i \lambda_i^2 \le \lambda_1\,\mathrm{tr}(\mathbf{G}) = n\lambda_1$ gives $\tau_0 \le 4n\lambda_1/d^2$, so $\lambda^+_{\min}(\mathbf{K}^{\mathrm{QKV}}_\infty) \le 8n\lambda_1\lambda_r/d^2$. Substituting this and $\sigma_\tau^2 \ge 32\lambda_1^4/d^4$:

$$H \ge \frac{32\lambda_1^4 d^{-4}\lambda_1^2}{2\varepsilon^2\cdot 64n^2\lambda_1^2\lambda_r^2 d^{-4}} = \frac{\lambda_1^4}{4\varepsilon^2 n^2\lambda_r^2} = \frac{\theta^2\kappa_d(\mathbf{G})^2}{4\varepsilon^2}, \qquad \theta := \lambda_1/n,$$

establishing $H = \Omega(\kappa_d(\mathbf{G})^2/\varepsilon^2)$ (Corollary C.11). If instead $\tau_0 \le \lambda_r^2$ (possible only for $\kappa_d(\mathbf{G}) \le d/2$, Corollary C.10), $\lambda^+_{\min}(\mathbf{K}^{\mathrm{QKV}}_\infty) \le 2\lambda_r^3$ and the same substitution gives $H \ge 4\kappa_d(\mathbf{G})^6/(d^4\varepsilon^2)$ (Corollary C.12), the trainable-attention analog of the frozen-model $\kappa_d(\mathbf{G})^6$ width scale, deflated by d^4 because the fluctuating quantity is the single scalar $\bar\tau_H$, whereas the frozen model fluctuates over n^2 kernel entries.

Sufficiency. By Chebyshev's inequality applied to $\bar\tau_H$ (variance σ_τ^2/H) with threshold $t = \varepsilon\lambda^+_{\min}(\mathbf{K}^{\mathrm{QKV}}_\infty)/\lambda_1$,

$$\Pr\left[\|\mathbf{K}_H - \mathbf{K}^{\mathrm{QKV}}_\infty\|_2 > \varepsilon\lambda^+_{\min}(\mathbf{K}^{\mathrm{QKV}}_\infty)\right] \le \frac{\sigma_\tau^2\lambda_1^2}{H\varepsilon^2\lambda^+_{\min}(\mathbf{K}^{\mathrm{QKV}}_\infty)^2} \le \delta$$

whenever $H \ge \sigma_\tau^2\lambda_1^2/(\delta\varepsilon^2\lambda^+_{\min}(\mathbf{K}^{\mathrm{QKV}}_\infty)^2)$. Using $\sigma_\tau^2 \le \frac{8\lambda_1^2\tau_0}{d^2}$ (from $\sum_i\lambda_i^4 \le \lambda_1^2\sum_i\lambda_i^2$) and $\lambda^+_{\min}(\mathbf{K}^{\mathrm{QKV}}_\infty) \ge \tau_0\lambda_r$, together with $\tau_0 \ge 4\lambda_1^2/d^2$, this bound simplifies to $H \ge 2\kappa_d(\mathbf{G})^2/(\delta\varepsilon^2)$ in the lifted regime, matching the necessity exponent exactly and giving $H^\star = \Theta(\kappa_d(\mathbf{G})^2/\varepsilon^2)$.

For the conditioning values measured in this paper ($\kappa_d(\mathbf{G}) \approx 9.3\times10^8$ for MNIST and 4.4×10^7 for CIFAR-10, both lifted since $\kappa_d(\mathbf{G}) > d/2$; $\kappa_d(\mathbf{G}) \approx 10^7$ for the contextual embedding, $d = 768$), taking $\varepsilon = 0.1$ and $\theta \approx 0.3$ gives $H \gtrsim 10^{15}$–10^{18} for the image datasets and $H \gtrsim 10^{14}$ for the embedding, against practical head counts of order 10^2 (the largest deployed transformers use at most a few hundred attention heads). □

Corollary C.11 (Lifted regime: quadratic obstruction)**.** *In the lifted regime, $H = \Omega(\theta^2\kappa_d(\mathbf{G})^2/\varepsilon^2)$ with $\theta = \lambda_1/n$, as derived in the necessity argument above.*

Corollary C.12 (Cubic regime: sextic obstruction survives)**.** *If $\tau_0 \le \lambda_r^2$ (possible only for $\kappa_d(\mathbf{G}) \le d/2$), $H = \Omega(\kappa_d(\mathbf{G})^6/(d^4\varepsilon^2))$, as derived in the necessity argument above.*

Corollary C.13 (Independent initialization: quadratic obstruction)**.** *Under independent Gaussian initialization (Definition 4.8; any $s, p, \sigma_q, \sigma_k, \sigma_v > 0$), write $r = \operatorname{rank}(\mathbf{G})$ and $\rho := \sigma_q^2\sigma_k^2/(\sigma_q^2\sigma_k^2 + \sigma_k^2\sigma_v^2 + \sigma_q^2\sigma_v^2) \in (0,1)$. For the relative spectral criterion $\|\mathbf{K}_H - \mathbf{K}_\infty^{\mathrm{QKV},(\mathrm{I})}\|_2/\lambda_{\min}^+(\mathbf{K}_\infty^{\mathrm{QKV},(\mathrm{I})}) \le \varepsilon$: (i) if the criterion holds with probability exceeding $1 - \frac{1}{26256}$, then $H \ge \rho^2\kappa_d(\mathbf{G})^2/(p\varepsilon^2)$; (ii) for any $\delta \in (0,1)$, the criterion holds with probability at least $1-\delta$ whenever $H \ge 87\, r^2\kappa_d(\mathbf{G})^2/(\delta\varepsilon^2)$. Hence $H^\star = \Theta(\kappa_d(\mathbf{G})^2/\varepsilon^2)$, the constants depending only on (r,p) and the initialization variance ratios and on neither $\kappa_d(\mathbf{G})$ nor ε. Exponent two persists under independent initialization, with no alignment or lifted-regime hypothesis.*

Proof. Throughout, $m_2 := \operatorname{tr}(\mathbf{M}^2) = \|\mathbf{G}\|_F^2$ and $m_4 := \operatorname{tr}(\mathbf{M}^4)$, so that $m_4 \le m_2^2$, $(\operatorname{tr}\mathbf{M})^2 \le r\, m_2$ (Cauchy–Schwarz over the r nonzero eigenvalues), and $m_2 \le r\lambda_1^2$. By Theorem C.8, $\mathbf{K}_H = \frac{1}{H}\sum_{h=1}^H \mathbf{K}^{(h)}$ with per-head kernels

$$\mathbf{K}^{(h)} = \mathbf{A}_h\mathbf{G}\mathbf{A}_h^\top + \|\mathbf{g}_h\|_2^2\,\mathbf{G} + \frac{\|\mathbf{h}_h\|_2^2}{s^2}\,\mathbf{X}\mathbf{W}_Q^h(\mathbf{W}_Q^h)^\top\mathbf{X}^\top,$$

i.i.d. and positive semidefinite across heads under independent initialization. The proof of Theorem C.9 gives $\mathbb{E}[\mathbf{K}^{(h)}] = c_{\mathrm{I}}\mathbf{G}$, so the strong law of large numbers identifies $\mathbf{K}_\infty^{\mathrm{QKV},(\mathrm{I})} = c_{\mathrm{I}}\mathbf{G}$ and $\lambda_{\min}^+(\mathbf{K}_\infty^{\mathrm{QKV},(\mathrm{I})}) = c_{\mathrm{I}}\lambda_r$ exactly. Write $\Delta_H := \mathbf{K}_H - c_{\mathrm{I}}\mathbf{G} = \frac{1}{H}\sum_h(\mathbf{K}^{(h)} - c_{\mathrm{I}}\mathbf{G})$, an average of H i.i.d. centered random matrices.

Necessity: scalar reduction and the per-head second moment. Since $\mathbf{G}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$, the vector $\mathbf{v}_1 := \lambda_1^{-1/2}\mathbf{X}^\top\mathbf{u}_1$ is a unit eigenvector of $\mathbf{M}$ with eigenvalue λ_1. Fix a head, drop its index, and set $\mathbf{q} := \mathbf{W}_Q^\top\mathbf{v}_1 \in \mathbb{R}^p$, $\mathbf{B} := \mathbf{W}_K^\top\mathbf{M}^2\mathbf{W}_K \in \mathbb{R}^{p\times p}$, $\beta := \|\mathbf{M}\mathbf{w}_V\|_2^2$. Substituting $\mathbf{X}^\top\mathbf{u}_1 = \sqrt{\lambda_1}\,\mathbf{v}_1$ into the three blocks and using $\mathbf{X}^\top\mathbf{G}\mathbf{X} = \mathbf{M}^2$ gives, exactly,

$$Y := \mathbf{u}_1^\top\mathbf{K}^{(1)}\mathbf{u}_1 = \frac{\lambda_1}{s^2}\Big[\mathbf{q}^\top(\mathbf{B} + \beta\mathbf{I}_p)\mathbf{q} + \|\mathbf{W}_K^\top\mathbf{M}\mathbf{w}_V\|_2^2\Big],$$

where $\mathbf{q} \sim \mathcal{N}(0, \sigma_q^2\mathbf{I}_p)$ is independent of $(\mathbf{W}_K, \mathbf{w}_V)$ because $\mathbf{v}_1$ is a fixed unit vector. Taking expectations recovers $\mathbb{E}Y = c_{\mathrm{I}}\lambda_1$, consistent with $\mathbb{E}[\mathbf{K}^{(1)}] = c_{\mathrm{I}}\mathbf{G}$. The law of total variance and the Gaussian quadratic-form identity $\operatorname{Var}(\mathbf{q}^\top\mathbf{\Phi}\mathbf{q} \mid \mathbf{\Phi}) = 2\sigma_q^4\|\mathbf{\Phi}\|_F^2$ give

$$\sigma_1^2 := \operatorname{Var}(Y) \;\ge\; \mathbb{E}\big[\operatorname{Var}(Y \mid \mathbf{W}_K, \mathbf{w}_V)\big] = \frac{2\sigma_q^4\lambda_1^2}{s^4}\,\mathbb{E}\|\mathbf{B} + \beta\mathbf{I}_p\|_F^2 \;\ge\; \frac{2\sigma_q^4\lambda_1^2}{s^4}\,\mathbb{E}\|\mathbf{B}\|_F^2,$$

the last step because $\mathbf{B} \succeq 0$ and $\beta \ge 0$ make the cross term $2\beta\operatorname{tr}(\mathbf{B})$ nonnegative. Cauchy–Schwarz on the eigenvalues of $\mathbf{B}$ and Jensen give $\mathbb{E}\|\mathbf{B}\|_F^2 \ge \frac{1}{p}\,\mathbb{E}[(\operatorname{tr}\mathbf{B})^2] \ge \frac{1}{p}(\mathbb{E}\operatorname{tr}\mathbf{B})^2 = p\,\sigma_k^4 m_2^2$, hence

$$\sigma_1^2 \;\ge\; \frac{2p\,\sigma_q^4\sigma_k^4 m_2^2\lambda_1^2}{s^4} \;=\; \frac{2\rho^2}{p}\,(c_{\mathrm{I}}\lambda_1)^2 :$$

the per-head fluctuation of the top-direction Rayleigh quotient is a constant fraction of its mean, the constant depending only on p and the variance ratios.

Necessity: Paley–Zygmund. Let $\xi_h := \mathbf{u}_1^\top\mathbf{K}^{(h)}\mathbf{u}_1 - c_{\mathrm{I}}\lambda_1$, i.i.d. centered with variance σ_1^2, and $S := \sum_{h=1}^H \xi_h$, so that $\mathbf{u}_1^\top\Delta_H\mathbf{u}_1 = S/H$ and $\|\Delta_H\|_2 \ge |S|/H$. Each ξ_h is a polynomial of degree four in the jointly independent Gaussian entries of $(\mathbf{W}_Q^h, \mathbf{W}_K^h, \mathbf{w}_V^h)$, so Gaussian hypercontractivity ($\|P\|_4 \le 3^{k/2}\|P\|_2$ for degree $k \le 4$) gives $\beta_4 := \mathbb{E}\xi^4/(\mathbb{E}\xi^2)^2 \le 9^4 = 6561$, whence $\mathbb{E}S^2 = H\sigma_1^2$ and $\mathbb{E}S^4 \le (\beta_4 + 3)H^2\sigma_1^4 \le 6564\,H^2\sigma_1^4$. The Paley–Zygmund inequality applied to $W = S^2$ gives $\Pr[W \ge \frac{1}{2}\mathbb{E}W] \ge \frac{1}{4}(\mathbb{E}W)^2/\mathbb{E}W^2 \ge \frac{1}{26256}$, on which event $\|\Delta_H\|_2 \ge |S|/H \ge \sigma_1/\sqrt{2H}$. If the relative spectral criterion held with probability exceeding $1 - \frac{1}{26256}$, its success event would intersect this one, forcing $\varepsilon\, c_{\mathrm{I}}\lambda_r \ge \sigma_1/\sqrt{2H}$, i.e.

$$H \;\ge\; \frac{\sigma_1^2}{2\varepsilon^2 c_{\mathrm{I}}^2\lambda_r^2} \;\ge\; \frac{\rho^2}{p}\cdot\frac{\lambda_1^2}{\varepsilon^2\lambda_r^2} \;=\; \frac{\rho^2\,\kappa_d(\mathbf{G})^2}{p\,\varepsilon^2},$$

establishing (i). Unlike the aligned case, no factor $\theta = \lambda_1/n$ appears, since $\lambda_{\min}^+(\mathbf{K}_\infty^{\mathrm{QKV},(\mathrm{I})}) = c_{\mathrm{I}}\lambda_r$ is exact rather than upper-bounded.

Sufficiency. Bounding the operator norm by the Frobenius norm and applying Chebyshev, $\Pr[\|\Delta_H\|_2 > \varepsilon c_{\mathrm{I}}\lambda_r] \le \mathbb{E}\|\Delta_H\|_F^2/(\varepsilon c_{\mathrm{I}}\lambda_r)^2$, and independence across heads gives $\mathbb{E}\|\Delta_H\|_F^2 = V_F/H$ with $V_F := \mathbb{E}\|\mathbf{K}^{(1)} - c_{\mathrm{I}}\mathbf{G}\|_F^2$. Write $c_1 := p\sigma_q^2\sigma_k^2 m_2/s^2$, $c_2 := p\sigma_k^2\sigma_v^2 m_2/s^2$, $c_3 := p\sigma_q^2\sigma_v^2 m_2/s^2$, so $c_1 + c_2 + c_3 = c_{\mathrm{I}}$ and each $c_i \le c_{\mathrm{I}}$; bound the three centered blocks separately via $\|\sum_{i=1}^3 \mathbf{Z}_i\|_F^2 \le 3\sum_i \|\mathbf{Z}_i\|_F^2$ and $\mathrm{Var} \le$ second moment. Three Gaussian identities are used: $\mathbb{E}(\mathbf{w}^\top\mathbf{P}\mathbf{w})^2 = \sigma^4[(\operatorname{tr}\mathbf{P})^2 + 2\|\mathbf{P}\|_F^2]$ for $\mathbf{w} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ and symmetric $\mathbf{P}$; $\mathbb{E}\|\mathbf{z}\|_2^4 = p(p+2)\tau^2$ for $\mathbf{z} \sim \mathcal{N}(0, \tau\mathbf{I}_p)$; and, for $\mathbf{W} \in \mathbb{R}^{d\times p}$ with i.i.d. $\mathcal{N}(0,\sigma^2)$ entries and symmetric $\mathbf{B}_0 \in \mathbb{R}^{p\times p}$, $\mathbf{N} \in \mathbb{R}^{d\times d}$, the Wick identity

$$\mathbb{E}\operatorname{tr}\big(\mathbf{W}\mathbf{B}_0\mathbf{W}^\top\mathbf{N}\mathbf{W}\mathbf{B}_0\mathbf{W}^\top\mathbf{N}\big) = \sigma^4\big[(\operatorname{tr}\mathbf{B}_0)^2\operatorname{tr}(\mathbf{N}^2) + \operatorname{tr}(\mathbf{B}_0^2)\big(\operatorname{tr}(\mathbf{N}^2) + (\operatorname{tr}\mathbf{N})^2\big)\big],$$

obtained from the three pairings of the four copies of $\mathbf{W}$. *Query block:* conditioning on $\mathbf{w}_V$, $\mathbf{W}_K^\top\mathbf{M}\mathbf{w}_V \sim \mathcal{N}(0, \sigma_k^2\beta\mathbf{I}_p)$ and $\mathbb{E}\beta^2 = \sigma_v^4(m_2^2 + 2m_4) \le 3\sigma_v^4 m_2^2$, so $\mathrm{Var}(\|\mathbf{g}\|_2^2) \le \sigma_k^4\sigma_v^4 m_2^2\,[3p(p+2) - p^2]/s^4 \le 8c_2^2$, and this block contributes at most $8c_2^2\,\|\mathbf{G}\|_F^2 = 8c_2^2 m_2 \le 8r(c_2\lambda_1)^2$. *Key block:* independence of $\mathbf{w}_V$ and $\mathbf{W}_Q$ and the Wick identity with $\mathbf{B}_0 = \mathbf{I}_p$, $\mathbf{N} = \mathbf{M}$ give

$$\mathbb{E}\Big\|\frac{\beta}{s^2}\mathbf{X}\mathbf{W}_Q\mathbf{W}_Q^\top\mathbf{X}^\top\Big\|_F^2 \le \frac{3\sigma_v^4 m_2^2}{s^4}\,\sigma_q^4\big[p^2 m_2 + p\big(m_2 + (\operatorname{tr}\mathbf{M})^2\big)\big] \le \frac{6(r+p)}{p}\,c_3^2 m_2 \le 12r^2(c_3\lambda_1)^2,$$

using $(\operatorname{tr}\mathbf{M})^2 \le r m_2$, $p + 1 + r \le 2(r+p)$, $m_2 \le r\lambda_1^2$, and $(r+p)/p \le r + 1 \le 2r$. *Value block:* $\mathbf{A}\mathbf{G}\mathbf{A}^\top = s^{-2}\mathbf{X}\mathbf{W}_Q\mathbf{B}\mathbf{W}_Q^\top\mathbf{X}^\top$, so $\|\mathbf{A}\mathbf{G}\mathbf{A}^\top\|_F^2 = s^{-4}\operatorname{tr}(\mathbf{W}_Q\mathbf{B}\mathbf{W}_Q^\top\mathbf{M}\mathbf{W}_Q\mathbf{B}\mathbf{W}_Q^\top\mathbf{M})$; the Wick identity over $\mathbf{W}_Q$, then $\mathbb{E}(\operatorname{tr}\mathbf{B})^2 = \sigma_k^4(p^2 m_2^2 + 2pm_4)$ and $\mathbb{E}\operatorname{tr}(\mathbf{B}^2) = \sigma_k^4(pm_2^2 + (p^2+p)m_4)$ over $\mathbf{W}_K$, together with $m_4 \le m_2^2$, $(\operatorname{tr}\mathbf{M})^2 \le r m_2$, $p^2 + 2p \le 3p^2$, and $2 + r \le 3r$, give

$$\mathbb{E}\|\mathbf{A}\mathbf{G}\mathbf{A}^\top\|_F^2 \le \frac{\sigma_q^4\sigma_k^4}{s^4}(p^2+2p)m_2^2\big[m_2 + m_2 + r m_2\big] \le 9rc_1^2 m_2 \le 9r^2(c_1\lambda_1)^2.$$

Collecting, $V_F \le 3\big[9r^2 + 8r + 12r^2\big](c_{\mathrm{I}}\lambda_1)^2 \le 87\,r^2(c_{\mathrm{I}}\lambda_1)^2$, so

$$\Pr\big[\|\Delta_H\|_2 > \varepsilon c_{\mathrm{I}}\lambda_r\big] \;\le\; \frac{87\,r^2\kappa_d(\mathbf{G})^2}{H\varepsilon^2} \;\le\; \delta \qquad \text{whenever} \qquad H \;\ge\; \frac{87\,r^2\kappa_d(\mathbf{G})^2}{\delta\varepsilon^2},$$

establishing (ii). The two bounds match in the exponents of $\kappa_d(\mathbf{G})$ and ε and differ only by the factor $87pr^2/(\rho^2\delta)$, which depends on neither, so $H^\star = \Theta(\kappa_d(\mathbf{G})^2/\varepsilon^2)$ in the same sense as Theorem 4.10. Since the necessity constant scales as $1/p$, the invariant form of the obstruction is the total key/query parameter budget $Hp = \Omega(\rho^2\kappa_d(\mathbf{G})^2/\varepsilon^2)$. Widening each head trades H for p but cannot remove the exponent. □

Proof of Proposition 4.11. For a single head with parameter point $\theta = (\mathbf{W}_Q, \mathbf{W}_K, \mathbf{w}_V)$, define $\tau_Q := \|\mathbf{W}_K^\top\mathbf{M}\mathbf{w}_V\|_2^2/s^2$ and $\tau^\pm := \tau_Q + \lambda_{\max\,/\,\min}(\mathbf{W}_Q\mathbf{W}_Q^\top)\|\mathbf{M}\mathbf{w}_V\|_2^2/s^2$ (τ^+ uses $\lambda_{\max}$, τ^- uses $\lambda_{\min}$). By the block decomposition of Theorem C.8, for every $\mathbf{v} \in \mathbb{R}^n$, $\mathbf{v}^\top\mathbf{X}\mathbf{W}_Q\mathbf{W}_Q^\top\mathbf{X}^\top\mathbf{v} \in [\lambda_{\min}(\mathbf{W}_Q\mathbf{W}_Q^\top), \lambda_{\max}(\mathbf{W}_Q\mathbf{W}_Q^\top)] \cdot \mathbf{v}^\top\mathbf{G}\mathbf{v}$, so

$$\mathbf{A}\mathbf{G}\mathbf{A}^\top + \tau^-\mathbf{G} \;\preceq\; \mathbf{K} \;\preceq\; \mathbf{A}\mathbf{G}\mathbf{A}^\top + \tau^+\mathbf{G}$$

in the PSD order, with $\operatorname{range}(\mathbf{K}) = \operatorname{range}(\mathbf{G})$ whenever $\tau^- > 0$. Evaluating the lower sandwich at $\mathbf{u}_1$ gives $\lambda_{\max}(\mathbf{K}) \ge a_1 + \tau^-\lambda_1$, and evaluating the upper sandwich at $\mathbf{u}_r \in \operatorname{range}(\mathbf{K})$ gives $\lambda_{\min}^+(\mathbf{K}) \le a_r + \tau^+\lambda_r$; dividing gives the Rayleigh bound $\kappa^+(\mathbf{K}) \ge (a_1 + \tau^-\lambda_1)/(a_r + \tau^+\lambda_r)$. If $a_1 \ge \kappa_d(\mathbf{G})^2(a_r + \tau^+\lambda_r)$, this bound gives $\kappa^+(\mathbf{K}) \ge \kappa_d(\mathbf{G})^2$, i.e. $p(\mathbf{K}) \ge 2$.

On the aligned family $\mathbf{A} = c\mathbf{G}$, $\mathbf{W}_Q\mathbf{W}_Q^\top = \omega\mathbf{I}$, $\tau^- = \tau^+ =: \tau$ exactly (the key block reduces to $\omega\|\mathbf{M}\mathbf{w}_V\|_2^2/s^2 \cdot \mathbf{G}$), so $\mathbf{K} = c^2\mathbf{G}^3 + \tau\mathbf{G}$ exactly, sharing the eigenbasis of $\mathbf{G}$ with eigenvalues $c^2\lambda_i^3 + \tau\lambda_i$. With $t = \tau/(c^2\lambda_r^2)$,

$$\kappa^+(\mathbf{K}) = \kappa_d(\mathbf{G}) \cdot \frac{\kappa_d(\mathbf{G})^2 + t}{1 + t},$$

identical in form to the phase-transition ratio of Theorem C.9. Taking $\log_{\kappa_d(\mathbf{G})}$ gives the stated exponent formula, and the condition $a_1 \ge \kappa_d(\mathbf{G})^2(a_r + \tau^+\lambda_r)$ reduces, after dividing by λ_1^3 and substituting $a_i = c^2\lambda_i^3$, exactly to $R = a_1/a_r > \kappa_d(\mathbf{G})^2$ up to the additive term. Monotonicity of

$p(\mathbf{K})$ in t follows since $\partial_t[\log(\kappa_d(\mathbf{G})^2+t)-\log(1+t)] = 1/(\kappa_d(\mathbf{G})^2+t) - 1/(1+t) < 0$ for $\kappa_d(\mathbf{G}) > 1$. The threshold $p \geq 2 \iff (\kappa_d(\mathbf{G})^2+t)/(1+t) \geq \kappa_d(\mathbf{G}) \iff t \leq \kappa_d(\mathbf{G})$ follows by clearing denominators, and $p(0) = 1 + \log(\kappa_d(\mathbf{G})^2)/\log\kappa_d(\mathbf{G}) = 3$.

The head output is $f = \frac{1}{s}\mathbf{X}\mathbf{W}_Q\mathbf{W}_K^\top\mathbf{M}\mathbf{w}_V$, so $\mathbf{M}\mathbf{w}_V = 0$ would force $f \equiv 0$; a nonzero head therefore has $\mathbf{M}\mathbf{w}_V \neq 0$, and if additionally $\lambda_{\min}(\mathbf{W}_Q\mathbf{W}_Q^\top) > 0$ then $\tau \geq \lambda_{\min}(\mathbf{W}_Q\mathbf{W}_Q^\top)\|\mathbf{M}\mathbf{w}_V\|_2^2/s^2 > 0$, so $t > 0$ strictly and $p(\mathbf{K}) < p(0) = 3$ strictly by monotonicity: the exact cube is unattained at any functioning checkpoint.

For the scale invariance of the checkable condition, consider $\theta_\gamma := (\gamma\mathbf{W}_Q, \gamma\mathbf{W}_K, \gamma^{-2}\mathbf{w}_V)$, $\gamma > 0$; since the head is trilinear of degree $(1,1,1)$ in $(\mathbf{W}_Q, \mathbf{W}_K, \mathbf{w}_V)$, $f(\cdot;\theta_\gamma) = f(\cdot;\theta)$ for every γ, while $a_1, a_r \mapsto \gamma^4 a_1, \gamma^4 a_r$ (so R is invariant) and $\tau^\pm \mapsto \gamma^{-2}\tau^\pm$. Hence $R > \kappa_d(\mathbf{G})^2$ is necessary for condition $R > \kappa_d(\mathbf{G})^2$ to hold anywhere on the orbit (the inequality $a_1 > \kappa_d(\mathbf{G})^2 a_r$ is γ-invariant) and, together with $\gamma^6 \geq \kappa_d(\mathbf{G})^2\tau^+\lambda_r/(a_1 - \kappa_d(\mathbf{G})^2 a_r)$, sufficient for it to be reachable at some point of the orbit. On the whitening orbit $\mathbf{W}_Q\mathbf{W}_K^\top \propto \mathbf{M}^{-1}$, $\mathbf{A} \propto \mathbf{U}\mathbf{U}^\top$ (a scaled projector onto $\mathrm{range}(\mathbf{X}^\top)$), so $a_1/a_r = \lambda_1/\lambda_r = \kappa_d(\mathbf{G}) < \kappa_d(\mathbf{G})^2$: condition $R > \kappa_d(\mathbf{G})^2$ fails on the entire whitening orbit, as claimed.

Since the reparameterization $\theta \mapsto \theta_\gamma$ leaves the computed function unchanged but is not a gradient-flow trajectory, satisfiability of the checkable condition by some checkpoint computing a given function does not by itself imply reachability by a particular optimizer from a given initialization. Gradient flow under the trilinear head conserves standard balancedness invariants of multilinear parameterizations (e.g. $\|\mathbf{W}_Q\|_F^2 - \|\mathbf{W}_K\|_F^2$ and $\|\mathbf{W}_K\|_F^2 - \|\mathbf{w}_V\|_2^2$), which the map $\theta \mapsto \theta_\gamma$ violates for $\gamma \neq 1$, whereas adaptive and spectral update rules (AdamW, Muon, Shampoo) do not preserve these invariants and can move the effective scale ratio independently of direction. Whether a given trained checkpoint satisfies $R > \kappa_d(\mathbf{G})^2$ is accordingly read off directly from Definition 4.8's parameters and $\mathbf{G}$, regardless of which of these update rules produced it. □

C.7 Random Key Projections: Transfer of the Cubic Conditioning

Proposition C.14 (Transfer of Cubic Conditioning Under Random Key Projections)**.** *Adopt the setting of Proposition C.4 with $\mathbf{W}_K \in \mathbb{R}^{d\times d_k}$ i.i.d. $\mathcal{N}(0,\sigma^2)$, independent of $\mathbf{X}$, and write $r = \mathrm{rank}(\mathbf{X})$. In the* over-complete *regime, fix $t \geq 0$ with $\sqrt{d_k} > \sqrt{r} + t$, set $\rho := (\sqrt{r}+t)/\sqrt{d_k} \in (0,1)$ and $F := ((1+\rho)/(1-\rho))^2$; then, with probability at least $1 - 2e^{-t^2/2}$,*

$$F^{-1}\kappa_d(\mathbf{G}) \leq \kappa_d(\mathbf{G}_W) \leq F\,\kappa_d(\mathbf{G}), \qquad F^{-3}\kappa_d(\mathbf{G})^3 \leq \kappa_d(\mathbf{G}_W^3) \leq F^3\kappa_d(\mathbf{G})^3, \tag{33}$$

so the cubic amplification transfers up to the factor $F^3 = 1 + O(\sqrt{r/d_k})$. The standard regime is $d_k \leq r$, which covers the single-head case $d_k = d$ with $n \geq d$ and every multi-head configuration $d_k = d/h$; there, $\mathrm{rank}(\mathbf{G}_W) = d_k < r$ almost surely, so the index-r effective condition number is $\kappa_d(\mathbf{G}_W) = +\infty$ almost surely, only a one-sided bound survives at the reduced index d_k, and no matching lower bound holds. At $d_k = 1$, $\kappa^+(\mathbf{G}_W) = 1$ identically regardless of $\kappa_d(\mathbf{G})$, a rigorous whitening counterexample.

In the setting above, $\mathbf{W}_K \in \mathbb{R}^{d\times d_k}$ has i.i.d. $\mathcal{N}(0,\sigma^2)$ entries, independent of $\mathbf{X}$. Write $\mathbf{M}_K := \mathbf{W}_K\mathbf{W}_K^\top \in \mathbb{R}^{d\times d}$ (distinct from $\mathbf{M} = \mathbf{X}^\top\mathbf{X}$ used elsewhere), $\mathbf{G}_W = \mathbf{X}\mathbf{M}_K\mathbf{X}^\top$, and $r = \mathrm{rank}(\mathbf{X})$. Two standard facts about Gaussian random matrices are used without reproof. First, the singular values of an $N \times p$ ($N \geq p$) standard Gaussian matrix concentrate within $O(1)$ of $\sqrt{N} \pm \sqrt{p}$ with Gaussian tail probability $e^{-t^2/2}$ on each side (Davidson & Szarek, 2001; Vershynin, 2018). Second, the smallest singular value of a square $d \times d$ standard Gaussian matrix satisfies a two-sided small-ball bound of order $d^{-1/2}$, sharply (Rudelson & Vershynin, 2008; Edelman, 1988). Weyl monotonicity ($\mathbf{A} \preceq \mathbf{B} \Rightarrow \lambda_i(\mathbf{A}) \leq \lambda_i(\mathbf{B})$ for all i, by the Courant–Fischer variational characterization) and the Ostrowski congruence inequality ($\lambda_{\min}(\mathbf{S}^\top\mathbf{S})\lambda_i(\mathbf{A}) \leq \lambda_i(\mathbf{S}\mathbf{A}\mathbf{S}^\top) \leq \lambda_{\max}(\mathbf{S}^\top\mathbf{S})\lambda_i(\mathbf{A})$, obtained by conjugating the Loewner bounds on $\mathbf{S}^\top\mathbf{S}$ by $\mathbf{A}^{1/2}$ and applying Weyl monotonicity to the congruent pair $\mathbf{S}\mathbf{A}\mathbf{S}^\top$ and $\mathbf{A}^{1/2}\mathbf{S}^\top\mathbf{S}\mathbf{A}^{1/2}$, which share a spectrum) are used repeatedly below.

Lemma C.15 (Row-space restriction)**.** *Let $\mathbf{V} \in \mathbb{R}^{d\times r}$ be an orthonormal basis of $\mathrm{Row}(\mathbf{X})$. Then $\mathbf{W}_K^\top\mathbf{V} \in \mathbb{R}^{d_k\times r}$ has i.i.d. $\mathcal{N}(0,\sigma^2)$ entries. If $d_k \geq r$, then for every $t \geq 0$, with probability at least $1 - 2e^{-t^2/2}$,*

$$\sigma^2(\sqrt{d_k}-\sqrt{r}-t)_+^2\|\mathbf{u}\|_2^2 \;\leq\; \mathbf{u}^\top\mathbf{M}_K\mathbf{u} \;\leq\; \sigma^2(\sqrt{d_k}+\sqrt{r}+t)^2\|\mathbf{u}\|_2^2 \qquad \textit{for all } \mathbf{u} \in \mathrm{Row}(\mathbf{X}) \textit{ simultaneously.}$$

If $d_k < r$, then $\ker(\mathbf{M}_K) \cap \mathrm{Row}(\mathbf{X}) \neq \{0\}$ for every draw of $\mathbf{W}_K$, so no positive lower bound on $\mathbf{u}^\top \mathbf{M}_K \mathbf{u}$ over $\mathrm{Row}(\mathbf{X})$ is possible.

Proof. Writing $\mathbf{W}_K$'s columns as i.i.d. $\mathcal{N}(0, \sigma^2 \mathbf{I}_d)$ vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{d_k}$, each row of $\mathbf{W}_K^\top \mathbf{V}$ is $\mathbf{w}_a^\top \mathbf{V} \sim \mathcal{N}(0, \sigma^2 \mathbf{V}^\top \mathbf{V}) = \mathcal{N}(0, \sigma^2 \mathbf{I}_r)$ (using $\mathbf{V}^\top \mathbf{V} = \mathbf{I}_r$ and that $\mathbf{V}$ is fixed relative to $\mathbf{W}_K$), and the rows are independent since the columns of $\mathbf{W}_K$ are. This proves the first claim. For $d_k \geq r$, applying the Davidson–Szarek fact to $\mathbf{W}_K^\top \mathbf{V}/\sigma$ (a $d_k \times r$, i.e. tall, standard Gaussian matrix) gives, with probability $\geq 1 - 2e^{-t^2/2}$, every singular value in $\sigma \cdot [(\sqrt{d_k} - \sqrt{r} - t)_+, \sqrt{d_k} + \sqrt{r} + t]$. Writing $\mathbf{u} = \mathbf{V}\alpha \in \mathrm{Row}(\mathbf{X})$ with $\|\alpha\|_2 = \|\mathbf{u}\|_2$, $\mathbf{u}^\top \mathbf{M}_K \mathbf{u} = \|\mathbf{W}_K^\top \mathbf{u}\|_2^2 = \|(\mathbf{W}_K^\top \mathbf{V})\alpha\|_2^2 \in [s_{\min}^2, s_{\max}^2]\|\mathbf{u}\|_2^2$, giving the stated window. For $d_k < r$: $\dim \ker(\mathbf{M}_K) \geq d - d_k$, so $\dim(\ker(\mathbf{M}_K) \cap \mathrm{Row}(\mathbf{X})) \geq (d - d_k) + r - d = r - d_k > 0$ by the dimension formula, and any nonzero $\mathbf{u}$ in the intersection has $\mathbf{u}^\top \mathbf{M}_K \mathbf{u} = \|\mathbf{W}_K^\top \mathbf{u}\|_2^2 = 0$. □

Lemma C.16 (Congruence sandwich at the rank-r truncation)**.** *If $a\|\mathbf{u}\|_2^2 \leq \mathbf{u}^\top \mathbf{M}_K \mathbf{u} \leq b\|\mathbf{u}\|_2^2$ for all $\mathbf{u} \in \mathrm{Row}(\mathbf{X})$, with $0 < a \leq b$, then, deterministically, $a\mathbf{G} \preceq \mathbf{G}_W \preceq b\mathbf{G}$, $a\lambda_i(\mathbf{G}) \leq \lambda_i(\mathbf{G}_W) \leq b\lambda_i(\mathbf{G})$ for every i, $\mathrm{rank}(\mathbf{G}_W) = r$, and $\frac{a}{b}\kappa_d(\mathbf{G}) \leq \kappa_d(\mathbf{G}_W) \leq \frac{b}{a}\kappa_d(\mathbf{G})$.*

Proof. For $\mathbf{v} \in \mathbb{R}^n$, set $\mathbf{u} := \mathbf{X}^\top \mathbf{v} \in \mathrm{Row}(\mathbf{X})$; then $\mathbf{v}^\top \mathbf{G}_W \mathbf{v} = \mathbf{v}^\top \mathbf{X} \mathbf{M}_K \mathbf{X}^\top \mathbf{v} = \mathbf{u}^\top \mathbf{M}_K \mathbf{u} \in [a, b]\|\mathbf{u}\|_2^2 = [a, b]\, \mathbf{v}^\top \mathbf{G} \mathbf{v}$, giving $a\mathbf{G} \preceq \mathbf{G}_W \preceq b\mathbf{G}$ for arbitrary $\mathbf{v}$; no invertibility of $\mathbf{M}_K$ or restriction on the sign of $n - d$ is used. Weyl monotonicity applied to both inequalities gives the eigenvalue sandwich, with equality of zero eigenvalues for $i > r$. Taking $i = r$: $\lambda_r(\mathbf{G}_W) \geq a\lambda_r(\mathbf{G}) > 0$, so $\mathrm{rank}(\mathbf{G}_W) \geq r$; combined with $\mathrm{rank}(\mathbf{G}_W) \leq \mathrm{rank}(\mathbf{X}) = r$, this gives $\mathrm{rank}(\mathbf{G}_W) = r$. The condition-number bound follows by dividing the $i = 1$ and $i = r$ inequalities. □

Proof of Proposition C.14. Over-complete regime. Since $\sqrt{d_k} > \sqrt{r} + t$ gives $d_k > r$, Lemma C.15 applies with $a = \sigma^2(\sqrt{d_k} - \sqrt{r} - t)^2 = d_k \sigma^2 (1 - \rho)^2 > 0$ and $b = \sigma^2(\sqrt{d_k} + \sqrt{r} + t)^2 = d_k \sigma^2 (1 + \rho)^2$ on an event of probability $\geq 1 - 2e^{-t^2/2}$; Lemma C.16 then gives the spectrum-wise sandwich and $\kappa_d(\mathbf{G}_W) \in [F^{-1}, F]\kappa_d(\mathbf{G})$ with $b/a = F$, on the *same* event for both directions, so no further union bound is needed. Since $\mathbf{G}_W$ is PSD, $\mathbf{G}_W^3$ has eigenvalues $\lambda_i(\mathbf{G}_W)^3$ with the same eigenvectors, so $\kappa_d(\mathbf{G}_W^3) = \kappa_d(\mathbf{G}_W)^3 \in [F^{-3}, F^3]\kappa_d(\mathbf{G})^3$, cubing the two-sided bound.

Standard regime. Write $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ (compact SVD, $\mathbf{V} \in \mathbb{R}^{d \times r}$ orthonormal) and $\mathbf{N} := \mathbf{V}^\top \mathbf{M}_K \mathbf{V} = (\mathbf{W}_K^\top \mathbf{V})^\top (\mathbf{W}_K^\top \mathbf{V})$, so $\mathbf{G}_W = \mathbf{U}(\mathbf{\Sigma}\mathbf{N}\mathbf{\Sigma})\mathbf{U}^\top$, and $\mathrm{rank}(\mathbf{G}_W) = \mathrm{rank}(\mathbf{W}_K^\top \mathbf{V}) = \min(d_k, r)$ almost surely (a $d_k \times r$ Gaussian matrix has full rank almost surely, its complement being a Lebesgue-null algebraic subvariety). If $d_k < r$, $\mathrm{rank}(\mathbf{G}_W) = d_k < r$, so $\lambda_r(\mathbf{G}_W) = 0$ and the index-r ratio $\kappa_d(\mathbf{G}_W) = +\infty$ almost surely. At the reduced index $m = \min(d_k, r)$, the Ostrowski congruence inequality applied with $\mathbf{S} = \mathbf{\Sigma}$ gives $\kappa^+(\mathbf{G}_W) = \lambda_1(\mathbf{\Sigma}\mathbf{N}\mathbf{\Sigma})/\lambda_m(\mathbf{\Sigma}\mathbf{N}\mathbf{\Sigma}) \leq \kappa_d(\mathbf{G}) \cdot \kappa^+(\mathbf{N})$, and, for $d_k = r$, the Rudelson–Vershynin/Edelman facts give $\kappa^+(\mathbf{N}) = \Theta(r^2)$ with probability bounded away from 0, so $\kappa^+(\mathbf{G}_W) = \kappa_d(\mathbf{G}) \cdot O(r^2)$, unbounded, and no constant-factor transfer is provable at the critical aspect ratio. No lower bound of the form $\kappa^+(\mathbf{G}_W) \geq g(\kappa_d(\mathbf{G}))$ with $g > 1$ can hold in this regime: at $d_k = 1$, $\mathbf{G}_W = \sigma^2(\mathbf{X}\mathbf{w})(\mathbf{X}\mathbf{w})^\top$ is rank one almost surely for $\mathbf{w} \sim \mathcal{N}(0, \mathbf{I}_d)$ (since $\mathbf{X}\mathbf{w} \neq 0$ almost surely), so $\kappa^+(\mathbf{G}_W) = 1$ identically regardless of $\kappa_d(\mathbf{G})$, the claimed whitening counterexample. □

C.8 NTK Correspondence and Stability

Theorem C.17 (NTK Correspondence for Sequential Architecture)**.** *The sequential architecture Input → Attention → ReLU MLP has Neural Tangent Kernel:*

$$\mathbf{K}_{\textit{sequential}}(\mathbf{X}, \mathbf{X}') = \mathbf{K}_{\textit{ReLU}}(f^{\textit{att}}(\mathbf{X}), f^{\textit{att}}(\mathbf{X}'))$$

where the attention transformation is parameter-free.

Proof. For the sequential architecture, the forward pass computes: $g(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{\sqrt{m}} \sum_{r=1}^m a_r \sigma(w_r^T f^{\text{att}}(\mathbf{x}))$

The Neural Tangent Kernel is:

$$\begin{aligned}\mathbf{K}_{\text{sequential}}(\mathbf{x}, \mathbf{x}') &= \langle \nabla_{\boldsymbol{\theta}} g(\mathbf{x}; \boldsymbol{\theta}), \nabla_{\boldsymbol{\theta}} g(\mathbf{x}'; \boldsymbol{\theta}) \rangle \\ &= \frac{1}{m} \sum_{r=1}^{m} a_r^2 \sigma'(w_r^T f^{\text{att}}(\mathbf{x})) \sigma'(w_r^T f^{\text{att}}(\mathbf{x}')) \langle f^{\text{att}}(\mathbf{x}), f^{\text{att}}(\mathbf{x}') \rangle \end{aligned} \tag{34}$$

Since the attention transformation f^{att} is parameter-free, it acts as a fixed feature map. In the infinite-width limit, this converges to the standard ReLU NTK applied to the attention-transformed features. □

C.9 Classification within Kernel Taxonomy

Definition C.18 (Parameter-Free Gram-Induced Kernel Classification)**.** The parameter-free linearized attention kernel defines a subclass of data-dependent kernels:

Homogeneous Fourth-Order Kernels: $K_{\text{homo}}(x, y) = (x^T y)^4$ with uniform structure across all interactions.

Inhomogeneous Fourth-Order Kernels: $K_{\text{inhomo}}(x, y) = (x^T y + c)^4$ including mixed-degree terms.

Parameterized Fourth-Order Kernels: Require learned transformation parameters with $O(d^2)$ or $O(d^4)$ complexity.

Parameter-Free Data-Mediated Class: K_{LinAttn} belongs to:

$$\mathcal{K}^{(4)}_{\text{param-free}} = \{K(x, y; X) : K \text{ is 4th-order, parameter-free, data-mediated}\}$$

Remark C.19 (Computational Advantages of Parameter-Free Fourth-Order Design). The parameter-free structure provides:

1. **Memory efficiency**: $O(nd)$ storage vs. $O(d^4)$ for general fourth-order kernels
2. **Deterministic computation**: Fourth-order interactions computed directly from data geometry
3. **Transitive interactions**: Paths $(\mathbf{x}_i \to \mathbf{x}_k \to \mathbf{x}_\ell \to \mathbf{x}_j)$ enable multi-step similarity propagation
4. **Theoretical tractability**: Fixed structure enables exact mathematical analysis

C.10 Fourth-Order Representational Power Analysis

Theorem C.20 (RKHS Characterization of Linearized Attention Kernel)**.** *The RKHS $\mathcal{H}_{\text{LinAttn}}$ of the linearized attention kernel, restricted to the training set, is:*

$$\mathcal{H}_{\text{LinAttn}} = \left\{ f : f(\mathbf{x}) = \sum_{j=1}^{n} \alpha_j K_{\text{LinAttn}}(\mathbf{x}_j, \mathbf{x}), \quad \boldsymbol{\alpha} \in \mathbb{R}^n, \quad \|\boldsymbol{\alpha}\|^2_{\mathbf{K}_n} = \boldsymbol{\alpha}^T \mathbf{K}_n \boldsymbol{\alpha} < \infty \right\}$$

where $\mathbf{K}_n \in \mathbb{R}^{n \times n}$ is the kernel matrix with $[\mathbf{K}_n]_{ij} = K_{\text{LinAttn}}(\mathbf{x}_i, \mathbf{x}_j) = [\mathbf{G}^3]_{ij}$, and the RKHS norm of f is $\|f\|^2_{\mathcal{H}} = \boldsymbol{\alpha}^T \mathbf{K}_n \boldsymbol{\alpha}$. Each function in the RKHS evaluates as:

$$f(\mathbf{x}) = \sum_{j=1}^{n} \alpha_j \sum_{k,\ell=1}^{n} (\mathbf{x}_j^T \mathbf{x}_k)(\mathbf{x}_k^T \mathbf{x}_\ell)(\mathbf{x}_\ell^T \mathbf{x})$$

using a single coefficient vector $\boldsymbol{\alpha} \in \mathbb{R}^n$ (one per training point), not a matrix.

Remark C.21 (Linearity in Test Input). Despite the transitive triple-sum structure in the RKHS representation, $f(\mathbf{x})$ is *linear* in the test input $\mathbf{x}$ (the only degree-1 factor in $\mathbf{x}$ is the rightmost $\mathbf{x}_\ell^T \mathbf{x}$). The data-dependent complexity lies in the coefficient structure, where each training point $\mathbf{x}_j$ contributes through a three-step chain $\mathbf{x}_j \to \mathbf{x}_k \to \mathbf{x}_\ell \to \mathbf{x}$, with weights determined by $\mathbf{G}^2$. The representational power relative to standard polynomial kernels comes from the data-adaptive weighting of the linear basis functions $\{\mathbf{x}_\ell^T \mathbf{x}\}_{\ell=1}^n$, while the degree in $\mathbf{x}$ stays one.

C.11 Influence Function Stability

Remark C.22 (NTK influence computation and numerical stability). The NTK-based influence computation of Zhang & Zhang (2022), built on the NTK correspondence (Jacot et al., 2018; Lee et al., 2019), avoids the Hessian-inverse-vector products of IHVP methods and instead works through the regularized kernel $\mathbf{K} + \lambda\mathbf{I}$, whose conditioning is bounded by $\kappa(\mathbf{K} + \lambda\mathbf{I}) \le \lambda_{\max}(\mathbf{K})/\lambda + 1$. Influence scores and their leave-one-out updates are rank-one modifications of $(\mathbf{K} + \lambda\mathbf{I})^{-1}$ computed in closed form without retraining. The kernel is held fixed in these updates, following the fixed-kernel (resolvent) definition of influence, so the feature map is not re-derived when a training point is removed, which matches how influence functions approximate leave-one-out retraining without carrying it out. The quantitative stability guarantee is the resolvent bound of Theorem C.23 below.

Theorem C.23 (Influence Function Stability)**.** *Let $\mathbf{K} \in \mathbb{R}^{n\times n}$ be a positive semidefinite NTK matrix with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_n \ge 0$ and regularization $\lambda > 0$. Then $\kappa(\mathbf{K} + \lambda\mathbf{I}) \le \lambda_1/\lambda + 1$, and for $\|\Delta\mathbf{K}\|_2 \le \epsilon < \lambda$,*

$$\|(\mathbf{K} + \lambda\mathbf{I})^{-1} - (\mathbf{K} + \Delta\mathbf{K} + \lambda\mathbf{I})^{-1}\|_2 \le \frac{\epsilon}{\lambda(\lambda - \epsilon)}\,.$$

Proof. The resolvent identity gives $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ for $A = \mathbf{K} + \lambda\mathbf{I}$ and $B = \mathbf{K} + \Delta\mathbf{K} + \lambda\mathbf{I}$. Since $\mathbf{K} \succeq 0$, $\|A^{-1}\|_2 \le 1/\lambda$; since $\|\Delta\mathbf{K}\|_2 \le \varepsilon < \lambda$, Weyl's inequality gives $\lambda_{\min}(B) \ge \lambda - \varepsilon$, so $\|B^{-1}\|_2 \le 1/(\lambda - \varepsilon)$. Taking operator norms in the resolvent identity, $\|A^{-1}(B - A)B^{-1}\|_2 \le \|A^{-1}\|_2\|B - A\|_2\|B^{-1}\|_2 \le \varepsilon/(\lambda(\lambda - \varepsilon))$, which is the stated bound. The conditioning bound $\kappa(\mathbf{K} + \lambda\mathbf{I}) \le \lambda_1/\lambda + 1$ follows from $\lambda_{\max}(\mathbf{K} + \lambda\mathbf{I}) = \lambda_1 + \lambda$ and $\lambda_{\min}(\mathbf{K} + \lambda\mathbf{I}) \ge \lambda$. □

C.12 Proof of Proposition C.24

Proposition C.24 (Data-Dependent Kernel Sensitivity)**.** *The linearized attention kernel K_{LinAttn} exhibits* data-dependent *sensitivity to input perturbations that depends on the Gram matrix structure. For perturbation δ with $\|\delta\|_2 \le \epsilon$:*

$$|K_{\text{LinAttn}}(\mathbf{x}_i + \delta, \mathbf{x}_j) - K_{\text{LinAttn}}(\mathbf{x}_i, \mathbf{x}_j)| \le 6\lambda_1(\mathbf{G})^2\,\epsilon \tag{35}$$

where $\mathbf{G} = \mathbf{X}\mathbf{X}^T$ and $\lambda_1(\mathbf{G})$ is its spectral norm. The key distinction from polynomial kernels $K_{\text{poly}}(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y})^p$ (which have $O(\epsilon)$ sensitivity independent of data) is that K_{LinAttn} sensitivity grows with the spectral structure of the dataset through $\lambda_1(\mathbf{G})$.

Remark C.25 (Connection to Feature Learning). Although $\mathbf{X}$ is fixed during training, $f^{\text{att}}(\mathbf{X}) = \mathbf{X}\mathbf{X}^T\mathbf{X}$ encodes global pairwise relationships. When MLP weights are updated, gradients propagate through this globally-coupled representation, causing the empirical NTK to evolve in ways that depend on the full correlation structure, the hallmark of feature learning, in contrast to frozen-kernel lazy training.

Proof. Write $\mathbf{M} := \mathbf{X}^T\mathbf{X} \in \mathbb{R}^{d\times d}$, so that $K_{\text{LinAttn}}(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{M}\mathbf{x}_i, \mathbf{M}\mathbf{x}_j\rangle = \mathbf{x}_i^T\mathbf{M}^2\mathbf{x}_j$ and $\|\mathbf{M}\|_2 = \lambda_1(\mathbf{G})$. Perturbing $\mathbf{x}_i$ changes both the explicit factor and $\mathbf{M}$; the change in $\mathbf{M}$ is $\Delta\mathbf{M} = \mathbf{x}_i\delta^T + \delta\mathbf{x}_i^T + \delta\delta^T$, a rank-$\le 2$ update of the $d \times d$ operator, for which the spectral-norm bound genuinely holds: $\|\Delta\mathbf{M}\|_2 \le 2\|\mathbf{x}_i\|_2\|\delta\|_2 + \|\delta\|_2^2 \le 2\epsilon + \epsilon^2$ (the analogous bound on the $n \times n$ Gram $\mathbf{G}$ itself does not hold in general). Hence $\|\Delta(\mathbf{M}^2)\|_2 = \|\Delta\mathbf{M}\,\mathbf{M} + \mathbf{M}\,\Delta\mathbf{M} + \Delta\mathbf{M}\Delta\mathbf{M}\|_2 \le 2\lambda_1(\mathbf{G})\|\Delta\mathbf{M}\|_2 + O(\epsilon^2) \le 4\lambda_1(\mathbf{G})\epsilon + O(\epsilon^2)$. Therefore, to first order,

$$\begin{aligned}|\Delta K_{ij}| &= |\delta^T\mathbf{M}^2\mathbf{x}_j + \mathbf{x}_i^T\Delta(\mathbf{M}^2)\mathbf{x}_j| \le \|\delta\|_2\|\mathbf{M}^2\mathbf{x}_j\|_2 + \|\mathbf{x}_i\|_2\|\Delta(\mathbf{M}^2)\|_2\|\mathbf{x}_j\|_2 \\ &\le \lambda_1(\mathbf{G})^2\epsilon + 4\lambda_1(\mathbf{G})\epsilon \le 6\lambda_1(\mathbf{G})^2\epsilon,\end{aligned}$$

using $\lambda_1(\mathbf{G}) \ge 1$ (unit-norm rows give $\operatorname{tr}\mathbf{G} = n$, and $\lambda_1(\mathbf{G})$ is at least the average eigenvalue $n/r \ge 1$ over the $r \le n$ positive eigenvalues). For polynomial kernels, $|\nabla_{\mathbf{x}}(\mathbf{x}^T\mathbf{y})^p| \cdot \epsilon = O(\epsilon)$, independent of the dataset spectrum. □

C.13 Proof of Spectral Amplification (Theorem 4.4)

The proof relies on the following identity.

Lemma C.26 (Spectral Transfer). *For any* $\mathbf{X} \in \mathbb{R}^{n\times d}$ *with* $\mathrm{rank}(\mathbf{X}) = r$ *and any* $k \geq 0$*:*

$$\mathbf{X}(\mathbf{X}^T\mathbf{X})^k\mathbf{X}^T = (\mathbf{X}\mathbf{X}^T)^{k+1}$$

Proof. Let $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ be the compact SVD with $\mathbf{U} \in \mathbb{R}^{n\times r}$, $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$, $\mathbf{V} \in \mathbb{R}^{d\times r}$, $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}_r$. Then $\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T$ and $(\mathbf{X}^T\mathbf{X})^k = \mathbf{V}\mathbf{\Sigma}^{2k}\mathbf{V}^T$, so:

$$\mathbf{X}(\mathbf{X}^T\mathbf{X})^k\mathbf{X}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \cdot \mathbf{V}\mathbf{\Sigma}^{2k}\mathbf{V}^T \cdot \mathbf{V}\mathbf{\Sigma}\mathbf{U}^T = \mathbf{U}\mathbf{\Sigma}^{2k+2}\mathbf{U}^T = (\mathbf{X}\mathbf{X}^T)^{k+1}$$

using $\mathbf{V}^T\mathbf{V} = \mathbf{I}_r$ at each step. □

Proof of Theorem 4.4. The attention output admits the SVD factorization $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{X}^T\mathbf{X} = \mathbf{U}\mathbf{\Sigma}^3\mathbf{V}^T$, where $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ is the compact SVD with $\mathbf{U}^T\mathbf{U} = \mathbf{I}_r$ ($r = \mathrm{rank}(\mathbf{X})$). By Lemma C.26 with $k = 1$, the Gram matrix of the transformed features is

$$\tilde{\mathbf{G}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T = \mathbf{U}\mathbf{\Sigma}^6\mathbf{U}^T = (\mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^T)^3 = \mathbf{G}^3.$$

Since $\mathbf{G}$ has $r = \mathrm{rank}(\mathbf{X})$ positive eigenvalues $\lambda_1 \geq \cdots \geq \lambda_r > 0$, the transformed Gram matrix $\tilde{\mathbf{G}}$ has positive eigenvalues λ_i^3 with effective condition number $\kappa_d(\tilde{\mathbf{G}}) = \lambda_1^3/\lambda_r^3 = \kappa_d(\mathbf{G})^3$.

Consider now the finite-width NTK. The (i, j)-th entry is $[\mathbf{K}_m]_{ij} = \frac{1}{m}\sum_{r=1}^m Z_r^{(ij)}$ where $Z_r^{(ij)} = \mathbf{1}[\mathbf{w}_r^T\tilde{\mathbf{x}}_i > 0]\,\mathbf{1}[\mathbf{w}_r^T\tilde{\mathbf{x}}_j > 0]\,\langle\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j\rangle$. Defining the centered matrix $\mathbf{S}_r$ with entries $[\mathbf{S}_r]_{ij} = Z_r^{(ij)} - \mathbb{E}[Z_r^{(ij)}]$, the deviation from the infinite-width kernel is $\mathbf{K}_m - \mathbf{K}_\infty = \frac{1}{m}\sum_{r=1}^m \mathbf{S}_r$, a sum of m independent centered random matrices. Each satisfies $\|\mathbf{S}_r\|_2 \leq 2\|\mathbf{G}_r\|_2 \leq 2\lambda_1^3$. To see this, note that $\mathbf{G}_r = \mathbf{H}_r\mathbf{H}_r^T$ where $\mathbf{H}_r$ equals $\tilde{\mathbf{X}}$ with rows zeroed by the indicator $\mathbf{1}[\mathbf{w}_r^T\tilde{\mathbf{x}}_i > 0]$; row deletion cannot increase the spectral norm, so $\|\mathbf{H}_r\|_2 \leq \|\tilde{\mathbf{X}}\|_2 = \lambda_1^{3/2}$ and $\|\mathbf{G}_r\|_2 = \|\mathbf{H}_r\|_2^2 \leq \lambda_1^3$. The matrix variance statistic $\sigma^2 = \frac{1}{m}\|\mathbb{E}[\mathbf{S}_1^2]\|_2$ satisfies $\sigma^2 \leq \lambda_1^6/(2m)$, bounding the variance by its operator norm instead of the trace. Writing $\mathbf{S}_1 = \mathbf{G}_1 - \mathbf{K}_{\infty,\mathrm{seq}}$ with $\mathbb{E}[\mathbf{G}_1] = \mathbf{K}_{\infty,\mathrm{seq}}$, the variance is $\mathbb{E}[\mathbf{S}_1^2] = \mathbb{E}[\mathbf{G}_1^2] - \mathbf{K}_{\infty,\mathrm{seq}}^2 \preceq \mathbb{E}[\mathbf{G}_1^2]$; since $\mathbf{G}_1 \succeq 0$ with $\|\mathbf{G}_1\|_2 \leq \lambda_1^3$ gives $\mathbf{G}_1^2 \preceq \lambda_1^3\mathbf{G}_1$, taking expectations yields $\mathbb{E}[\mathbf{G}_1^2] \preceq \lambda_1^3\mathbf{K}_{\infty,\mathrm{seq}}$, so $\|\mathbb{E}[\mathbf{S}_1^2]\|_2 \leq \lambda_1^3\|\mathbf{K}_{\infty,\mathrm{seq}}\|_2$. The arc-cosine decomposition $\mathbf{K}_{\infty,\mathrm{seq}} = \mathbf{A} \circ \mathbf{G}^3$ established below has $\max_i[\mathbf{A}]_{ii} = \frac{1}{2}$, so the Schur upper bound $\lambda_{\max}(\mathbf{A} \circ \mathbf{B}) \leq (\max_i[\mathbf{A}]_{ii})\lambda_{\max}(\mathbf{B})$ for $\mathbf{A}, \mathbf{B} \succeq 0$ (Horn & Johnson, 1994) gives $\|\mathbf{K}_{\infty,\mathrm{seq}}\|_2 \leq \frac{1}{2}\lambda_1^3$, hence $\|\mathbb{E}[\mathbf{S}_1^2]\|_2 \leq \frac{1}{2}\lambda_1^6$ with no factor of n (the earlier trace relaxation $\|\cdot\|_F^2 \leq n\|\cdot\|_2^2$ was pure rank slack). Applying the matrix Bernstein inequality (Tropp, 2012) yields

$$\|\mathbf{K}_m - \mathbf{K}_\infty\|_2 \leq C_0\left(\sqrt{\frac{\lambda_1^6\log n}{m}} + \frac{\lambda_1^3\log n}{m}\right). \tag{36}$$

For $m \gg \log n$, the first term dominates: $\|\mathbf{K}_m - \mathbf{K}_\infty\|_2 = O(\lambda_1^3\sqrt{\log n/m})$.

For the NTK approximation to be meaningful, the error must be small relative to $\lambda_{\min}(\mathbf{K}_\infty)$. The infinite-width NTK of the two-layer ReLU on transformed features $\tilde{\mathbf{x}}_i$ decomposes as $\mathbf{K}_{\infty,\mathrm{seq}} = \mathbf{A} \circ \mathbf{G}^3$ (Hadamard product), where $[\mathbf{A}]_{ij} = (\pi - \arccos(\hat{\tilde{\mathbf{x}}}_i^T\hat{\tilde{\mathbf{x}}}_j))/(2\pi)$ is the arc-cosine kernel evaluated on row-normalized features $\hat{\tilde{\mathbf{x}}}_i = \tilde{\mathbf{x}}_i/\|\tilde{\mathbf{x}}_i\|_2$. Both $\mathbf{A}$ and $\mathbf{G}^3$ are PSD. The diagonal entries satisfy $[\mathbf{A}]_{ii} = 1/2$ for all i (since $\hat{\tilde{\mathbf{x}}}_i^T\hat{\tilde{\mathbf{x}}}_i = 1$ gives $\arccos(1) = 0$). For $n \leq d$ with $\mathrm{rank}(\mathbf{X}) = n$, $\mathbf{G}^3$ is full rank and the Schur product inequality (Horn & Johnson, 1994) ($\lambda_{\min}(\mathbf{A} \circ \mathbf{B}) \geq \min_i[\mathbf{A}]_{ii} \cdot \lambda_{\min}(\mathbf{B})$ for $\mathbf{A}, \mathbf{B} \succeq 0$) gives $\lambda_{\min}(\mathbf{K}_{\infty,\mathrm{seq}}) \geq \min_i[\mathbf{A}]_{ii} \cdot \lambda_{\min}(\mathbf{G}^3) = \frac{1}{2}\lambda_r^3$, and the convergence condition $\|\mathbf{K}_m - \mathbf{K}_\infty\|_2/\lambda_{\min}(\mathbf{K}_\infty) \leq \epsilon$ requires

$$\frac{\lambda_1^3\sqrt{\log n/m}}{\lambda_r^3/2} \leq \epsilon \quad\Longrightarrow\quad m \geq \frac{4\kappa_d(\mathbf{G})^6\log n}{\epsilon^2}.$$

For $n > d$, $\mathbf{G}^3$ is rank-deficient ($\lambda_{\min}(\mathbf{G}^3) = 0$), so this form of the Schur bound is vacuous. The valid direction is instead $\lambda_{\min}(\mathbf{A} \circ \mathbf{G}^3) \geq \lambda_{\min}(\mathbf{A}) \cdot \min_i[\mathbf{G}^3]_{ii}$, where $\min_i[\mathbf{G}^3]_{ii} = \min_i\|\tilde{\mathbf{x}}_i\|^2 \geq 1$ unconditionally (Cauchy–Schwarz with $\|\mathbf{x}_i\| = 1$), and, under the transformed-feature incoherence $\tilde{\mu} = \max_{i\neq j}|\hat{\mathbf{G}}_{ij}| < 1$,

$$\lambda_{\min}(\mathbf{A}) \geq \frac{1}{12\pi}\sqrt{\frac{2\log(1/\tilde{\mu})}{\log(2(n-1)) + 2\log(1/\tilde{\mu})}} > 0,$$

obtained by expanding the arc-cosine kernel $\mathbf{A}$ into non-negative-coefficient Hadamard powers of $\hat{\mathbf{G}}$ (each PSD with unit diagonal) and bounding the tail by Gershgorin. Then $\mathbf{K}_{\infty,\text{seq}}$ is strictly positive definite ($\lambda^+_{\min} = \lambda_{\min}$), and the convergence condition requires the honest sufficient width

$$m = \Omega\Big(\lambda_1^6 \frac{\log n}{\log(1/\tilde{\mu})} \frac{\log n}{\varepsilon^2}\Big) = \Omega\Big(\kappa_d(\mathbf{G})^6 \lambda_r^6 \frac{\log n}{\log(1/\tilde{\mu})} \frac{\log n}{\varepsilon^2}\Big),$$

differing from the κ_d^6 form by the factor $\lambda_r^6 \log n / \log(1/\tilde{\mu})$, which is $\ll 1$ when $\lambda_r \ll 1$. The physical infeasibility, driven by the λ_1^6 scale, is unchanged. For 2L-ReLU on raw unit-norm inputs the kernel entries $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ are $O(1)$ and carry no spectral amplification, so the $O(n)$ entry count sets the variance statistic, and the analogous argument gives $\|\mathbf{K}_m - \mathbf{K}_\infty\|_2 = O(\sqrt{n \log n / m})$ with $\lambda_{\min}(\mathbf{K}_\infty) = \Theta(1)$, so $m = O(n \log n / \epsilon^2)$ suffices without spectral amplification. The factor n that the operator-norm bound removes for attention, where $\|\mathbf{K}_{\infty,\text{seq}}\|_2$ carries the amplified scale λ_1^3, is genuine here, so the two widths are governed by different mechanisms, sample count for ReLU and cubic conditioning for attention.

Finally, when row-normalization is applied, the effective Gram matrix becomes $\hat{\mathbf{G}} = \mathbf{D}^{-1}\mathbf{G}^3\mathbf{D}^{-1}$ where $\mathbf{D} = \text{diag}(\|\tilde{\mathbf{x}}_1\|, \ldots, \|\tilde{\mathbf{x}}_n\|)$. Under the μ-incoherence condition (Assumption 4.3), incoherence delocalizes the eigenvectors, so the leverage scores $[\mathbf{U}]_{ik}^2$ are nearly uniform across rows and the diagonal entries $[\mathbf{G}^3]_{ii} = \sum_k \lambda_k^3 [\mathbf{U}]_{ik}^2$ have a spread across rows that is small relative to their common mean $\frac{1}{n}\text{tr}(\mathbf{G}^3)$, vanishing as the incoherence level $\mu/\sqrt{d} \to 0$. Hence $\mathbf{D}$ is approximately a scalar multiple of $\mathbf{I}$ and $\kappa_d(\hat{\mathbf{G}}) = \kappa_d(\mathbf{G})^3(1 + o(1))$. □

Remark C.27 (Rank-deficient case ($n > d$)). For $n \le d$ the Schur floor $\lambda_{\min}(\mathbf{K}_{\infty,\text{seq}}) \ge \frac{1}{2}\lambda_r^3$ yields the clean $\kappa_d(\mathbf{G})^6$ width certificate. For $n > d$ (the regime of all experiments) $\mathbf{G}^3$ is rank-deficient and that step is vacuous. The floor instead holds via $\lambda_{\min}(\mathbf{A}) \cdot \min_i [\mathbf{G}^3]_{ii}$ under the transformed-feature incoherence $\tilde{\mu} < 1$, under which $\mathbf{K}_{\infty,\text{seq}} = \mathbf{A} \circ \mathbf{G}^3$ is generically *full rank*, since the Hadamard product with the arc-cosine kernel restores the rank that $\mathbf{G}^3$ loses. The failure mode is then the collapse $\lambda_{\min} \to 0$ as transformed features coalesce ($\tilde{\mu} \to 1$), with no null space present, and the honest sufficient width for $n > d$ carries the extra factor $\lambda_r^6 \log n / \log(1/\tilde{\mu})$ relative to $\kappa_d(\mathbf{G})^6$. Measured on the controlled datasets (unit-norm rows, 1000 examples per class), $\tilde{\mu} \approx 0.99998$ (MNIST) and 0.99995 (CIFAR-10), just below one, so $\mathbf{K}_{\infty,\text{seq}}$ stays full rank while the small $\log(1/\tilde{\mu})$ inflates the $n > d$ width further. Here $\kappa_d(\mathbf{G}) = \lambda_1/\lambda_r$ is the ratio of the largest to smallest positive eigenvalue.

C.14 Proof of Theorem 4.5

Proof. The diagonal entry $[\mathbf{K}_{m,\text{seq}}]_{11} = \frac{1}{m}\sum_{r=1}^m \xi_{r,1}$ is an average of m i.i.d. Bernoulli$(1/2)$ variables ($\xi_{r,1} = \mathbf{1}[\mathbf{w}_r^T \hat{\mathbf{x}}_1 > 0]$, symmetric by Gaussian sign invariance), with mean $[\mathbf{K}_{\infty,\text{seq}}]_{11} = 1/2$. Since a single entry lower-bounds the spectral norm of the deviation, $\|\mathbf{K}_{m,\text{seq}} - \mathbf{K}_{\infty,\text{seq}}\|_2 \ge |[\mathbf{K}_{m,\text{seq}}]_{11} - 1/2|$, so $\mathbb{E}\|\mathbf{K}_{m,\text{seq}} - \mathbf{K}_{\infty,\text{seq}}\|_2^2 \ge \text{Var}([\mathbf{K}_{m,\text{seq}}]_{11}) = 1/(4m)$. Applying the Paley–Zygmund inequality to $W = \big(\sum_{r=1}^m (\xi_{r,1} - \frac{1}{2})\big)^2$, whose summands are mean-zero, bounded by $1/2$, with variance $1/4$, gives $\Pr[W \ge \frac{1}{2}\mathbb{E}W] \ge (\mathbb{E}W)^2/(4\,\mathbb{E}W^2) \ge 1/16$; on this event $|[\mathbf{K}_{m,\text{seq}}]_{11} - 1/2| \ge (8m)^{-1/2}$. If the relative spectral criterion held with probability exceeding $15/16$, this event and the criterion's success event would intersect, forcing $\varepsilon\lambda_{\min}(\mathbf{K}_{\infty,\text{seq}}) \ge (8m)^{-1/2}$, which rearranges to (8). □

The following construction establishes that the exponent $\kappa_d(\mathbf{G})^3$ in Theorem 4.5 is tight and cannot be improved to a lower power.

Proof of the extremal exponent. Fix $\kappa = \kappa_d(\mathbf{G}) > 1$ and consider the two-point family $\mathbf{x}_{1,2} = (\cos a, \pm \sin a)$ with $\cos 2a = (1 - \kappa^{-1})/(1 + \kappa^{-1})$, so that $\mathbf{G}$ has eigenvalues $1 \pm \cos 2a$ and $\kappa_d(\mathbf{G}) = \kappa$ exactly. Diagonalizing $\mathbf{G}^3$ gives the exact identity $\hat{G}_{12} = [\mathbf{G}^3]_{12}/[\mathbf{G}^3]_{11} = (1 - \kappa^{-3})/(1 + \kappa^{-3})$, so the transformed angle satisfies $\tilde{\theta}_{12} = \arccos \hat{G}_{12} = 2\kappa^{-3/2}(1 + o(1))$. Since $[\mathbf{A}]_{12} = (\pi - \tilde{\theta}_{12})/(2\pi)$ and $[\hat{\mathbf{G}}]_{ii} = 1$, the 2×2 kernel matrix $\mathbf{K}_{\infty,\text{seq}}$ has smallest eigenvalue

$$\lambda_{\min}(\mathbf{K}_{\infty,\text{seq}}) = \tfrac{1}{2}(1 - \cos\tilde{\theta}_{12}) + \tfrac{\tilde{\theta}_{12}}{2\pi}\cos\tilde{\theta}_{12} = (1 + o(1))\,\kappa^{-3/2}/\pi.$$

Substituting into (8) gives the necessity $m \geq \pi^2\kappa^3/(8\varepsilon^2)\,(1+o(1))$. A matching sufficiency on the same family follows from Chebyshev's inequality applied to the exactly computable two-entry deviation, $\mathbb{E}\|\mathbf{K}_{m,\text{seq}} - \mathbf{K}_{\infty,\text{seq}}\|_F^2 \leq 1/m$, giving $m = O(\kappa^3/\varepsilon^2)$: the true convergence threshold of this family is $\Theta(\kappa^3/\varepsilon^2)$, so the cubic exponent is attained and cannot be improved to a lower power. For $n \leq d$, Cauchy interlacing on every principal 2×2 submatrix of $\mathbf{G}$ gives $\theta_{ij}^{\text{abs}} \geq \sqrt{2\lambda_r}$ for the raw pair, and the angle-distortion bound of the row-normalization map (a Lipschitz diffeomorphism of the unit sphere with modulus in $[\kappa^{-1},\kappa]$) gives $\tilde{\theta}_{\min}^{\text{abs}} \geq \sqrt{2}\,\kappa^{-3/2}$ unconditionally, hence $\lambda_{\min}(\mathbf{K}_{\infty,\text{seq}}) = \Omega(\kappa^{-3/2}/\sqrt{\log n})$; no correct necessity bound for the row-normalized architecture can therefore exceed $O(\kappa^3\,\text{poly}(n)/\varepsilon^2)$, and an exponent of 6 is unattainable as a necessity statement, though it remains the correct sufficiency constant for the unnormalized construction of Theorem 4.4, whose diagonal entries $[\mathbf{G}^3]_{ii}$ carry the extra factor of κ^3 that the unit diagonal of $\hat{\mathbf{G}}$ normalizes away. □

C.15 Multi-Class Extension Proofs

Theorem C.28 (Multi-Class Kernel Ridge Regression Equivalence)**.** *The vectorized implementation with one-hot encoding correctly extends to multi-class classification without architectural modifications.*

Proof. For C classes, the one-hot encoded label matrix $\mathbf{Y} \in \{0,1\}^{n\times C}$ satisfies $\sum_{c=1}^{C}[\mathbf{Y}]_{ic} = 1$ for all i.

The multi-class kernel ridge regression objective is:

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^{n\times C}} \frac{1}{2}\|\mathbf{Y} - \mathbf{K}\boldsymbol{\alpha}\|_F^2 + \frac{\lambda}{2}\text{Tr}(\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha})$$

Taking the gradient and setting to zero yields the closed-form solution: $\boldsymbol{\alpha}^* = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{Y}$

This demonstrates that the same kernel matrix $\mathbf{K}$ applies to all classes simultaneously, with the solution operating column-wise across the label matrix $\mathbf{Y}$. □

C.16 Finite-Width Approximation Analysis

Theorem C.29 (Finite-Width NTK Approximation)**.** *For networks with width m, the finite-width NTK approximates the infinite-width kernel with squared error:*

$$\mathbb{E}[|\epsilon_m(\mathbf{x},\mathbf{x}')|^2] \leq \frac{C}{m}$$

for some constant C depending on the input distribution and network architecture, where $\epsilon_m(\mathbf{x},\mathbf{x}') = K_m(\mathbf{x},\mathbf{x}') - K_\infty(\mathbf{x},\mathbf{x}')$.

Proof. The finite-width NTK entry is the sample mean $K_m(\mathbf{x},\mathbf{x}') = \frac{1}{m}\sum_{r=1}^{m} Z_r$ of m i.i.d. bounded random variables $Z_r = \sigma'(\mathbf{w}_r^T\mathbf{x})\sigma'(\mathbf{w}_r^T\mathbf{x}')\langle\mathbf{x},\mathbf{x}'\rangle$ with $\mathbb{E}[Z_r] = K_\infty(\mathbf{x},\mathbf{x}')$. By the variance of a sample mean: $\mathbb{E}[|\epsilon_m|^2] = \text{Var}(Z_r)/m \leq C/m$ where $C = \text{Var}(Z_1) \leq \langle\mathbf{x},\mathbf{x}'\rangle^2 \leq 1$ for unit-norm inputs. For the sequential architecture with attention preprocessing, the attention transformation f^{att} is deterministic and parameter-free, so the same argument applies to the transformed features $\tilde{\mathbf{x}}_i$ with C scaling as $\|\tilde{\mathbf{G}}\|_2^2 = \lambda_1^6$ (the squared leading eigenvalue of the transformed Gram matrix). □

C.17 Attention-Specific Finite-Width Analysis

Theorem C.30 (Attention Feature Concentration)**.** *Let $\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathbb{R}^d$ be fixed datapoints with $\|\mathbf{x}_i\|_2 = 1$ for all i. Then the attention-transformed features satisfy the deterministic bound $\|[\mathbf{X}\mathbf{X}^T\mathbf{X}]_i\|_2 \leq n$ for all i. If additionally $\mathbf{x}_k \sim \text{Uniform}(\mathbb{S}^{d-1})$ i.i.d., then $\mathbb{E}[\|[\mathbf{X}\mathbf{X}^T\mathbf{X}]_i\|_2^2] = 1 + \Theta(n/d) + \Theta(n^2/d^2)$, so for $n > d$ the dominant term is $\Theta(n^2/d^2)$, not $O(n/d)$.*

Proof. The i-th transformed feature is $[\mathbf{X}\mathbf{X}^T\mathbf{X}]_i = \sum_{k=1}^{n}(\mathbf{x}_i^T\mathbf{x}_k)\mathbf{x}_k$ (including $k=i$, which contributes the leading 1). Under normalization $\|\mathbf{x}_k\|_2 = 1$, the triangle inequality gives the

deterministic bound

$$\|[\mathbf{X}\mathbf{X}^T\mathbf{X}]_i\|_2 \leq \sum_{k=1}^{n} |\mathbf{x}_i^T\mathbf{x}_k| \|\mathbf{x}_k\|_2 \leq n.$$

For the second moment, $\mathbb{E}\|[\mathbf{X}\mathbf{X}^T\mathbf{X}]_i\|_2^2 = \sum_{k,\ell=1}^{n} \mathbb{E}[(\mathbf{x}_i^T\mathbf{x}_k)(\mathbf{x}_i^T\mathbf{x}_\ell)(\mathbf{x}_k^T\mathbf{x}_\ell)]$. The diagonal terms $k = \ell$ give $\sum_k \mathbb{E}[(\mathbf{x}_i^T\mathbf{x}_k)^2] = 1+(n-1)/d = 1+\Theta(n/d)$ (the $k = i$ term contributes exactly 1; each $k \neq i$ term contributes $\mathbb{E}[(\mathbf{x}_i^T\mathbf{x}_k)^2] = 1/d$ under the uniform-sphere assumption). The off-diagonal terms $k \neq \ell$ (both $\neq i$) do *not* vanish in expectation: integrating over $\mathbf{x}_k$ first, $\mathbb{E}_{\mathbf{x}_k}[(\mathbf{x}_i^T\mathbf{x}_k)(\mathbf{x}_k^T\mathbf{x}_\ell)] = (\mathbf{x}_i^T\mathbf{x}_\ell)/d$, so each such term equals $\mathbb{E}[(\mathbf{x}_i^T\mathbf{x}_\ell)^2]/d = 1/d^2 > 0$; summing over the $\Theta(n^2)$ ordered pairs $k \neq \ell$ gives $\Theta(n^2/d^2)$. Combining, $\mathbb{E}\|[\mathbf{X}\mathbf{X}^T\mathbf{X}]_i\|_2^2 = 1 + \Theta(n/d) + \Theta(n^2/d^2)$; for $n > d$ (the regime of both datasets used) the $\Theta(n^2/d^2)$ term dominates $\Theta(n/d)$. The downstream use of this theorem (Proposition C.31) relies only on the deterministic bound $\|[\mathbf{X}\mathbf{X}^T\mathbf{X}]_i\|_2 \leq n$, which is unaffected by this correction. □

Proposition C.31 (Sequential Architecture Operator-Norm Bounds)**.** *Under the conditions of Theorem 4.4, the matrix Bernstein inequality applied to the ReLU layer on transformed features $\tilde{\mathbf{x}}_i$ (with $\|\tilde{\mathbf{x}}_i\|_2 \leq n$ for unit-norm inputs from Theorem C.30) yields:*

$$\|\mathbf{K}_{m,seq} - \mathbf{K}_{\infty,seq}\|_2 \leq C_0 \left(\lambda_1^3 \sqrt{\frac{\log n}{m}} + \frac{\lambda_1^3 \log n}{m} \right)$$

where C_0 is the universal constant from the matrix Bernstein inequality (Tropp, 2012).

Proof. The entries of $\mathbf{K}_{m,\text{seq}}$ are sample means of i.i.d. terms $Z_r^{(ij)} = \mathbf{1}[\mathbf{w}_r^T\tilde{\mathbf{x}}_i > 0]\,\mathbf{1}[\mathbf{w}_r^T\tilde{\mathbf{x}}_j > 0]\,\langle\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j\rangle$. Since $|\langle\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j\rangle| \leq [\tilde{\mathbf{G}}]_{ij} \leq \lambda_1^3$ and each indicator is bounded in $[0, 1]$, each centered matrix $\mathbf{S}_r$ satisfies $\|\mathbf{S}_r\|_2 \leq 2\lambda_1^3$ and matrix variance statistic $\|\frac{1}{m}\sum_r \mathbb{E}[\mathbf{S}_r\mathbf{S}_r^T]\|_2 \leq \lambda_1^6/(2m)$, using the operator-norm bound $\|\mathbb{E}[\mathbf{S}_1^2]\|_2 \leq \lambda_1^3\|\mathbf{K}_{\infty,\text{seq}}\|_2 \leq \frac{1}{2}\lambda_1^6$ from the proof of Theorem 4.4 in place of the looser trace bound. Applying the matrix Bernstein inequality (Tropp, 2012) to $\mathbf{K}_{m,\text{seq}} - \mathbf{K}_{\infty,\text{seq}} = \frac{1}{m}\sum_r \mathbf{S}_r$ gives the stated bound. This is identical to the bound derived in the proof of Theorem 4.4. The sequential-architecture structure introduces no additional deviation beyond the spectral amplification already captured by λ_1^3. □

D Formal Influence Malleability Theory

Definition D.1 (Formal Influence Malleability Measure)**.** Following the kernel-ridge view of Definition 3.3, take the influence vector to be the dual-coefficient vector $\boldsymbol{\mathcal{I}} = \boldsymbol{\alpha} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$, whose entry α_j weights training point j in the predictor $f = \sum_j \alpha_j K(\cdot, \mathbf{x}_j)$ and from which the leave-one-out influence I_j of Definition 3.3 is recovered through the same resolvent. The influence malleability measure is

$$\mu_M(f) = \mathbb{E}_{j,\,\delta\sim\Delta}\left[|\alpha_j(\mathbf{x}_j) - \alpha_j(\mathbf{x}_j + \delta)|\right],$$

the mean absolute change of these influence weights when a training input is perturbed by $\|\delta\|_p \leq \varepsilon$, with the index j drawn uniformly and Δ the perturbation distribution.

Theorem D.2 (Malleability-Kernel Eigenspectrum Connection)**.** *For the NTK regime with kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$ having eigendecomposition $\mathbf{K} = \sum_i \lambda_i \mathbf{v}_i\mathbf{v}_i^T$ with $\lambda_1 \geq \cdots \geq \lambda_n > 0$, regularization $\lambda > 0$, label vector $\mathbf{y}$, and perturbation budget $\varepsilon > 0$ with kernel Lipschitz constant L_K, letting $\boldsymbol{\alpha} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$, the following holds under the convention that μ_M is evaluated with Δ concentrating on the per-point adversarial direction (adversarial malleability μ_M^{adv}, which uses the same adversarial perturbation regime as the empirical Flip Rate since PGD maximises over the ε-ball):*

$$\begin{aligned}\mu_M(f) \;&\leq\; \frac{6\varepsilon L_K\|\boldsymbol{\alpha}\|_2}{n}\sum_{j=1}^{n}\|(\mathbf{K}+\lambda\mathbf{I})^{-1}_{j,:}\|_2 \;\leq\; \frac{6\varepsilon L_K\|\boldsymbol{\alpha}\|_2}{\sqrt{n}}\big\|(\mathbf{K}+\lambda\mathbf{I})^{-1}\big\|_F \\ &\leq\; \frac{6\varepsilon L_K\|\boldsymbol{\alpha}\|_2}{\lambda} \;\leq\; \frac{6\varepsilon L_K\|\mathbf{y}\|_2}{\lambda^2}.\end{aligned}$$

This is an upper bound and caps how malleable f can be, though a small right-hand side does not certify that μ_M is itself small. If instead a pointwise *lower-gradient condition holds,* $\min_j \|\partial(\mathbf{K}\boldsymbol{\alpha})/\partial \mathbf{x}_j\|_2 \geq \ell > 0$*, the lower bound* $\mu_M(f) \geq \varepsilon\ell\,\|(\mathbf{K}+\lambda\mathbf{I})^{-1}\|_F/n$ *follows with* ℓ *replacing* L_K *and the reverse estimate* $\sum_j \|(\mathbf{K}+\lambda\mathbf{I})^{-1}_{j,:}\|_2 \geq \|(\mathbf{K}+\lambda\mathbf{I})^{-1}\|_F$ *in place of the Cauchy–Schwarz step. Absent such a pointwise condition, no nontrivial lower bound on* μ_M *holds in general (a sparse-*$\boldsymbol{\alpha}$ *construction on orthonormal data with* $K(\mathbf{x},\mathbf{x}') = (\mathbf{x}^T\mathbf{x}')^2$ *gives* $\mu_M = \Theta(\varepsilon/n)$*, well below the* $\Theta(\varepsilon/\sqrt{n})$ *rate the chain above would suggest if misread as a lower bound).*

Proof. Let $\boldsymbol{\alpha} = (\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{y}$. For each training point j, let $\boldsymbol{\delta}_j^*$ be the perturbation of $\mathbf{x}_j$ that maximises $|\Delta\alpha_j|$, i.e., the change in the j-th component of $\boldsymbol{\alpha}$ when $\mathbf{x}_j \mapsto \mathbf{x}_j + \boldsymbol{\delta}_j$. By the chain rule, resolvent identity, and Cauchy–Schwarz,

$$|\Delta\alpha_j^*| = \varepsilon\left\|e_j^T(\mathbf{K}+\lambda\mathbf{I})^{-1}\,\nabla_{\mathbf{x}_j}K(\cdot,\mathbf{x}_j)\,\boldsymbol{\alpha}\right\|_2 \leq \varepsilon\,\|(\mathbf{K}+\lambda\mathbf{I})^{-1}_{j,:}\|_2 \cdot 6L_K\|\boldsymbol{\alpha}\|_2,$$

using $\|\nabla_{\mathbf{x}_j}K(\cdot,\mathbf{x}_j)\boldsymbol{\alpha}\|_2 \leq 6L_K\|\boldsymbol{\alpha}\|_2$, since a training-point perturbation moves $\mathbf{K} = \mathbf{G}^3$ through three product slots, giving $\|\Delta\mathbf{K}\|_2 \leq 3\|\mathbf{G}\|_2^2\|\Delta\mathbf{G}\|_2 = 6\lambda_1(\mathbf{G})^{5/2}\varepsilon$ with $\|\Delta\mathbf{G}\|_2 \leq 2\lambda_1(\mathbf{G})^{1/2}\varepsilon$ and $L_K = \lambda_1(\mathbf{G})^{5/2}$. Since $\mu_M(f) = \mathbb{E}_j[|\Delta\alpha_j^*|]$ (average over uniform draw of training index j):

$$\mu_M(f) = \frac{1}{n}\sum_{j=1}^n |\Delta\alpha_j^*| \leq \frac{6\varepsilon L_K\|\boldsymbol{\alpha}\|_2}{n}\sum_{j=1}^n \|(\mathbf{K}+\lambda\mathbf{I})^{-1}_{j,:}\|_2.$$

For non-negative $a_j = \|(\mathbf{K}+\lambda\mathbf{I})^{-1}_{j,:}\|_2$, Cauchy–Schwarz gives $\sum_j a_j \leq \sqrt{n}\sqrt{\sum_j a_j^2} = \sqrt{n}\,\|(\mathbf{K}+\lambda\mathbf{I})^{-1}\|_F$. The Frobenius norm satisfies $\|(\mathbf{K}+\lambda\mathbf{I})^{-1}\|_F^2 = \sum_i 1/(\lambda_i+\lambda)^2 \leq n/\lambda^2$ (each term is at most $1/\lambda^2$ since $\lambda_i \geq 0$), giving $\|(\mathbf{K}+\lambda\mathbf{I})^{-1}\|_F \leq \sqrt{n}/\lambda$. Chaining these:

$$\mu_M(f) \leq \frac{6\varepsilon L_K\|\boldsymbol{\alpha}\|_2}{n}\cdot\sqrt{n}\,\|(\mathbf{K}+\lambda\mathbf{I})^{-1}\|_F \leq \frac{6\varepsilon L_K\|\boldsymbol{\alpha}\|_2}{\sqrt{n}}\,\|(\mathbf{K}+\lambda\mathbf{I})^{-1}\|_F \leq \frac{6\varepsilon L_K\|\boldsymbol{\alpha}\|_2}{\lambda}.$$

Since $\|\boldsymbol{\alpha}\|_2 \leq \|\mathbf{y}\|_2/\lambda$ (as $\mathbf{K} \succeq 0$), the final bound is $\mu_M(f) \leq 6\varepsilon L_K\|\mathbf{y}\|_2/\lambda^2$. The pointwise-lower-bound clause follows by replacing the Cauchy–Schwarz step $\|\nabla_{\mathbf{x}_j}K(\cdot,\mathbf{x}_j)\boldsymbol{\alpha}\|_2 \leq 6L_K\|\boldsymbol{\alpha}\|_2$ with the assumed pointwise lower bound ℓ, which reverses each subsequent inequality. □

Remark D.3 (Tighter bound under isotropic label alignment). When the label vector has isotropic projection onto the kernel eigenbasis ($|\mathbf{v}_i^T\mathbf{y}|^2 = \|\mathbf{y}\|_2^2/n$ for all i, the typical/average-case assumption), the bound tightens to $\mu_M(f) \leq \frac{\varepsilon L_K\|\mathbf{y}\|_2}{n}\sum_i 1/(\lambda_i+\lambda)^2$. The worst-case bound above uses $\lambda\mathbf{I}$-only control on the resolvent and is distribution-free over label vectors with fixed $\|\mathbf{y}\|_2$.

Corollary D.4 (Regularization Controls Malleability)**.** *Both available bounds on malleability are upper bounds, and both are monotone decreasing in* λ*: the bound of Theorem D.2,* $\mu_M(f) \leq 6\varepsilon L_K\|\mathbf{y}\|_2/\lambda^2$*, and the bound of Theorem D.17,* $\mu_M(f) \leq 6\varepsilon L_K\|\mathbf{y}\|_2/\lambda^2$*, both diverge as* $\lambda \to 0$ *and converge to* 0 *as* $\lambda \to \infty$*, confirming that heavy regularization suppresses the* ceiling *on malleability. Neither bound, being one-sided, certifies that the malleability measure itself grows as* $\lambda \to 0$*; absent a pointwise lower-gradient condition (Theorem D.2), no such claim is established. The optimal* λ *balancing bias and malleability (from Theorem D.10) is data-dependent and need not equal the median eigenvalue in general.*

Proposition D.5 (Two-Sided Linearized Malleability Bound for the Cube Kernel)**.** *Let* $\mathbf{X} \in \mathbb{R}^{n\times d}$ *have unit-norm rows* $\mathbf{x}_j^\top$*, and let* $\mathbf{G} = \mathbf{X}\mathbf{X}^\top = \sum_{i=1}^r \lambda_i\mathbf{u}_i\mathbf{u}_i^\top$ *be the input Gram matrix with eigenvalues* $\lambda_1 \geq \cdots \geq \lambda_r > 0$*, orthonormal eigenvectors* $\mathbf{u}_i \in \mathbb{R}^n$*, and rank* $r = \operatorname{rank}\mathbf{X}$*. Let* $\mathbf{K} = \mathbf{G}^3$ *be the linearized-attention kernel matrix, fix a ridge* $\lambda > 0$ *and labels* $\mathbf{y} \in \mathbb{R}^n$*, and set* $\mathbf{R} = (\mathbf{K}+\lambda\mathbf{I})^{-1}$*,* $\boldsymbol{\alpha} = (\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{y} = \mathbf{R}\mathbf{y}$*, and* $L_K = \lambda_1^{5/2}$ *(the kernel Lipschitz constant of Theorem D.17). Write* $u_{i,j} = \langle\mathbf{u}_i,\mathbf{e}_j\rangle$*,* $a_1 = \langle\mathbf{u}_1,\boldsymbol{\alpha}\rangle = \langle\mathbf{u}_1,\mathbf{y}\rangle/(\lambda_1^3+\lambda)$*,* $\eta = |\langle\mathbf{u}_1,\mathbf{y}\rangle|/\|\mathbf{y}\|_2 \in [0,1]$*, and* $\kappa_d = \lambda_1/\lambda_r$*.*

For a perturbation $\mathbf{x}_j \mapsto \mathbf{x}_j + \boldsymbol{\delta}$*, let* $\Delta\alpha_j^{\mathrm{lin}}(\boldsymbol{\delta}) = -\mathbf{e}_j^\top\mathbf{R}\,(\mathrm{D}_j\mathbf{K}[\boldsymbol{\delta}])\,\boldsymbol{\alpha}$ *denote the first-order (linear-in-*$\boldsymbol{\delta}$*) change of the* j*-th influence coefficient, where* $\mathrm{D}_j\mathbf{K}[\boldsymbol{\delta}]$ *is the differential of* $\mathbf{K} = \mathbf{G}^3$ *in the direction that moves only row* j *of* $\mathbf{X}$ *by* $\boldsymbol{\delta}$*. Define the* linearized malleability

$$\mu_M^{\mathrm{lin}} = \frac{1}{n}\sum_{j=1}^n \max_{\|\boldsymbol{\delta}\|_2\leq\varepsilon}\left|\Delta\alpha_j^{\mathrm{lin}}(\boldsymbol{\delta})\right|,$$

which linearizes the adversarial malleability μ_M^{adv} of Theorem D.2 and coincides with it up to $O(\varepsilon^2)$ as $\varepsilon \to 0$ for fixed $(\mathbf{X}, \mathbf{y}, \lambda)$.

Assume the non-degeneracy condition

$$\langle \mathbf{u}_1, \mathbf{y} \rangle \neq 0, \qquad \text{and } \forall j: \ \operatorname{sign}(u_{1,j} B_j) = \operatorname{sign}(a_1) \ \text{ or } \ u_{1,j} B_j = 0, \tag{ND}$$

where $B_j = (\mathbf{G}^2 \boldsymbol{\alpha})_j + \lambda_1 (\mathbf{G}\boldsymbol{\alpha})_j + \lambda_1^2 \alpha_j$. Then

$$\frac{\varepsilon\, \eta\, \lambda_1^{5/2}\, \|\mathbf{y}\|_2}{(\lambda_1^3 + \lambda)^2} \ \le \ \mu_M^{\mathrm{lin}} \ \le \ \frac{6\, \varepsilon\, \lambda_1^{5/2}\, \|\mathbf{y}\|_2}{\lambda^2}. \tag{37}$$

Moreover the per-point influence gradient $\mathbf{g}_j = \mathbf{X}^\top \mathbf{h}_j$ (constructed in the proof), for which $\max_{\|\boldsymbol{\delta}\|_2 \le \varepsilon} |\Delta\alpha_j^{\mathrm{lin}}(\boldsymbol{\delta})| = \varepsilon \|\mathbf{g}_j\|_2$, satisfies the uniform lower bound

$$\|\mathbf{g}_j\|_2 \ \ge \ \ell \ := \ \frac{\eta\, \lambda_1^{5/2}\, \|\mathbf{y}\|_2}{(\lambda_1^3 + \lambda)^2} \qquad \text{for every } j, \tag{38}$$

so ℓ certifies the per-point floor $\max_{\|\boldsymbol{\delta}\|_2 \le \varepsilon} |\Delta\alpha_j^{\mathrm{lin}}| \ge \varepsilon \ell$ and hence $\mu_M^{\mathrm{lin}} \ge \varepsilon \ell$ directly, a stronger conclusion than the $\|\mathbf{R}\|_F/n$ form of the conditional clause of Theorem D.2.

Proof. Differential of the kernel. Perturbing row j by $\boldsymbol{\delta}$ moves the Gram matrix to first order by $\Delta\mathbf{G} = \mathbf{e}_j \mathbf{w}^\top + \mathbf{w} \mathbf{e}_j^\top$ with $\mathbf{w} = \mathbf{X}\boldsymbol{\delta}$, since only entries in row and column j change and the (j,j) entry changes by $2\mathbf{x}_j^\top \boldsymbol{\delta}$. The product rule for $\mathbf{K} = \mathbf{G}^3$ gives $\mathrm{D}_j\mathbf{K}[\boldsymbol{\delta}] = \Delta\mathbf{G}\, \mathbf{G}^2 + \mathbf{G}\, \Delta\mathbf{G}\, \mathbf{G} + \mathbf{G}^2\, \Delta\mathbf{G}$. Because $\|\Delta\mathbf{G}\|_2 \le 2\sqrt{\lambda_1}\, \|\boldsymbol{\delta}\|_2$ and $\|\mathbf{G}\|_2 = \lambda_1$, one has $\|\mathrm{D}_j\mathbf{K}[\boldsymbol{\delta}]\|_2 \le 3\lambda_1^2 \cdot 2\sqrt{\lambda_1}\, \|\boldsymbol{\delta}\|_2 = 6\lambda_1^{5/2} \|\boldsymbol{\delta}\|_2$.

Per-point gradient. With $\mathbf{r}_j = \mathbf{R}\mathbf{e}_j$, expanding $\Delta\alpha_j^{\mathrm{lin}}(\boldsymbol{\delta}) = -\mathbf{r}_j^\top (\mathrm{D}_j\mathbf{K}[\boldsymbol{\delta}])\boldsymbol{\alpha}$ and collecting the terms linear in $\boldsymbol{\delta}$ yields $\Delta\alpha_j^{\mathrm{lin}}(\boldsymbol{\delta}) = -\boldsymbol{\delta}^\top \mathbf{g}_j$ with $\mathbf{g}_j = \mathbf{X}^\top \mathbf{h}_j$ and

$$\mathbf{h}_j = R_{jj}\, \mathbf{G}^2\boldsymbol{\alpha} + (\mathbf{G}\mathbf{R})_{jj}\, \mathbf{G}\boldsymbol{\alpha} + (\mathbf{G}^2\mathbf{R})_{jj}\, \boldsymbol{\alpha} + (\mathbf{G}^2\boldsymbol{\alpha})_j\, \mathbf{r}_j + (\mathbf{G}\boldsymbol{\alpha})_j\, \mathbf{G}\mathbf{r}_j + \alpha_j\, \mathbf{G}^2\mathbf{r}_j,$$

so $\max_{\|\boldsymbol{\delta}\|_2 \le \varepsilon} |\Delta\alpha_j^{\mathrm{lin}}(\boldsymbol{\delta})| = \varepsilon \|\mathbf{g}_j\|_2$. Since $\mathbf{X}\mathbf{v}_1 = \sqrt{\lambda_1}\, \mathbf{u}_1$ for the top right singular vector $\mathbf{v}_1$ (unit norm), testing $\mathbf{g}_j = \mathbf{X}^\top \mathbf{h}_j$ against $\mathbf{v}_1$ gives $\|\mathbf{g}_j\|_2 \ge |\mathbf{v}_1^\top \mathbf{g}_j| = \sqrt{\lambda_1}\, |Q_j|$ with $Q_j = \langle \mathbf{u}_1, \mathbf{h}_j \rangle$. Using $\mathbf{G}^m \mathbf{u}_1 = \lambda_1^m \mathbf{u}_1$ and $\langle \mathbf{u}_1, \mathbf{r}_j \rangle = u_{1,j}/(\lambda_1^3 + \lambda)$,

$$Q_j = a_1 A_j + \frac{u_{1,j} B_j}{\lambda_1^3 + \lambda}, \qquad A_j = \lambda_1^2 R_{jj} + \lambda_1 (\mathbf{G}\mathbf{R})_{jj} + (\mathbf{G}^2\mathbf{R})_{jj}.$$

Uniform floor. Because $\mathbf{G}$ and $\mathbf{R}$ commute and are positive semidefinite, $\mathbf{G}\mathbf{R}$ and $\mathbf{G}^2\mathbf{R}$ are positive semidefinite, whence $(\mathbf{G}\mathbf{R})_{jj} \ge 0$ and $(\mathbf{G}^2\mathbf{R})_{jj} \ge 0$ and therefore $A_j \ge \lambda_1^2 R_{jj}$. Since $\lambda_{\max}(\mathbf{K}) = \lambda_1^3$, one has $\mathbf{K} + \lambda\mathbf{I} \preceq (\lambda_1^3 + \lambda)\mathbf{I}$, so $\mathbf{R} \succeq (\lambda_1^3 + \lambda)^{-1}\mathbf{I}$ and hence $R_{jj} \ge 1/(\lambda_1^3 + \lambda)$ *uniformly* in j, giving $A_j \ge \lambda_1^2/(\lambda_1^3 + \lambda)$. Under (ND) the two contributions to Q_j share the sign of a_1 (or the second vanishes), so $|Q_j| \ge |a_1| A_j \ge |a_1| \lambda_1^2/(\lambda_1^3 + \lambda)$. With $|a_1| = \eta \|\mathbf{y}\|_2/(\lambda_1^3 + \lambda)$ this yields, for every j,

$$\|\mathbf{g}_j\|_2 \ge \sqrt{\lambda_1}\, |Q_j| \ge \frac{\lambda_1^{5/2} |a_1|}{\lambda_1^3 + \lambda} = \frac{\eta\, \lambda_1^{5/2}\, \|\mathbf{y}\|_2}{(\lambda_1^3 + \lambda)^2} = \ell,$$

which is (38). Averaging $\max_{\|\boldsymbol{\delta}\|_2 \le \varepsilon} |\Delta\alpha_j^{\mathrm{lin}}| = \varepsilon \|\mathbf{g}_j\|_2 \ge \varepsilon\ell$ over j gives the lower bound in (37), with no factor $1/n$.

Ceiling. Writing $\Delta\boldsymbol{\alpha}^{\mathrm{lin}} = -\mathbf{R}\, (\mathrm{D}_j\mathbf{K}[\boldsymbol{\delta}])\, \boldsymbol{\alpha}$ and using $\|\mathbf{R}\|_2 \le 1/\lambda$, $\|\boldsymbol{\alpha}\|_2 \le \|\mathbf{y}\|_2/\lambda$, and the differential bound above,

$$|\Delta\alpha_j^{\mathrm{lin}}(\boldsymbol{\delta})| \le \|\Delta\boldsymbol{\alpha}^{\mathrm{lin}}\|_2 \le \|\mathbf{R}\|_2\, \|\mathrm{D}_j\mathbf{K}[\boldsymbol{\delta}]\|_2\, \|\boldsymbol{\alpha}\|_2 \le \frac{1}{\lambda} \cdot 6\lambda_1^{5/2} \varepsilon \cdot \frac{\|\mathbf{y}\|_2}{\lambda} = \frac{6\varepsilon \lambda_1^{5/2} \|\mathbf{y}\|_2}{\lambda^2}$$

for $\|\boldsymbol{\delta}\|_2 \le \varepsilon$, and averaging over j gives the upper bound in (37). □

Remark D.6 (Role of the Lipschitz constant and regime-qualified match). The Lipschitz constant $L_K = \lambda_1^{5/2}$ enters the floor in (37) as well as the ceiling, so the same spectral quantity that caps malleability also drives it upward. The bound is not claimed to be tight in λ_1 across all λ, since the floor carries $(\lambda_1^3 + \lambda)^{-2}$ while the ceiling carries λ^{-2}. In the regime $\lambda \gtrsim \lambda_1^3$, where $(\lambda_1^3 + \lambda)^2$ is comparable to λ^2, floor and ceiling agree in their λ_1-dependence up to the factor $6/\eta$. The ratio of ceiling to floor equals $6(\lambda_1^3 + \lambda)^2/(\eta\lambda^2) = 6(\kappa_d^3 + 1)^2/\eta$ at $\lambda = \lambda_r^3$, which scales as κ_d^6/η. Whether this reflects a genuine spread of μ_M^{lin} or residual slack in the resolvent and Cauchy–Schwarz steps remains unresolved, so the ratio is reported as slack of unknown tightness.

Remark D.7 (Small-ridge behavior is rank-dependent). For full-rank $\mathbf{G}$ ($r = n$), the matrix $\mathbf{K} = \mathbf{G}^3$ is invertible with $\|\mathbf{R}\|_2 \le 1/\lambda_r^3$ uniformly in λ, so the ceiling argument above gives $\mu_M^{\text{lin}} \le 6\varepsilon\lambda_1^{5/2}\|\mathbf{y}\|_2/\lambda_r^6$, which stays bounded as $\lambda \to 0^+$. This finiteness is a consequence of the upper bound and does not follow from the floor. In the regime $n > r$ studied in this work ($\mathbf{G}$ rank-deficient), the null space of $\mathbf{K}$ contributes $R_{jj} \ge \nu_j/\lambda$ with $\nu_j = \|\mathbf{P}_{\text{null}}\mathbf{e}_j\|_2^2$ and $\sum_j \nu_j = n - r > 0$, so at least one such term diverges and μ_M^{lin} diverges as $\lambda \to 0^+$ even for aligned labels, so no finite-malleability reading is available there.

Corollary D.8 (Two-Sided Bound for Top-Eigenvector-Aligned Labels)**.** *Assume* (ND) *together with* $\eta \ge \eta_0 > 0$*. Then*

$$\frac{\varepsilon\,\eta_0\,\lambda_1^{5/2}\,\|\mathbf{y}\|_2}{(\lambda_1^3+\lambda)^2} \;\le\; \mu_M^{\text{lin}} \;\le\; \frac{6\,\varepsilon\,\lambda_1^{5/2}\,\|\mathbf{y}\|_2}{\lambda^2}.$$

The exactly-aligned case $\mathbf{y} = c\,\mathbf{u}_1$ *(*$\eta = 1$*) satisfies* (ND) unconditionally*, since* $\boldsymbol{\alpha} = \beta\mathbf{u}_1$ *with* $\beta = c/(\lambda_1^3 + \lambda)$ *gives* $B_j = 3\lambda_1^2\beta u_{1,j}$*, so* $u_{1,j}B_j = 3\lambda_1^2\beta u_{1,j}^2$ *carries the sign of* $a_1 = \beta$*. The per-point contribution then sharpens to the coefficient* $1 + 3u_{1,j}^2$*, in which the uniform* A*-term supplies the* 1 *and the aligned* B*-term supplies the* $3u_{1,j}^2$*, whose average over* j *(using* $\sum_j u_{1,j}^2 = 1$*) gives the refined floor*

$$\mu_M^{\text{lin}} \;\ge\; \left(1 + \tfrac{3}{n}\right)\frac{\varepsilon\,\lambda_1^{5/2}\,\|\mathbf{y}\|_2}{(\lambda_1^3+\lambda)^2}.$$

Condition (ND) *is a joint condition on* $(\mathbf{X}, \mathbf{y}, \lambda)$*, and near-alignment* $\eta \to 1$ *alone does not imply it, since at small ridge the bulk contribution to* B_j *is amplified by* $1/(\lambda_r^3 + \lambda)$*. It is guaranteed only in the exactly-aligned case* $\mathbf{y} \propto \mathbf{u}_1$*, so the two-sided bound holds for aligned-label data, and generic labels or spiked inputs alone do not guarantee it.*

Theorem D.9 (Architecture-Dependent Malleability Gap via Sensitivity)**.** *Let* $K_{\textit{LinAttn}}$ *be the Gram-induced linearized attention kernel and* $K_{\textit{poly}}$ *a standard degree-*p *polynomial kernel. The intrinsic influence sensitivities (Definition D.18) satisfy:*

$$\frac{\mathcal{S}_{\text{att}}}{\mathcal{S}_{\text{poly}}} = \Theta\left(\frac{\lambda_1(\mathbf{G})^{5/2}}{\sqrt{n}}\right),$$

using the exact kernel Lipschitz constant $L_K^{\text{att}} = \lambda_1(\mathbf{G})^{5/2}$ *(Theorem D.17), attained along the top right singular vector of* $\mathbf{X}$*; since the numerator is attained exactly, this is an equality of orders for the worst-case perturbation direction, not a one-sided estimate. Since Theorem D.17 gives* $\mu_M(f) \le 6\varepsilon\mathcal{S}(f)$ *for any architecture, the ratio of the sensitivity-derived malleability upper bounds satisfies:*

$$\frac{\mu_M^{\text{UB}}(f_{\text{att}})}{\mu_M^{\text{UB}}(f_{\text{poly}})} \;=\; \frac{6\varepsilon\mathcal{S}_{\text{att}}}{6\varepsilon\mathcal{S}_{\text{poly}}} \;=\; O\left(\frac{\lambda_1(\mathbf{G})^{5/2}}{\sqrt{n}}\right).$$

Proof. By the proof of Theorem D.17, the exact kernel Lipschitz constant of linearized attention is $L_K^{\text{att}} = \lambda_1(\mathbf{G})^{5/2}$, attained exactly along the top right singular vector of $\mathbf{X}$; the polynomial-kernel constant is $L_K^{\text{poly}} = O(\sqrt{n})$ (Proposition D.19). Substituting into Definition D.18 with $\|\mathbf{y}\|_2$ and λ identical for both architectures gives $\mathcal{S}_{\text{att}}/\mathcal{S}_{\text{poly}} = L_K^{\text{att}}/L_K^{\text{poly}} = \Theta(\lambda_1(\mathbf{G})^{5/2}/\sqrt{n})$; because the numerator is exact, the ratio of sensitivities is an equality of orders, not a one-sided bound. The malleability upper bound $\mu_M^{\text{UB}}(f) \coloneqq 6\varepsilon\mathcal{S}(f)$ follows from Theorem D.17. Taking the ratio cancels $6\varepsilon\|\mathbf{y}\|_2/\lambda^2$ from numerator and denominator, leaving only the L_K ratio. This last step bounds only the ratio of malleability *upper bounds*, so it does not by itself establish, and should not be read as predicting, any particular empirical malleability ratio. □

Theorem D.10 (Influence-Robustness of Test Error). *Let $f(\mathbf{x}) = \mathbf{k}(\mathbf{x})^T(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$ be a KRR predictor with $\|\mathbf{k}(\mathbf{x})\|_2 \le B$ for all $\mathbf{x}$; B is carried explicitly, not normalized to 1, since the attention kernel does not satisfy $\|\mathbf{k}(\mathbf{x})\|_2 \le 1$ (Proposition C.24 and the proof of Theorem D.17 give $\sup_{\mathbf{x}} \|\mathbf{k}(\mathbf{x})\|_2$ up to $\lambda_1(\mathbf{G})^{5/2}$). Let f' be the predictor after training points are perturbed, inducing $\Delta\boldsymbol{\mathcal{I}} = \boldsymbol{\mathcal{I}}' - \boldsymbol{\mathcal{I}}$, define the* root-mean-square influence malleability $\bar{\mu}_2(f)^2 = \mathbb{E}_\delta\big[\|\Delta\boldsymbol{\mathcal{I}}(\delta)\|_2^2\big]/n$, *and write* $\mathrm{TE}(g) = \mathbb{E}_{\mathbf{x}_{\mathrm{test}}}[(g(\mathbf{x}_{\mathrm{test}}) - f^*(\mathbf{x}_{\mathrm{test}}))^2]$ *for the test error of predictor g. Then:*

$$\big|\mathrm{TE}(f') - \mathrm{TE}(f)\big| \;\le\; 2B\sqrt{n\,\mathrm{TE}(f)}\;\bar{\mu}_2(f) \;+\; B^2 n\,\bar{\mu}_2(f)^2.$$

The perturbed-data test error is thus controlled by the clean-data error $\mathrm{TE}(f)$ plus a term that grows with malleability: low malleability implies robust test error. As before, $\mu_M(f) \le \bar{\mu}_M(f) \le \mu_M^{\max}(f)$ (per-entry ℓ_2 malleability $\bar{\mu}_M(f) := \mathbb{E}_\delta[\|\Delta\boldsymbol{\mathcal{I}}(\delta)\|_2/\sqrt{n}]$), and by Jensen's inequality $\bar{\mu}_M(f) \le \bar{\mu}_2(f)$.

Proof. For a test point, $f'(\mathbf{x}_{\mathrm{test}}) - f(\mathbf{x}_{\mathrm{test}}) = \mathbf{k}(\mathbf{x}_{\mathrm{test}})^T\Delta\boldsymbol{\mathcal{I}}$, so pointwise $|f'(\mathbf{x}_{\mathrm{test}}) - f(\mathbf{x}_{\mathrm{test}})| \le \|\mathbf{k}(\mathbf{x}_{\mathrm{test}})\|_2\|\Delta\boldsymbol{\mathcal{I}}\|_2 \le B\|\Delta\boldsymbol{\mathcal{I}}\|_2$. Expanding

$$\mathrm{TE}(f') = \mathbb{E}\big[(f - f^* + (f' - f))^2\big] = \mathrm{TE}(f) + 2\,\mathbb{E}[(f - f^*)(f' - f)] + \mathbb{E}[(f' - f)^2],$$

Cauchy–Schwarz gives $|\mathbb{E}[(f - f^*)(f' - f)]| \le \sqrt{\mathrm{TE}(f)}\,\sqrt{\mathbb{E}[(f' - f)^2]}$, and the pointwise bound above gives $\mathbb{E}[(f' - f)^2] \le B^2\,\mathbb{E}\|\Delta\boldsymbol{\mathcal{I}}\|_2^2 = B^2 n\,\bar{\mu}_2(f)^2$. Substituting into the expansion and applying the triangle inequality to $|\mathrm{TE}(f') - \mathrm{TE}(f)|$ yields the stated bound.

Ordering $\mu_M \le \bar{\mu}_M$. By the norm inequality $\|\mathbf{v}\|_1 \le \sqrt{n}\|\mathbf{v}\|_2$: $\mu_M = \mathbb{E}[\|\Delta\boldsymbol{\mathcal{I}}\|_1]/n \le \mathbb{E}[\|\Delta\boldsymbol{\mathcal{I}}\|_2]/\sqrt{n} = \bar{\mu}_M$.

Ordering $\bar{\mu}_M \le \mu_M^{\max}$. By the norm inequality $\|\mathbf{v}\|_2 \le \sqrt{n}\|\mathbf{v}\|_\infty$: $\bar{\mu}_M = \mathbb{E}[\|\Delta\boldsymbol{\mathcal{I}}\|_2/\sqrt{n}] \le \mathbb{E}[\|\Delta\boldsymbol{\mathcal{I}}\|_\infty] = \mu_M^{\max}$.

Ordering $\bar{\mu}_M \le \bar{\mu}_2$. By Jensen's inequality applied to the convex map $t \mapsto t^2$: $\bar{\mu}_M(f) = \mathbb{E}_\delta[\|\Delta\boldsymbol{\mathcal{I}}\|_2/\sqrt{n}] \le \sqrt{\mathbb{E}_\delta[\|\Delta\boldsymbol{\mathcal{I}}\|_2^2]/n} = \bar{\mu}_2(f)$. □

Remark D.11 (Formal-Empirical Bridge: $\bar{\mu}_M$ vs. Flip Rate). Theorem D.10 defines $\bar{\mu}_M$ as an expectation over a smooth perturbation distribution on the ε-ball. The empirical Flip Rate (Definition 3.4) is measured under adversarial PGD perturbations, which *maximize* influence change over the same ball. A flip of point i is a sign change of its influence, which since I_i is a resolvent functional of the same influence vector $\boldsymbol{\mathcal{I}} = \boldsymbol{\alpha}$ (Definition 3.3) cannot occur unless $\boldsymbol{\alpha}$ moves, and requires a shift comparable to $|\alpha_i|$ itself. Since PGD maximises the influence shift over the ε-ball, the empirical flip probability dominates the smooth-distribution probability of the same sign-crossing event,

$$\mathrm{FlipRate}_\varepsilon^{\mathrm{PGD}} \;\ge\; \Pr_{\boldsymbol{\delta}\sim\mathrm{Unif}(\mathcal{B}_\varepsilon)}\big[\,\exists\, i \in \mathcal{H}_\tau :\; \mathrm{sign}(I_i(\mathbf{x}_i + \boldsymbol{\delta})) \ne \mathrm{sign}(I_i)\,\big].$$

Consequently, the ordering $\mu_M \le \bar{\mu}_M \le \mu_M^{\max}$ (which holds for smooth distributions) makes high $\bar{\mu}_M$ a *sufficient* driver of high PGD flip rates, not a necessary one, since a large $|\Delta\alpha_i|$ need not cross zero. The empirically observed gap between MLP-Attn and 2L-ReLU (Table 1) is explained by the higher intrinsic sensitivity $\mathcal{S}_{\mathrm{att}}$ driving higher $\bar{\mu}_M$, making adversarial manipulation structurally easier.

Corollary D.12 (Optimal Malleability). *Approximating $\mathrm{TE}(f)$ in Theorem D.10 by $\mathrm{Bias}^2(\mu)$, there exists an optimal root-mean-square malleability level $\bar{\mu}_2^*$ minimizing the resulting test-error robustness bound:*

$$\bar{\mu}_2^* = \arg\min_{\mu}\left[\mathrm{Bias}^2(\mu) + 2B\sqrt{n\,\mathrm{Bias}^2(\mu)}\;\mu + B^2 n\,\mu^2\right]$$

where the bias term depends on the kernel's spectral structure and $B = \sup_{\mathbf{x}}\|\mathbf{k}(\mathbf{x})\|_2$.

The bounds above control malleability through the kernel Lipschitz constant and the ridge λ. The conditioning of the influence operator itself carries a floor set by the data alone, and that floor is the same for the parameter-free proxy and for genuinely trainable attention.

Proposition D.13 (Malleability Floor for the Limiting Kernel). *Let $\mathbf{K}$ be either the parameter-free proxy kernel ($\mathbf{K} = \mathbf{G}^3$) or the limiting trainable kernel $\mathbf{K}_\infty^{\mathrm{QKV}}$ under either initialization scheme of*

Definition 4.8, and let $T_\lambda = (\mathbf{K} + \lambda \mathbf{I})^{-1}$ be the influence operator of Definition 3.3. Over the range of $\mathbf{K}$, the influence sensitivities $\{1/(\lambda_i(\mathbf{K}) + \lambda)\}_{i=1}^r$ have ridgeless dynamic range

$$\lim_{\lambda \to 0^+} \frac{\max_i 1/(\lambda_i(\mathbf{K}) + \lambda)}{\min_i 1/(\lambda_i(\mathbf{K}) + \lambda)} \;=\; \kappa^+(\mathbf{K}) \;\geq\; \kappa_d(\mathbf{G}), \tag{39}$$

with $\kappa^+(\mathbf{K}) = \kappa_d(\mathbf{G})$ under independent initialization and $\kappa^+(\mathbf{K}) = \kappa_d(\mathbf{G})^3$ for the proxy. The floor $\kappa_d(\mathbf{G})$ holds for each of these limiting kernels, with equality attained under independent initialization. The exponent p in $\kappa^+(\mathbf{K}) = \kappa_d(\mathbf{G})^p$ is the weight-dependent quantity ($p = 1$ under independent initialization, $p = 3$ for the proxy, and $p \in [2, 3)$ under aligned orthogonal initialization when the checkable condition $R > \kappa_d(\mathbf{G})^2$ of Proposition 4.11 holds).

Proof. The nonzero eigenvalues of $\mathbf{K}$ are strictly increasing functions of those of $\mathbf{G}$. Here $\mathbf{K} = \mathbf{G}^3$ for the proxy, while $\lambda_i(\mathbf{K}) = \lambda_i^3 + \tau_0 \lambda_i$ under aligned orthogonal initialization and $\lambda_i(\mathbf{K}) = c_{\mathrm{I}} \lambda_i$ under independent initialization (Theorem C.9). Hence $\mathbf{K}$ and $\mathbf{G}$ share an eigenbasis, and the displayed limit equals $\lambda_1(\mathbf{K})/\lambda_r(\mathbf{K}) = \kappa^+(\mathbf{K})$. The inequality $\kappa^+(\mathbf{K}) \geq \kappa_d(\mathbf{G})$ is Theorem 4.9. The exponent values follow from Theorem C.9 and Proposition 4.11. □

Remark D.14 (Reading of the floor). The ceiling on μ_M (Theorem D.2) and the influence-stability bound (Theorem C.23) both degrade as the influence operator becomes ill-conditioned. Proposition D.13 isolates the part of that ill-conditioning the data fixes: the dynamic range of the influence sensitivities of the limiting kernel cannot fall below $\kappa_d(\mathbf{G})$ under either initialization scheme. The malleability measured for the proxy (Table 1) therefore reflects a conditioning floor present in the trainable kernel as well, the two differing only in the exponent on $\kappa_d(\mathbf{G})$. In keeping with the one-sided nature of the bounds above, this establishes a shared lower bound on conditioning and does not predict a particular empirical malleability value for the trainable case.

Proposition D.15 (Sharp Conditional Bias Reduction via Data-Dependent Kernel)**.** *Let f^* be the target function, $\mathbf{G} = \sum_i \lambda_i \mathbf{u}_i \mathbf{u}_i^T$ the Gram matrix with kernel-matrix eigenvectors $\mathbf{u}_i \in \mathbb{R}^n$ (the left singular vectors of $\mathbf{X}$, not the right singular vectors), $w_i = \langle f^*, \mathbf{u}_i \rangle$, and $K_{\text{poly}}(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}')^3$ the degree-3 polynomial kernel, whose kernel matrix is the Hadamard cube $\mathbf{G}^{\circ 3}$. For kernel matrix $\mathbf{K}$ and regularization $\lambda > 0$, define the* shrinkage bias $\mathrm{Bias}^2(K) = \lambda^2 \|(\mathbf{K} + \lambda \mathbf{I})^{-1} f^*\|^2$ *and the poly-kernel Rayleigh quotient $\rho := f^{*T} \mathbf{G}^{\circ 3} f^* / \|f^*\|^2$. If*

$$\sum_i \frac{w_i^2}{(\lambda_i^3 + \lambda)^2} \;\leq\; \frac{\|f^*\|^2}{(\rho + \lambda)^2} \tag{S}$$

then $\mathrm{Bias}^2(K_{\text{LinAttn}}) \leq \mathrm{Bias}^2(K_{\text{poly}})$. Condition (S) holds in particular whenever f^* is supported on Gram modes with $\lambda_i^3 \geq \rho$, up to a residual of mass $\delta \leq \lambda^2/(\rho + \lambda)^2$ placed arbitrarily. The condition is sharp: on commuting (e.g., group-structured) data, where $\mathbf{G}$ and $\mathbf{G}^{\circ 3}$ share an eigenbasis, it reduces to the exact per-mode criterion $\lambda_k^3 \geq \mu_k$. For every target $f^* \in \ker \mathbf{G} \setminus \ker \mathbf{G}^{\circ 3}$ (a subspace of dimension $n - d$ whenever $n > \mathrm{rank}(\mathbf{X}) = d$, the regime studied here) the inequality strictly reverses instead, with ratio up to $(\lambda_{\min}(\mathbf{G}^{\circ 3}) + \lambda)^2/\lambda^2$; the claim holds directionally, conditional on (S).

Proof. The kernel matrix of K_{LinAttn} on training data is $\mathbf{G}^3$ (matrix cube of the Gram matrix $\mathbf{G} = \mathbf{X}\mathbf{X}^T$, Theorem 4.2), diagonal in $\{\mathbf{u}_i\}$ with eigenpairs $(\lambda_i^3, \mathbf{u}_i)$, so $\mathrm{Bias}^2(K_{\text{LinAttn}}) = \lambda^2 \sum_i w_i^2/(\lambda_i^3 + \lambda)^2$; condition (S) states exactly that this quantity is at most $\lambda^2 \|f^*\|^2/(\rho + \lambda)^2$. The degree-3 polynomial kernel matrix is $\mathbf{G}^{\circ 3}$ (Hadamard/element-wise cube), since $[K_{\text{poly}}]_{ij} = G_{ij}^3$, and generically does *not* share an eigenbasis with $\mathbf{G}^3$.

For K_{poly}, write f^* in the eigenbasis $\{\boldsymbol{\phi}_k\}$ of $\mathbf{G}^{\circ 3}$ with spectral weights $p_k = \tilde{w}_k^2/\|f^*\|^2$ ($\tilde{w}_k = \langle f^*, \boldsymbol{\phi}_k \rangle$) and eigenvalues μ_k, so that $\mathbb{E}_{k \sim p}[\mu_k] = f^{*T} \mathbf{G}^{\circ 3} f^*/\|f^*\|^2 = \rho$. Convexity of $x \mapsto (x + \lambda)^{-2}$ on $x \geq 0$ together with Jensen's inequality gives

$$\mathrm{Bias}^2(K_{\text{poly}}) = \lambda^2 \|f^*\|^2 \, \mathbb{E}_{k \sim p}\big[(\mu_k + \lambda)^{-2}\big] \;\geq\; \frac{\lambda^2 \|f^*\|^2}{\big(\mathbb{E}_{k \sim p}[\mu_k] + \lambda\big)^2} \;=\; \frac{\lambda^2 \|f^*\|^2}{(\rho + \lambda)^2}\,.$$

Chaining with condition (S) gives $\mathrm{Bias}^2(K_{\text{LinAttn}}) \leq \lambda^2 \|f^*\|^2/(\rho + \lambda)^2 \leq \mathrm{Bias}^2(K_{\text{poly}})$.

On commuting data, the shared eigenbasis makes Jensen's inequality an equality mode-by-mode, so the criterion reduces exactly to $\lambda_k^3 \geq \mu_k$ per occupied mode, establishing sharpness: the constant in condition (S) cannot be improved. For the reversal claim, when $n > \operatorname{rank}(\mathbf{X}) = d$, the linear dependence defining $\ker \mathbf{G}$ need not annihilate the cubic features $\mathbf{x}_i^{\otimes 3}$, so generically $\ker \mathbf{G} \not\subseteq \ker \mathbf{G}^{\circ 3}$ and $\mathbf{G}^{\circ 3}$ remains positive definite on this $(n-d)$-dimensional subspace. Any target f^* in this subspace carries all of its energy on the null eigenvalue of $\mathbf{G}^3$, giving $\operatorname{Bias}^2(K_{\text{LinAttn}}) = \lambda^2 \|f^*\|^2/\lambda^2 = \|f^*\|^2$, the maximum possible value, while $\operatorname{Bias}^2(K_{\text{poly}}) \leq \lambda^2 \|f^*\|^2/(\lambda_{\min}(\mathbf{G}^{\circ 3}) + \lambda)^2 < \|f^*\|^2$, giving a reversal by the stated ratio. Condition (S) is the weakest sound sufficient condition depending only on $(w_i, \{\lambda_i\}, \lambda, \rho)$: no condition restricted to this information class can certify a strictly larger set of targets without also certifying some target on which the inequality reverses. □

Remark D.16 (Completing the Dual Implications). Proposition D.15 combined with the Bias-Variance-Malleability Decomposition provides the theoretical basis for the dual implications of NTK non-convergence. The data-dependent kernel of linearized attention achieves *lower bias* (better task alignment) at the cost of *higher malleability* (sensitivity to perturbations). The net effect on generalization depends on data quality. When training data is clean, the bias reduction dominates and favors attention. When training data contains adversarial examples, the malleability cost dominates and favors rigid architectures. The Optimal Malleability corollary formalizes this tradeoff.

Theorem D.17 (Unified Sensitivity Bound for Predictions and Influence)**.** *Suppose $f(\mathbf{x}) = \mathbf{k}(\mathbf{x})^T(\mathbf{K} + \lambda \mathbf{I})^{-1}\mathbf{y}$ is a kernel ridge regression predictor with Gram matrix $\mathbf{K} \succeq 0$, regularization $\lambda > 0$, and kernel Lipschitz constant $L_K = \sup_{\|\boldsymbol{\delta}\|_2 \leq 1} \|\mathbf{k}(\mathbf{x}+\boldsymbol{\delta}) - \mathbf{k}(\mathbf{x})\|_2$. Let $\epsilon > 0$ be a perturbation budget. Then the prediction sensitivity satisfies*

$$\sup_{\|\boldsymbol{\delta}\|_2 \leq \epsilon} |f(\mathbf{x}+\boldsymbol{\delta}) - f(\mathbf{x})| \leq \frac{\epsilon L_K \|\mathbf{y}\|_2}{\lambda},$$

and perturbing any training point $\mathbf{x}_i$ by $\|\boldsymbol{\delta}\|_2 \leq \epsilon$ changes the influence vector $\boldsymbol{\mathcal{I}} = (\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{y}$ by at most

$$\|\Delta \boldsymbol{\mathcal{I}}\|_\infty \leq \frac{6\epsilon L_K \|\mathbf{y}\|_2}{\lambda^2}.$$

Both bounds scale as L_K/λ. For linearized attention, the exact kernel Lipschitz constant is $L_K^{\text{att}} = \lambda_1(\mathbf{G})^{5/2}$, attained along the top right singular vector of $\mathbf{X}$ (derived below). For standard ReLU, each of the n kernel entries $K_{\text{relu}}(\mathbf{x}+\boldsymbol{\delta}, \mathbf{x}_j)$ changes by $O(\|\boldsymbol{\delta}\|_2) = O(1)$, so $\|\Delta \mathbf{k}\|_2 = O(\sqrt{n})$ and $L_K^{\text{relu}} = O(\sqrt{n})$. The architecture-dependent gap in L_K simultaneously drives higher prediction sensitivity and higher influence instability for attention.

Proof. Write $\boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{y}$. Since $\mathbf{K} \succeq 0$, the operator norm satisfies $\|(\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\|_2 \leq 1/\lambda$, so $\|\boldsymbol{\alpha}\|_2 \leq \|\mathbf{y}\|_2/\lambda$. For the prediction bound, $f(\mathbf{x}+\boldsymbol{\delta}) - f(\mathbf{x}) = [\mathbf{k}(\mathbf{x}+\boldsymbol{\delta}) - \mathbf{k}(\mathbf{x})]^T \boldsymbol{\alpha}$, whence Cauchy–Schwarz gives $|f(\mathbf{x}+\boldsymbol{\delta}) - f(\mathbf{x})| \leq \epsilon L_K \|\boldsymbol{\alpha}\|_2 \leq \epsilon L_K \|\mathbf{y}\|_2/\lambda$.

For the influence bound, perturbing $\mathbf{x}_i$ to $\mathbf{x}_i + \boldsymbol{\delta}$ yields a perturbed Gram matrix $\mathbf{K}' \succeq 0$ (as $\mathbf{K}'$ is itself a valid kernel matrix), with $\|\Delta \mathbf{K}\|_2 \leq 2\epsilon L_K$ in general, or, for the transductive attention kernel, $\|\Delta \mathbf{K}\|_2 = \|\Delta(\mathbf{G}^3)\|_2 \leq 6\lambda_1(\mathbf{G})^{5/2}\epsilon$, since a training-point perturbation gives $\|\Delta \mathbf{G}\|_2 \leq 2\lambda_1(\mathbf{G})^{1/2}\epsilon$ and hence $\|\Delta(\mathbf{G}^3)\|_2 \leq 3\|\mathbf{G}\|_2^2 \|\Delta \mathbf{G}\|_2 = 6\lambda_1(\mathbf{G})^{5/2}\epsilon$ (the perturbation moves every entry of $\mathbf{K} = \mathbf{G}^3$, beyond row and column i). By the resolvent identity, $\Delta \boldsymbol{\mathcal{I}} = -(\mathbf{K}' + \lambda \mathbf{I}_n)^{-1} \Delta \mathbf{K}\, \boldsymbol{\alpha}$. Submultiplicativity gives $\|\Delta \boldsymbol{\mathcal{I}}\|_2 \leq (1/\lambda)(6\epsilon L_K)(\|\mathbf{y}\|_2/\lambda)$ for the attention kernel, where $\|(\mathbf{K}' + \lambda \mathbf{I}_n)^{-1}\|_2 \leq 1/\lambda$ since $\mathbf{K}' \succeq 0$ and $\|\Delta \mathbf{K}\|_2 \leq 6\lambda_1(\mathbf{G})^{5/2}\epsilon = 6\epsilon L_K$ with $L_K = \lambda_1(\mathbf{G})^{5/2}$, whereas the generic rank-two argument, which confines the perturbation to row and column i, would give the smaller constant $2\epsilon L_K$ that does not apply here, since the transductive kernel $\mathbf{K} = \mathbf{G}^3$ moves every entry under a training-point perturbation. The entry-wise bound follows from $\|\Delta \boldsymbol{\mathcal{I}}\|_\infty \leq \|\Delta \boldsymbol{\mathcal{I}}\|_2$.

Exact value of L_K^{att}. For a test-point perturbation $\mathbf{x} \mapsto \mathbf{x} + \boldsymbol{\delta}$ with the training set $\mathbf{X}$ fixed, the linearized-attention kernel vector is $\mathbf{k}(\mathbf{x}) = \mathbf{X}\mathbf{M}^2\mathbf{x}$ with $\mathbf{M} := \mathbf{X}^T\mathbf{X}$ (Proposition C.24). Writing the compact SVD $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, $\mathbf{M} = \mathbf{V}\boldsymbol{\Sigma}^2\mathbf{V}^T$ and $\mathbf{M}^2 = \mathbf{V}\boldsymbol{\Sigma}^4\mathbf{V}^T$, so $\mathbf{X}\mathbf{M}^2 = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T\mathbf{V}\boldsymbol{\Sigma}^4\mathbf{V}^T = \mathbf{U}\boldsymbol{\Sigma}^5\mathbf{V}^T$, giving $L_K^{\text{att}} = \|\mathbf{X}\mathbf{M}^2\|_2 = \sigma_1(\mathbf{X})^5 = \lambda_1(\mathbf{G})^{5/2}$, attained exactly along the top right singular vector of $\mathbf{X}$. For the two-layer ReLU arc-cosine NTK, each of the n kernel entries is $O(1)$-Lipschitz on the sphere, giving $L_K^{\text{relu}} = O(\sqrt{n})$ in quadrature. □

Definition D.18 (Intrinsic Influence Sensitivity)**.** For a kernel ridge regression predictor with Gram matrix $\mathbf{K} \succeq 0$, regularization $\lambda > 0$, labels $\mathbf{y}$, and kernel Lipschitz constant L_K, the *intrinsic influence sensitivity* is

$$\mathcal{S}(\mathbf{K}, \lambda, \mathbf{y}) \;=\; \frac{L_K \, \|\mathbf{y}\|_2}{\lambda^2}\,.$$

This quantity upper-bounds the per-unit-perturbation change in the influence vector $\boldsymbol{\mathcal{I}} = (\mathbf{K} + \lambda \mathbf{I}_n)^{-1}\mathbf{y}$ up to the absolute constant 6 (Theorem D.17), independent of any evaluation protocol (threshold τ, perturbation budget ε, or attack type).

Proposition D.19 (Architecture-Dependent Sensitivity Gap)**.** *Under the conditions of Theorem D.17 and Proposition C.24, the intrinsic influence sensitivity satisfies:*

1. ***Linearized attention:*** $\mathcal{S}_{\text{att}} = \Theta\big(\lambda_1(\mathbf{G})^{5/2}\,\|\mathbf{y}\|_2 \,/\, \lambda^2\big)$*, using the exact kernel Lipschitz constant* $L_K^{\text{att}} = \lambda_1(\mathbf{G})^{5/2}$ *established in the proof of Theorem D.17.*
2. ***Two-layer ReLU:*** $\mathcal{S}_{\text{relu}} = O\big(\sqrt{n}\,\|\mathbf{y}\|_2 \,/\, \lambda^2\big)$*, since each of the* n *kernel entries changes by* $O(1)$*, so* $L_K^{\text{relu}} = O(\sqrt{n})$*.*

Consequently, the sensitivity gap scales as

$$\frac{\mathcal{S}_{\text{att}}}{\mathcal{S}_{\text{relu}}} \;=\; \Omega\!\left(\frac{\lambda_1(\mathbf{G})^{5/2}}{\sqrt{n}}\right),$$

and, since L_K^{att} *is attained exactly, this order is tight for the worst-case perturbation direction. For natural-image data with a dominant mean component,* $\lambda_1(\mathbf{G}) = \Theta(n)$*, so the gap is* $\Theta(n^2)$*, exceeding the value* $O(\sqrt{n}\,\lambda_1(\mathbf{G})) = \Theta(n^{3/2})$ *obtained from the earlier, incorrect derivation, so attention is strictly more sensitive than ReLU whenever* $\lambda_1(\mathbf{G}) \gg n^{1/5}$*.*

Proof. Substitute the kernel Lipschitz constants from Proposition C.24 into Definition D.18. For linearized attention, the test-point kernel vector is $\mathbf{k}(\mathbf{x}) = \mathbf{X}\mathbf{M}^2\mathbf{x}$ with $\mathbf{M} = \mathbf{X}^T\mathbf{X}$. Via the compact SVD $\mathbf{X} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$, $\mathbf{X}\mathbf{M}^2 = \mathbf{U}\boldsymbol{\Sigma}^5\mathbf{V}^T$, so $L_K^{\text{att}} = \|\mathbf{X}\mathbf{M}^2\|_2 = \sigma_1(\mathbf{X})^5 = \lambda_1(\mathbf{G})^{5/2}$ exactly (Theorem D.17). For the standard ReLU NTK, each entry satisfies $|K_{\text{relu}}(\mathbf{x}+\boldsymbol{\delta}, \mathbf{x}_j) - K_{\text{relu}}(\mathbf{x}, \mathbf{x}_j)| \le L_0\|\boldsymbol{\delta}\|_2$ where $L_0 = O(1)$ (the kernel gradient depends only on the local input pair and is bounded by $\|\mathbf{x}_j\|_2 = 1$). Summing n such $O(1)$ entries in quadrature gives $L_K^{\text{relu}} = \|\Delta\mathbf{k}\|_2/\epsilon = O(\sqrt{n})$. The ratio $\mathcal{S}_{\text{att}}/\mathcal{S}_{\text{relu}} = L_K^{\text{att}}/L_K^{\text{relu}}$ cancels the common $\|\mathbf{y}\|_2/\lambda^2$ factor. Since $L_K^{\text{att}} = \lambda_1(\mathbf{G})^{5/2}$ is exact while $L_K^{\text{relu}} = O(\sqrt{n})$ is an upper bound, the ratio is a lower bound $\Omega(\lambda_1(\mathbf{G})^{5/2}/\sqrt{n})$, and the threshold $\lambda_1(\mathbf{G}) \gg n^{1/5}$ follows from $\lambda_1(\mathbf{G})^{5/2}/\sqrt{n} \gg 1$. □

Remark D.20 (Intrinsic Sensitivity as a Diagnostic). The quantity $\mathcal{S}(\mathbf{K}, \lambda, \mathbf{y})$ provides a parameter-free characterization of influence malleability that does not depend on the evaluation protocol (choice of τ, ε, or attack). Its dependence on λ is analogous to robustness margins in adversarial learning, since the regularization strength controls the tradeoff between expressiveness and stability. Proposition D.19 shows that the architecture-dependent factor $n\,\|\mathbf{G}\|_2$ is the intrinsic source of the malleability gap, formalizing the empirical observation that attention's data-dependent kernel couples all training points and amplifies perturbation effects.

Proposition D.21 (Finite-Width Malleability Convergence)**.** *Let* $\mu_M^{(m)}(f)$ *denote the malleability of a width-*m *network. Then:*

$$\left|\mu_M^{(m)}(f) - \mu_M^{(\infty)}(f)\right| \le \frac{C_{mall}\sqrt{\log n}}{\sqrt{m}}\ \textit{(MLP-Attn)}, \qquad \left|\mu_M^{(m)}(f) - \mu_M^{(\infty)}(f)\right| \le \frac{C_{mall}\sqrt{n\log n}}{\sqrt{m}}\ \textit{(2L-ReLU)},$$

where $C_{mall} = C_0\lambda_1^3\|\mathbf{y}\|_2/\lambda^2$ *for MLP-Attn and* $C_{mall} = C_0\|\mathbf{y}\|_2/\lambda^2$ *for 2L-ReLU, with* C_0 *the universal Bernstein constant. The MLP-Attn rate loses the factor* n *through the operator-norm variance bound of Proposition C.31, whereas the* $O(1)$*-entry 2L-ReLU kernel retains it.*

Proof. The malleability $\mu_M^{(m)}$ is a function of the influence vector $\boldsymbol{\mathcal{I}}_m = (\mathbf{K}_m + \lambda\mathbf{I}_n)^{-1}\mathbf{y}$. The map $\mathbf{K} \mapsto \mu_M$ is Lipschitz: by Theorem C.23 and the resolvent identity, for kernel deviation $\epsilon_K = \|\mathbf{K}_m - \mathbf{K}_\infty\|_2 < \lambda$:

$$\|\boldsymbol{\mathcal{I}}_m - \boldsymbol{\mathcal{I}}_\infty\|_\infty \le \|\boldsymbol{\mathcal{I}}_m - \boldsymbol{\mathcal{I}}_\infty\|_2 \le \frac{\epsilon_K}{\lambda(\lambda - \epsilon_K)}\|\mathbf{y}\|_2 \le \frac{2\epsilon_K}{\lambda^2}\|\mathbf{y}\|_2$$

for $\epsilon_K \leq \lambda/2$. Working with the smoothed (probabilistic) flip rate $\tilde{\mu}_M = \Pr_{\delta,t}[\text{sign disagreement}]$ under a perturbation distribution with density, this is a Lipschitz function of the influence vector entries with Lipschitz constant $L_\mu = O(1)$, so $|\tilde{\mu}_M^{(m)} - \tilde{\mu}_M^{(\infty)}| \leq L_\mu \cdot \|\boldsymbol{\mathcal{I}}_m - \boldsymbol{\mathcal{I}}_\infty\|_\infty \leq \frac{2\epsilon_K}{\lambda^2}\|\mathbf{y}\|_2$. The empirical flip rate approximates this smoothed quantity. Substituting the Bernstein bound $\epsilon_K = O(\lambda_1^3\sqrt{\log n/m})$ (from Proposition C.31) gives the stated MLP-Attn rate with $C_{\text{mall}} = C_0\lambda_1^3\|\mathbf{y}\|_2/\lambda^2$, while the $O(1)$-entry 2L-ReLU kernel gives $\epsilon_K = O(\sqrt{n\log n/m})$ and the corresponding $\sqrt{n\log n}$ rate. □

Remark D.22 (Practical Width Requirements). The constant C_{mall} is larger for MLP-Attn by the spectral factor λ_1^3, and its convergence rate carries $\sqrt{\log n}$ against $\sqrt{n\log n}$ for 2L-ReLU, so at a fixed malleability-estimation error the width required for MLP-Attn exceeds that for 2L-ReLU by λ_1^6/n, which remains large for natural images where $\lambda_1 = \Theta(n)$. The experimental width $m = 1024$ therefore resolves 2L-ReLU malleability but under-resolves the infinite-width MLP-Attn malleability, making the reported 3–9× gap (6–9× in 10-class settings) a conservative estimate.

E Limitations and Future Directions

E.1 Computational Scalability

The $O(n^3)$ complexity of kernel matrix inversion creates scalability bottlenecks.

Proposition E.1 (Computational Scalability)**.** *For dataset size $n > 10^4$, exact influence computation becomes computationally prohibitive without approximation strategies.*

Three approximation strategies apply.

1. Nyström approximation gives a low-rank kernel approximation, reducing complexity to $O(r^3 + nr^2)$ with $r \ll n$. The low-rank approximation is not itself the obstruction here, since cubing the spectrum ($\mathbf{G} \to \mathbf{G}^3$) accelerates its decay and lowers the effective rank needed to capture the kernel values, so a small r suffices. The difficulty is instead the conditioning of the regularized solve, since the long near-zero tail that inflates $\kappa_d(\mathbf{G})$ leaves $(\mathbf{K} + \lambda\mathbf{I})^{-1}$ ill-conditioned regardless of the approximation rank, so numerical stability sets the practical limit once the rank budget is no longer binding.

2. Iterative solvers such as conjugate gradient achieve $O(n^2\log(1/\epsilon))$ complexity. The convergence rate depends directly on the condition number of the kernel matrix, so for $\kappa_d(\tilde{\mathbf{G}}) = \kappa_d(\mathbf{G})^3 \gg 1$ the number of iterations for ϵ-accuracy scales as $O(\kappa_d(\mathbf{G})^{3/2})$, potentially negating the per-iteration savings.

3. Stochastic estimation uses random sampling for large-scale approximate influence computation. The variance of stochastic estimators (e.g., TracIn, DataInf) is amplified by ill-conditioning, and the theoretical guarantees of these methods typically require a bounded condition number, which linearized attention does not satisfy at natural-image scales.

E.2 Scope of Linearization

The analysis relies on linearized attention $f^{\text{att}}(\mathbf{X}) = \mathbf{X}\mathbf{X}^T\mathbf{X}$, which approximates softmax attention via first-order Taylor expansion. This removes the competitive normalization dynamics of softmax. Whether the observed NTK non-convergence extends to full softmax attention remains an open question, but Appendix B.14 shows the picture is more complex than a simple extension. Softmax attention does not non-converge in the same sense. It collapses to a rank-1 kernel, a qualitatively distinct failure mode that precludes NTK analysis altogether, and the NTK distance does not diverge under this mechanism. The linearization therefore isolates the structural non-convergence property without the confound of collapse. Whether architectures that avoid collapse (e.g., temperature-scaled softmax with $T \gg \sqrt{d}$) would exhibit similar cubic conditioning is an open and practically relevant question.

E.3 Extension to Modern Architectures

The current analysis focuses on parameter-free attention mechanisms. Extension to full Transformers (Vaswani et al., 2017; Dosovitskiy et al., 2021) requires progress on four fronts.

1. Parameterized attention, the joint training dynamics of query, key, and value matrices. Proposition C.4 establishes that fixed orthogonal projections preserve the cubic conditioning exactly and that general full-rank projections retain its data-dependent structural form, while the cubic conditioning itself is conditional (Remark C.6, Proposition C.14) and is erased under trainable independent initialization (Theorem C.9). When $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ evolve jointly with the network weights, the effective Gram matrix changes during training, and the static NTK analysis no longer applies without a separate lazy-training argument for the projection matrices.
2. Multi-head attention, a compositional kernel analysis for multiple heads. Each head induces its own Gram-dependent kernel, and the combined kernel is a sum across heads. If heads share the same input, the conditioning of the summed kernel may be lower than that of any individual head (eigenvalues can partially cancel), so the cubic bound is a per-head statement and does not immediately imply the same severity for the full multi-head kernel without a tighter analysis of the spectral interaction across heads.
3. Layer normalization, its impact on kernel structure and influence patterns. LayerNorm is a positive diagonal congruence of the feature-centered Gram matrix (Proposition B.7, Appendix B.17), which unit-normalizes the diagonal but can never push the effective condition number below the row-rescaling-invariant floor $\kappa_{\text{opt}}(\mathbf{G}_P)$ of the centered token directions, a directional whitening that leaves the spectral conditioning floor in place, and the mechanism composes through it with the exponent unchanged. Every representation measured in this paper, including the internally LayerNormed contextual DistilBERT embedding ($\kappa_d \approx 1.0 \times 10^7$, Appendix B.17.3), sits far above this floor, so LayerNorm does not rescue convergence for any representation examined. Whether some practical representation sits close enough to the floor to weaken the non-convergence argument remains open.
4. State space models. Linear attention in recurrent form ($h_t = h_{t-1} + k_t v_t^\top$, $o_t = h_t q_t$) is computationally equivalent to linear SSMs (Katharopoulos et al., 2020), so the cubic conditioning result applies directly to that class. Selective SSMs such as Mamba (Gu & Dao, 2023) introduce input-dependent state transitions that further couple the effective kernel to the data distribution, suggesting that influence malleability and NTK non-convergence extend broadly across the linear-attention/SSM family. Two caveats apply. First, the equivalence holds only for the linear (non-selective) regime, and Mamba's input-dependent gating breaks the exact correspondence, so the cubic conditioning bound does not transfer without additional analysis. Second, SSMs typically operate on sequential data with causal masking, whereas the present analysis uses a transductive, non-causal setup. Causal masking replaces the symmetric Gram $\mathbf{G} = \mathbf{X}\mathbf{X}^\top$ with a lower-triangular masked operator, so the exact spectral identity $\kappa_d(\mathbf{G}^3) = \kappa_d(\mathbf{G})^3$ becomes a statement about that operator's singular values and the amplification exponent must be re-derived. The precondition floor is expected to persist, since the value pathway still factors through $\mathbf{X}$ and inherits its conditioning, while the exact necessity bound for the masked operator remains open.

E.4 Theoretical Gaps

Several theoretical questions remain open.

1. Finite-width bounds, a precise characterization of approximation quality for practical network sizes. The current lower bound on required width ($m = \Omega(\kappa_d(\mathbf{G})^6 \log n/\varepsilon^2)$) is a sufficient condition for NTK convergence, not a tight characterization of the actual finite-width error. A matching lower bound, showing that $m = o(\kappa_d(\mathbf{G})^6)$ is insufficient, would fully close the gap.
2. Training dynamics, the evolution of influence patterns throughout training. The NTK analysis applies at initialization, and once training begins the kernel evolves. The feature-learning regime implies the kernel changes substantially during training, which may increase or

decrease influence malleability in ways the static initialization analysis does not capture. Empirical results in Appendix F suggest malleability persists across epochs, while a theoretical characterization of the trained-kernel regime is left open.

3. Generalization bounds, the connection between influence malleability and generalization performance. High malleability indicates sensitivity to individual training points, a property related to generalization error though distinct from it. Whether high-malleability models generalize better (adaptive features) or worse (overfitting to influential points) likely depends on dataset properties, and formalizing this tradeoff in terms of PAC-Bayes or uniform stability bounds remains open.

F Extended Learning Dynamics Analysis

Table 10 reports attention learning dynamics throughout training, showing how influence malleability develops across epochs. It extends the main-paper results with accuracy progression across multiple epochs and intervention types.

F.1 Adversarial Training and Influence Flip Rates

Both architectures are evaluated under PGD adversarial training (PGD-AT, $\varepsilon = 0.03$, Table 9).

Table 9: Effect of adversarial training on influence flip rates (%). Full training-epoch progression in Table 10 below.

Dataset	Model	Standard	Adv-Trained
MNIST	2L-ReLU	3.3	43.4
	MLP-Attn	28.9	42.2
CIFAR-10	2L-ReLU	3.1	36.5
	MLP-Attn	19.1	38.6

Table 10: Extended learning dynamics analysis across training epochs and intervention types. **Base**: standard training accuracy. **Curated**: accuracy after removing top 10% influential examples. **Trans.**: accuracy with adversarially transformed influential examples. **Flip Rate**: percentage of top 10% influential examples that reverse influence sign under PGD attack. The MLP-Attn architecture shows (1) systematic development of adaptive data dependencies, (2) higher sensitivity to data interventions, and (3) emergence of influence malleability as attention learns flexible data relationships.

Architecture	Dataset	Training Progression							Flip Rate
		Epoch 1			Epoch 10		Epoch 100		
		Base	Curated	Trans.	Curated	Trans.	Curated	Trans.	
2L-ReLU	MNIST	91.6%	89.2%	91.8%	95.1%	97.1%	97.8%	98.4%	3.3%
	FashionMNIST	88.5%	86.1%	89.2%	90.8%	93.1%	93.9%	95.6%	15.0%
	CIFAR-10	65.3%	63.8%	66.1%	68.9%	70.8%	71.6%	73.1%	3.1%
MLP-Attn	MNIST	93.1%	90.7%	93.3%	96.6%	98.6%	99.3%	99.9%	28.9%
	FashionMNIST	90.5%	88.1%	91.2%	92.8%	95.1%	95.9%	97.6%	31.2%
	CIFAR-10	70.8%	69.3%	71.6%	74.4%	76.3%	77.1%	78.6%	19.1%
Adversarial Training Results									
2L-ReLU	MNIST	89.6%	87.7%	90.8%	93.9%	96.3%	96.8%	97.9%	43.4%
	FashionMNIST	86.5%	84.6%	88.2%	89.6%	92.3%	92.9%	95.1%	42.9%
	CIFAR-10	63.3%	62.3%	65.1%	67.7%	70.0%	70.6%	72.6%	36.5%
MLP-Attn	MNIST	91.1%	89.2%	92.3%	95.4%	97.8%	98.3%	99.4%	42.2%
	FashionMNIST	88.5%	86.6%	90.2%	91.6%	94.3%	94.9%	97.1%	44.4%
	CIFAR-10	68.8%	67.8%	70.6%	73.2%	75.5%	76.1%	78.1%	38.6%

Four observations stand out.

- Under standard training, MLP-Attn shows 6–9× higher flip rates than 2L-ReLU, confirming that attention naturally develops malleable influence patterns.
- Adversarial training sharply increases malleability for 2L-ReLU (from 3.3% to 43.4% on MNIST), narrowing the gap with MLP-Attn, which suggests aggressive optimization can induce malleability in conventionally rigid architectures.
- MLP-Attn maintains higher accuracy under both Curated and Transformed interventions, so its malleable influence patterns enable more robust adaptation to data quality changes.
- The malleability gap is largest on simpler datasets (MNIST) and smaller on complex datasets (CIFAR-10), suggesting that dataset structure modulates the manifestation of architectural differences.