How Much Can You Project? A Neural Tangent Kernel Theory of Intrinsic Dimension

Anonymous authors
Paper under double-blind review

Abstract

Neural networks with millions of parameters are routinely trained within a surprisingly small effective parameter space — a phenomenon formalized by Li et al. as *intrinsic dimension*, which has remained without theoretical explanation since its discovery. We provide the first such explanation for asymptotically wide networks, showing that intrinsic dimension emerges from the interplay of the Neural Tangent Kernel (NTK) and the Johnson–Lindenstrauss lemma. Specifically, we ask whether the NTK regime survives when training is restricted to a randomly chosen low-dimensional subspace of the full parameter space. We prove that it does when the subspace dimension d exceeds a threshold of order $\varepsilon^{-2} \log N$: a block-structured random Gaussian projection preserves the NTK entry-wise throughout training with high probability as network width grows. Crucially, this threshold is independent of the ambient parameter dimension D, which can be arbitrarily large. We complement this result with matching lower bounds: the ε^{-2} dependence is unavoidable for random Gaussian projections, while no linear map can beat the $\varepsilon^{-2} \log N$ threshold for the hardest gradient configurations. Thus, we identify a critical dimension d^* above which random subspace training preserves the NTK regime and below which the same mechanism fails. We also identify a distinct threshold at $d = N$: below it, the projected kernel is rank-deficient and a component of the training error is permanently frozen, with the effect determined by its alignment with the surviving eigenspace. Experiments on standard benchmarks support our theoretical predictions.

1 Introduction

A remarkable and poorly understood empirical phenomenon sits at the heart of modern deep learning: neural networks with millions (sometimes billions) of parameters appear, in practice, to optimize within a drastically smaller effective parameter space. Li et al. (2018) formalized this observation through the notion of *intrinsic dimension*: the minimum number of parameters d sufficient to reach a satisfactory solution on a given task, when training is restricted to a randomly chosen d-dimensional affine subspace of the full parameter space. Their experiments revealed that despite ambient dimensions D on the order of millions, the intrinsic dimension d required to solve complex vision and language tasks is often surprisingly small — sometimes in the hundreds. This finding is striking not merely as an empirical curiosity but as a fundamental statement about the geometry of neural network optimization landscapes: the loss surface, for all its apparent complexity in $\mathbb{R}^D$, can be effectively navigated from within a random low-dimensional slice.

Yet despite the significance of this observation — and the practical implications it carries for parameter-efficient training, model compression, and our basic understanding of why overparameterization helps rather than hurts — the intrinsic dimension phenomenon has remained almost entirely without theoretical explanation. Why should a randomly chosen subspace of dimension $d \ll D$ suffice for training? What governs the value of d, and why does it appear to be largely independent of the ambient dimension D, at least when the network depth L is fixed? Why does d grow so slowly with the number of training samples N? And what is the role of choosing a *random* subspace, as opposed to other structured or data-dependent restrictions? These questions have been left largely open, leaving a gap in the theoretical foundations of deep learning. The present work takes a first step toward closing this gap.

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.

Our approach. We study this problem in the regime where network width is very large, analyzing it through the lens of the Neural Tangent Kernel (NTK) framework Jacot et al. (2018), which characterizes the training dynamics of sufficiently wide neural networks in the infinite-width limit. The NTK framework establishes that, as network width grows, gradient flow converges to a linear dynamical system governed by a kernel K_∞ that remains approximately constant throughout training — a linearization that makes the dynamics tractable and is precisely what we exploit here. If the NTK regime persists when training is restricted to a low-dimensional random subspace, then the resulting projected dynamics remain approximately linear and hence analyzable; if it breaks down, the geometry of the subspace is too impoverished to support the original training dynamics. Our central finding is that a well-known tool — the Johnson–Lindenstrauss (JL) lemma (Johnson et al., 1984), which guarantees approximate preservation of pairwise distances under random projection — is, perhaps surprisingly, sufficient to explain most of the open questions raised above, and to do so quantitatively.

1.1 Our Contributions

We establish a rigorous theoretical characterization of the conditions under which subspace-restricted training preserves the NTK regime, providing what we believe is the first theoretical account of the intrinsic dimension phenomenon in the setting of asymptotically wide networks. To state our results, we need one piece of notation: we say the projected NTK is ε-*faithful* if it agrees with the infinite-width kernel K_∞ entry-by-entry to within a multiplicative tolerance ε — that is, if restricting training to the subspace does not meaningfully distort the kernel that governs the dynamics. The question is then: how large must the subspace dimension d be for this to hold with high probability?

Our main results answer this question from both sides. On the positive side, we prove that a subspace dimension scaling logarithmically in the number of training points N — independently of the ambient parameter dimension D, which can be arbitrarily large — suffices for the projected NTK to be ε-faithful throughout training, with high probability over the random projection (see Theorem 4.2, or more formally Theorem A.14). On the negative side, we prove that the same family of random projections cannot succeed below this threshold: the kernel distortion is bounded away from zero with constant probability, for any fixed configuration of gradient vectors, with no special structure assumed (see Section 4.2). Together, these two results identify a *critical subspace dimension* d^*, scaling as $\varepsilon^{-2} \log N$ up to depth factors, above which random subspace training provably preserves the NTK regime and below which the same mechanism provably fails. Our theoretical results are established in the infinite-width NTK regime; extending them to finite-width networks and explaining why empirical intrinsic dimensions are often substantially smaller than the worst-case bound remain important open questions.

2 Related Work

Theoretical work on low-dimensional neural network training lies at the intersection of infinite-width neural network theory, intrinsic dimension, and random dimensionality reduction. We briefly discuss each.

Neural Tangent Kernel and Infinite-Width Dynamics. The Neural Tangent Kernel (NTK), introduced by Jacot et al. (2018), shows that in the infinite-width limit, gradient flow is governed by a kernel that remains approximately constant during training, yielding tractable linear dynamics. Subsequent work established exact NTK formulas and convergence guarantees for various architectures (Arora et al., 2019; Du et al., 2019; Allen-Zhu et al., 2019). These results consider training in the full parameter space; we ask whether the same stable dynamics survive when training is restricted to a low-dimensional random subspace.

Intrinsic Dimension and Random Subspace Training. Li et al. (2018) provided landmark empirical evidence that neural networks can reach satisfactory solutions when training is restricted to random subspaces of dimension $d \ll D$, formalizing *intrinsic dimension*. This builds on earlier random-subspace optimization methods (Wang et al., 2013) and extends to pretrained language models, where pretraining can further reduce intrinsic dimension (Aghajanyan et al., 2020). Most closely related to our setting, Gressmann et al. (2020) found that layer-wise (block-diagonal) random pro-

jections substantially outperform a single global projection. We provide a theoretical explanation for this observation by showing that block-diagonal projections yield controlled, additive distortion of the NTK across layers. Our setting differs from deterministic low-rank adaptation such as LoRA (Hu et al., 2021) and sparse subnetworks studied by the Lottery Ticket Hypothesis (Frankle & Carbin, 2018): we consider continuous random subspaces and derive explicit dimension guarantees.

Dimension Reduction and the Johnson–Lindenstrauss Lemma. The Johnson–Lindenstrauss (JL) lemma (Johnson et al., 1984) provides the key dimensionality-reduction tool in our analysis. We use its inner-product form to control entry-wise distortion of the NTK Gram matrix, together with standard high-dimensional probability techniques (Vershynin, 2018). For lower bounds, we use sharp chi-squared tail estimates (Zhang & Zhou, 2020) and the optimality result of Larsen & Nelson (2017), which establishes the $\varepsilon^{-2} \log N$ barrier for linear dimension reduction. Two recent papers are closely related in tools but distinct in scope. Malladi et al. (2023) extend NTK theory to Adam via Tensor Programs and invoke the inner-product JL lemma once, at fixed parameters, to argue informally that subspace methods such as LoRA inherit the laziness of full fine-tuning; Li et al. (2026) show that zeroth-order optimization implicitly realizes a JL-type projection of the eNTK onto freshly redrawn perturbation directions each step. Both results are single-instant or single-step: neither fixes one random subspace before training and tracks its fidelity throughout, and Li et al. (2026)'s directions change every iteration, so no *persistent* subspace is defined in their setting. We instead fix the subspace once and prove a uniform-in-time guarantee across the entire trajectory, together with a depth-dependent critical dimension $d^* = \Theta(L^3 \varepsilon^{-2} \log N)$ and a separate rank threshold $d = N$ below which part of the training error is permanently frozen — neither identifiable from a single-instant or single-step analysis.

Optimization in Low-Dimensional Subspaces. Low-dimensional structure in neural network optimization has also been studied empirically and through projected kernel methods. Gur-Ari et al. (2018) observed concentration of gradient updates in low-dimensional Hessian eigenspaces, while Li et al. (2026) connected zeroth-order optimization to a projected empirical NTK. In contrast, we establish directly that the NTK dynamics remain stable under random low-dimensional subspace training, with a dimension threshold depending on the task size and accuracy but not the ambient parameter dimension.

3 Background and Preliminaries

This section introduces the concepts underlying our analysis. We first recall *intrinsic dimension*, then the *neural tangent kernel*, and finally formulate the problem precisely in Section 3.1. Consider a fully-connected feed-forward network of fixed depth $L \geq 1$, with n_ℓ neurons in layer ℓ, $\ell = 0, \ldots, L$. The network $f(\cdot, \cdot; \theta) : \mathbb{R}^{n_0} \to \mathbb{R}$ has parameters $\theta \in \mathbb{R}^D$, where layer ℓ contributes $N_\ell = (n_{\ell-1} + 1)n_\ell$ parameters and $D = \sum_{\ell=1}^{L} N_\ell$. We assume a nonlinear Lipschitz activation σ with constant $\mathcal{O}(1)$. We are given a training set $\mathcal{D} = (x_i, y_i)_{i \in [N]}$ and empirical loss $\mathcal{L}(\theta)$. [1]

The intrinsic dimension of a learning task. Li et al. (2018) showed empirically that highly overparameterized networks can often be trained successfully within a random low-dimensional affine subspace. Fixing an initialization θ_0, one draws $P \in \mathbb{R}^{D \times d}$ and restricts the parameters to

$$\theta(\phi) = \theta_0 + P\phi, \qquad \phi \in \mathbb{R}^d,$$

with $\phi = 0$ initially and both θ_0 and P fixed throughout training. The smallest $d \ll D$ for which this restricted training reaches a satisfactory solution is referred to as the *intrinsic dimension* of the task.

Empirically, d can be orders of magnitude smaller than D, with the effect becoming more pronounced after pre-training (Li et al., 2018; Aghajanyan et al., 2020). Layer-wise random projections have also been found to outperform a single global projection (Gressmann et al., 2020). These observations motivate the central theoretical question of this paper: under what conditions can training in such a random subspace preserve the dynamics of full-space training?

[1] We use scalar output for simplicity; the vector-valued case is treated in Appendix A.

The neural tangent kernel. For gradient flow, $\dot{\theta} = -\nabla_\theta \mathcal{L}(\theta)$, the evolution of network predictions is governed by the empirical neural tangent kernel (NTK) (Jacot et al., 2018).

Definition 3.1 (Empirical neural tangent kernel)**.** Let $g_i \triangleq \nabla_\theta f(x_i;\theta) \in \mathbb{R}^D$. The empirical NTK is the Gram matrix

$$K_{ij} = \big\langle \nabla_\theta f(x_i;\theta), \nabla_\theta f(x_j;\theta) \big\rangle = \langle g_i, g_j \rangle.$$

For squared loss, the residual $\vec{\Delta}^{(t)} = (f_{\theta^{(t)}}(x_i) - y_i)_{i=1}^N$ satisfies $\frac{d\vec{\Delta}^{(t)}}{dt} = -\frac{1}{N} K^{(t)} \vec{\Delta}^{(t)}$. In the infinite-width regime, the NTK remains asymptotically constant during training, $K^{(t)} \approx K^{(0)} =: K$, yielding $\vec{\Delta}^{(t)} = \exp\big(-\frac{t}{N} K\big)\, \vec{\Delta}^{(0)}$. Thus, wide-network training is governed by a fixed kernel and linear dynamics; when K is positive definite, the training error converges to zero (Du et al., 2019; Allen-Zhu et al., 2019). Our goal is to determine whether this kernel regime survives random low-dimensional projection.

3.1 Problem Formulation

We now formulate the question studied in the rest of the paper. Fix a training set $\mathcal{D} = (x_i, y_i)_{i\in[N]}$ and a depth-L network with fixed input and output dimensions n_0, n_L, while the hidden widths $n_1, \ldots, n_{L-1}$ tend to infinity. Then $D = \sum_\ell N_\ell \to \infty$. Instead of training in $\mathbb{R}^D$, draw a Gaussian projection $P \in \mathbb{R}^{D\times d}$ before training and restrict the parameters to

$$\theta(\phi) = \theta_0 + P\phi, \qquad \phi \in \mathbb{R}^d,$$

where θ_0 is the random initialization and only ϕ is trained. The projection is drawn independently of θ_0 and remains fixed. Our main construction uses a block-diagonal variant that projects each layer separately; see Section 4. Our central question is: In the unprojected infinite-width limit, the NTK converges at initialization to a deterministic kernel K_∞ and remains approximately constant during training. We ask whether the same regime survives the projection. Specifically, two properties are required:

1. **Initialization.** The projected kernel Θ^P, obtained by replacing each gradient g_i with $P^\top g_i$, is close to K_∞ at $t = 0$; equivalently, the projection approximately preserves the relevant gradient inner products.
2. **Stability throughout training.** For any fixed $T \geq 0$, as the hidden widths tend to infinity, $\Theta^P(t)$ remains close to $\Theta^P(0)$ uniformly for $t \in [0, T]$.

When both properties hold, projected training inherits the full-space NTK dynamics. The question is therefore how small d can be while they hold with high probability over P. For accuracy $\varepsilon \in (0,1)$ and failure probability $\delta \in (0,1)$, we define the corresponding critical dimension.

Definition 3.2 (Critical dimension)**.** The critical dimension $d^*(\varepsilon, \delta)$ is the smallest d such that, with probability at least $1 - \delta$ over the random projection P, properties (i) and (ii) hold simultaneously with relative entry-wise kernel error at most ε.

The key question is whether d^* remains bounded as $D \to \infty$. Our results show that it does: d^* depends on N, ε, and depth-dependent factors, but not on the ambient parameter dimension D. We do not optimize the polynomial dependence on L. **The converse:** Below the critical scale, the same Gaussian random-projection mechanism cannot, in general, preserve the kernel: the distortion remains bounded away from zero with constant probability. We establish this obstruction for Gaussian projections in Appendix B and for general linear maps in Appendix C.

4 A NTK Theory Approach

The whole analysis rests on a single classical tool, which we recall first.

Lemma 4.1 (Johnson–Lindenstrauss (Johnson et al., 1984))**.** *For any $0 < \varepsilon < 1$, $\delta \in (0,1)$, and any set V of n points in $\mathbb{R}^D$, let $d \geq C_{\mathrm{JL}}\, \varepsilon^{-2} \log(n/\delta)$ for a universal constant $C_{\mathrm{JL}} > 0$, and let $P \in \mathbb{R}^{d\times D}$ have i.i.d. entries $P_{ij} \sim \mathcal{N}(0, 1/d)$. Then with probability at least $1-\delta$, simultaneously for all $u, v \in V$,*

$$(1-\varepsilon)\|u - v\|^2 \;\leq\; \|Pu - Pv\|^2 \;\leq\; (1+\varepsilon)\|u - v\|^2.$$

The lemma says a random Gaussian map into dimension only $d = \mathcal{O}(\varepsilon^{-2} \log n)$ — growing *logarithmically* in the number of points — preserves every pairwise distance up to a factor $1 \pm \varepsilon$. Its relevance here is direct. The tangent kernel is the Gram matrix of the per-sample gradients, $K_{ij} = \langle g_i, g_j \rangle$ with $g_i = \nabla_\theta f(x_i; \theta)$; projecting the parameters replaces each gradient g_i by its image under the projection, so the kernel is preserved exactly when the projection preserves these inner products. Distances and inner products carry the same information, so the distance guarantee above translates directly into an inner-product guarantee of the same form (Lemma A.2 in the appendix). There is, however, one essential caveat: the lemma controls a *fixed* set of vectors, whereas during training the gradients move. Bridging that gap — from a guarantee at initialization to one that holds uniformly throughout training — is the technical heart of what follows.

Instead of projecting the full parameter vector through a single matrix, we project *each layer separately*. Concretely, write $\theta = (\theta^{(1)}, \ldots, \theta^{(L)})$ with $\theta^{(\ell)} \in \mathbb{R}^{N_\ell}$, and for each layer draw an independent Gaussian projection

$$P^{(\ell)} \in \mathbb{R}^{N_\ell \times d_\ell}, \qquad \big(P^{(\ell)}\big)_{ab} \sim \mathcal{N}\big(0, \tfrac{1}{d_\ell}\big),$$

restricting layer ℓ to the affine subspace $\theta^{(\ell)}(\phi^{(\ell)}) = \theta_0^{(\ell)} + P^{(\ell)}\phi^{(\ell)}$, $\phi^{(\ell)}(0) = 0$. The total trainable dimension is $d = \sum_{\ell=1}^{L} d_\ell$, with the symmetric split $d_\ell = d/L$. This block structure is what allows the kernel-regime analysis to proceed layer by layer, following the same inductive machinery as Jacot et al. (2018); we show that the cost of this restriction is at most a polynomial factor in L. Training evolves the reduced coordinates $\phi = (\phi^{(1)}, \ldots, \phi^{(L)})$ by gradient flow which explained below. [2]

The projected NTK replaces each gradient block $g_i^{(\ell)} \in \mathbb{R}^{N_\ell}$ by its projection $(P^{(\ell)})^\top g_i^{(\ell)}$:

$$\Theta_{ij}^{P} \;=\; \sum_{\ell=1}^{L} \Big\langle \big(P^{(\ell)}\big)^\top g_i^{(\ell)},\; \big(P^{(\ell)}\big)^\top g_j^{(\ell)} \Big\rangle.$$

Two basic properties orient the analysis. The projection is *unbiased*: $\mathbb{E}[P^{(\ell)}(P^{(\ell)})^\top] = I_{N_\ell}$ implies $\mathbb{E}[\Theta^P] = K$, so Θ^P is a random estimate of the true kernel whose accuracy is controlled by the block dimensions d_ℓ through Lemma 4.1. The projected kernel is also *rank-deficient* for small d: since $\Theta^P = \sum_\ell F_\ell F_\ell^\top$ with $F_\ell \in \mathbb{R}^{N \times d_\ell}$, we have $\operatorname{rank}(\Theta^P) \le d$, so for $d < N$ the projected kernel is singular. This singularity has its own consequences, which we analyze separately in Section 4.3. Under projected gradient flow the residual obeys $\frac{d\vec{\Delta}^{(t)}}{dt} = -\frac{1}{N}\Theta^P(t)\,\vec{\Delta}^{(t)}$, with the same linear form as the standard NTK dynamics.

4.1 Main Result: Time-Stability of the Projected Kernel

We establish stability in two stages, both carried out in full in Appendix A.

Stage I: closeness at initialization. Applying the inner-product Johnson–Lindenstrauss lemma to each layer's gradient blocks separately and combining the L resulting guarantees via a union bound shows that $\Theta^P(0)$ matches $K(0)$ entry-wise to within ε as soon as $d_\ell \gtrsim \varepsilon^{-2} \log(N/\delta)$ for each ℓ (Theorem A.3). The block structure is what makes this tractable: per-layer distortions *add* rather than compound, keeping the total dimension modest.

Stage II: the kernel barely moves. Closeness at initialization is not enough, because training changes the parameters and hence the gradients that define Θ^P. The saving grace is exactly the property that makes the ordinary NTK useful: in the wide-network limit the parameters move so little that the kernel is essentially frozen. Here, however, the standard argument does not carry over the rest: the projection $P^{(\ell)}$ that keeps the projected gradient the same size as the true gradient is precisely the map under which the *ambient* parameter displacement is no longer controlled — unlike full-space training, $\|\theta_P(t) - \theta_0\|_2$ can grow with width, so "*the parameters move little*" cannot be the starting point. What we show instead is that the *subspace coordinate* $\phi(t)$ itself stays confined to a

[2] The matrix $P^{(\ell)}$ maps reduced coordinates up to the ambient block, $\mathbb{R}^{d_\ell} \to \mathbb{R}^{N_\ell}$; the object acting on gradients is its transpose $(P^{(\ell)})^\top : \mathbb{R}^{N_\ell} \to \mathbb{R}^{d_\ell}$, which is exactly the Johnson–Lindenstrauss map of Lemma 4.1. The single-matrix view is the one-block special case; we keep it for intuition but use the block construction throughout.

fixed, width-independent ball throughout training, and that this confinement, together with a uniform bound on the network's second derivative furnished by the NTK parametrization, is enough to freeze the kernel.

Establishing the confinement is where the real difficulty lies: bounding $\phi(t)$ requires the very Hessian bound that is only valid once $\phi(t)$ is known to be confined. We break this circularity with a stopping-time argument: we fix a candidate confinement radius from data available at initialization alone, assume it holds up to the first time it might fail, and show that on this (possibly empty) interval the true bound is strictly tighter than what was assumed. A continuity argument then rules out the interval ever ending early, so the assumed confinement in fact holds for all $t \geq 0$ (Lemma A.10). The block structure of P is what makes even this tractable: because each layer has its own independent projection matrix, the projected Hessian sandwiched against $P^{(\ell)}$ inherits the same $1/\sqrt{\min_\ell n_\ell}$ suppression as the unprojected one, layer by layer, rather than the $O(N_\ell/d_\ell)$ blow-up a single global projection would incur. With confinement in hand, the drift of $\Theta^P(t)$ away from $\Theta^P(0)$ vanishes at the same $1/\sqrt{\min_\ell n_\ell}$ rate as in the full-space case (Theorem A.12). Chaining the two stages gives the central guarantee of the paper.

Theorem 4.2 ((Informal) Main Result). *Fix $\varepsilon, \delta \in (0,1)$, $T > 0$, and a dataset of size $N \geq 1$. Consider a fully-connected network of depth $L \geq 2$ with hidden widths $n_1, \ldots, n_{L-1}$, and let K_∞ denote the limiting unprojected NTK matrix as $\min_i n_i \to \infty$. Suppose training is restricted to a block-structured random subspace of total dimension d, with each layer assigned an independent Gaussian projection and an appropriately chosen layer-wise learning rate. If*

$$d \;\geq\; \Omega\left(\frac{L^3}{\varepsilon^2}\log\left(\frac{NL}{\delta}\right)\right),$$

then, as the network's width tends to infinity, with probability at least $1-\delta$,

$$\sup_{0\leq t\leq T}\;\max_{1\leq i,j\leq N}\left|\Theta^P_{ij}(t) - K_\infty(i,j)\right| \;\leq\; \mathcal{O}(\varepsilon).$$

The formal version, with explicit constants and a careful treatment of the probability space as width tends to infinity, is given by Theorem A.14 (aggregation of Theorem A.3 and Theorem A.12).

It is worth reading the dimension requirement factor by factor. The ε^{-2} is the Johnson–Lindenstrauss rate: halving the allowed distortion costs four times the dimension. The $\log(NL/\delta)$ is the mild logarithmic price of demanding the guarantee for all N data points and L layers at once, with confidence $1-\delta$. The L^3 is the cost of the block construction: controlling the total distortion to within ε requires each of the L layers to contribute at most ε/L, which demands $(L/\varepsilon)^2$ dimensions per layer, and summing over L layers gives L^3/ε^2. The conclusion is that a subspace of this size keeps the network in the kernel regime, so projected training tracks full-space training to within ε throughout.

4.2 Matching Lower Bounds and the Critical Dimension

Theorem 4.2 gives a sufficient condition for preserving the NTK under random projection. We now show that its $\varepsilon^{-2}\log N$ dependence is unavoidable in the worst case, with the ε^{-2} obstruction already arising from a single gradient vector. For a nonzero vector $u \in \mathbb{R}^D$, a random Gaussian projection yields a chi-squared random variable, and hence

$$\mathbb{P}\big[\|Pu\|_2^2 \leq (1-\varepsilon)\|u\|_2^2\big] \geq c\exp(-C_\varepsilon d\varepsilon^2)$$

for absolute constants $c, C_\varepsilon > 0$, independently of D and u. Consequently, when $d = \mathcal{O}(\varepsilon^{-2})$, the projection fails to preserve even a single gradient norm with probability bounded away from zero. Thus the ε^{-2} dependence cannot be improved for the Gaussian projection family; see Appendix B and Theorem B.3. The $\log N$ dependence requires a stronger, worst-case argument. We construct configurations of N gradient vectors for which *any* linear projection satisfying the kernel-approximation condition of Theorem A.3 must have $d = \Omega(\varepsilon^{-2}\log N)$. This follows from the optimality theorem of Larsen & Nelson (2017) for Johnson–Lindenstrauss embeddings, transferred to inner products via polarization; see Appendix C and Theorem C.3. This is a worst-case statement and does not assert that typical NTK gradients attain the lower bound.

Theorem 4.3 ((Informal) Lower Bounds). *Fix $\varepsilon \in (0,1)$. The following hold:*

- **Random Gaussian projections.** *For any nonzero gradient vector g_i, a random Gaussian projection of dimension $d = \mathcal{O}(\varepsilon^{-2})$ fails to preserve $|g_i|_2^2$ to relative accuracy ε with probability bounded away from zero, independently of D and N.*

- **Worst-case linear maps.** *There exist configurations of N gradient vectors in $\mathbb{R}^D$ for which any linear projection satisfying the kernel-approximation condition of Theorem A.3 must have $d = \Omega(\varepsilon^{-2} \log N)$.*

Thus, ε^{-2} is unavoidable for Gaussian projections, while $\varepsilon^{-2} \log N$ is unavoidable in the worst case for arbitrary linear maps.

Together, these bounds show that the upper bound has the optimal $\varepsilon^{-2} \log N$ dimension-reduction scaling in the worst case, up to the depth factor L^3, universal constants, and logarithmic dependence on $1/\delta$. In particular, $d^*(\varepsilon) = \Theta\big(\varepsilon^{-2} \log N\big)$, up to these factors. Whether typical NTK gradient configurations admit smaller dimensions through structured or data-dependent projections remains open.

4.3 Empirical Risk Residuals

Theorem 4.2 guarantees that the projected kernel is well-approximated, but approximation is only half of what learning needs: the kernel must also be able to *drive the error to zero*. Here the projection has a second effect, invisible to the approximation analysis above, which is a matter of rank. In the standard NTK regime, K_{NTK} is strictly positive definite under mild conditions on the data and architecture, so every mode of the residual decays and the network reaches zero empirical risk. Restricting optimization to a d-dimensional subspace changes the spectrum fundamentally: the effective kernel $K_{\mathrm{eff}} = \sum_\ell \eta_\ell F_\ell F_\ell^\top$ with $F_\ell \in \mathbb{R}^{N \times d_\ell}$ satisfies $\mathrm{rank}(K_{\mathrm{eff}}) \le d$, so whenever $d < N$ the kernel has at least $N - d$ exactly zero eigenvalues and is only positive *semi*-definite. Decomposing the initial error as

$$\vec{\Delta}^{(0)} = \vec{\Delta}_{\parallel}^{(0)} + \vec{\Delta}_{\perp}^{(0)}, \qquad \vec{\Delta}_{\parallel}^{(0)} \in \mathrm{Col}(K_{\mathrm{eff}}), \quad \vec{\Delta}_{\perp}^{(0)} \in \mathsf{Ker}(K_{\mathrm{eff}}),$$

the two parts evolve completely differently. The column-space component decays exponentially at a rate set by the smallest nonzero eigenvalue of K_{eff}; the null-space component satisfies $\mathrm{d}\vec{\Delta}_{\perp}^{(t)}/\mathrm{d}t = -\frac{1}{N} K_{\mathrm{eff}} \vec{\Delta}_{\perp}^{(t)} = 0$ and does not move at all. The training error therefore settles at a strict residual floor,

$$\lim_{t \to \infty} \vec{\Delta}^{(t)} \;=\; \vec{\Delta}_{\perp}^{(0)}.$$

This floor is not a death sentence for every task. If the targets are such that the initial error has no component in $\mathsf{Ker}(K_{\mathrm{eff}})$ (i.e., the presence of an *inductive bias*), then $\vec{\Delta}_{\perp}^{(0)} = 0$ and the network drives the empirical risk to zero despite the extreme compression. Convergence under a projected subspace is, in the end, a question of *alignment*: whether the target lives in the surviving eigenspace of the projected kernel.

Two distinct thresholds therefore govern projected training and should not be conflated. The critical dimension $d^* \sim \varepsilon^{-2} \log N$ of Section 4.1 is a property of the *kernel geometry* and controls whether the NTK regime holds at all. The rank threshold $d \ge N$ is what a *worst-case* task needs before the error can reach zero, whereas a well-aligned target can succeed with far fewer dimensions. Section 5 tests both predictions directly.

5 Experimental Results

We test the three quantitative predictions of our theory in turn: that the block-projected tangent kernel $\Theta^P(0)$ approximates the true NTK $K(0)$ with an accuracy that improves as d grows (Theorem A.3); that $\Theta^P(t)$ stays essentially frozen along subspace training, placing the network in the kernel regime (Theorem 4.2); and that, as a consequence, training inside a random d-dimensional subspace recovers the full-parameter solution once d is moderate. Throughout, the headline finding is that the dimension needed is a *small* fraction of the ambient parameter count. Kernel-approximation diagnostics are deferred to Appendix D.2.

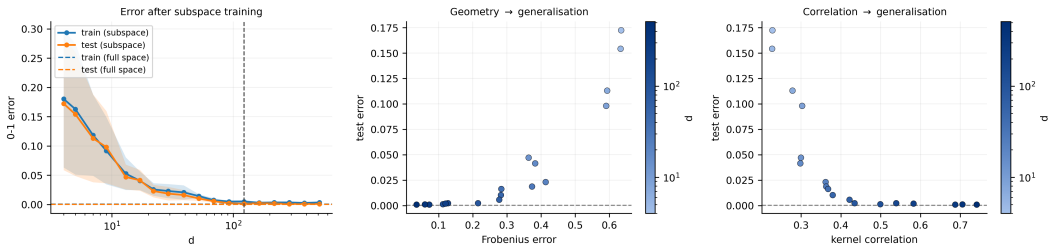


Figure 1: **Subspace training recovers MNIST generalisation at low dimension** ($m = 1024$, $N = 512$, $D = 804{,}865$; mean ± 1 std over five projection seeds). Subspace training drives both train and test 0–1 error to zero at $d \approx 40$, far below the ambient parameter dimension D. The scatter plots show that test error decreases as the projected kernel approaches the full-space kernel, linking task performance to kernel geometry.

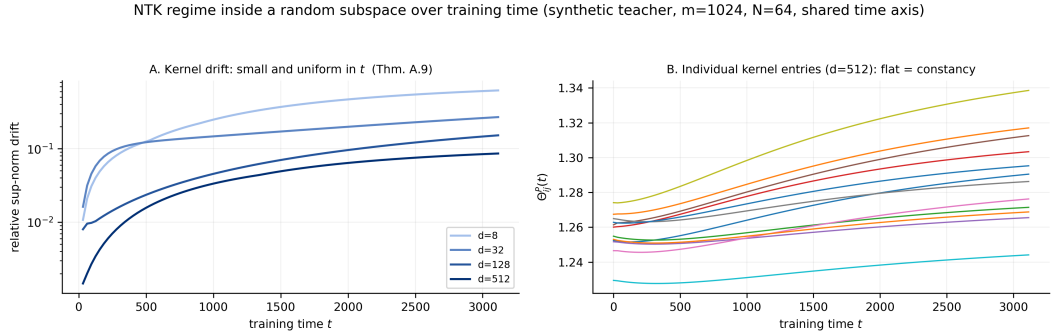


Figure 2: **The projected kernel is nearly constant along training** (synthetic teacher, $m = 1024$, $N = 64$). *A:* relative kernel drift remains small and decreases with d. *B:* individual entries $\Theta^P_{ij}(t)$ at $d = 512$ remain within a few percent throughout training.

Setup. We evaluate one-hidden-layer sigmoid networks in the NTK parameterisation, trained by full-batch gradient descent on squared loss, using the block-structured Gaussian subspaces described in Section 4. Experiments use MNIST 0 vs. 1, a synthetic teacher, and a 1-D regression task. We use the layer-wise learning-rate correction and kernel-time normalisation described in Appendix D.1; further details on architectures, datasets, projection sampling, optimisation, precision, and random seeds are provided in Appendix D.1.

5.1 Generalisation at Low Subspace Dimension

Figure 1 shows that subspace training recovers the MNIST 0 vs. 1 task at a dimension far below the ambient parameter count. With $m = 1024$, $N = 512$, and $D = 804{,}865$, both train and test 0–1 errors reach zero at approximately $d \approx 40$. This is only a tiny fraction of the full parameter dimension and is substantially below the kernel critical dimension $d^* \approx 123$ obtained from the kernel-approximation criterion. The scatter plots further connect task performance to kernel geometry: test error decreases as the projected kernel becomes closer to the full-space kernel and its entrywise correlation increases. Thus, for this task, successful generalisation requires substantially less dimension than would be needed to accurately approximate the entire kernel uniformly. The underlying kernel-approximation measurements are reported in Figure 4.

5.2 Kernel Stability over the Training Path

Figure 2 verifies that subspace training remains in the kernel regime on the synthetic teacher task ($m = 1024$, $N = 64$). The relative drift of $\Theta^P(t)$ decreases substantially with d, while individual kernel entries remain nearly constant throughout training. At $d = 512$, the entrywise variation is only a few percent. Thus, despite potentially substantial parameter evolution, the tangent kernel

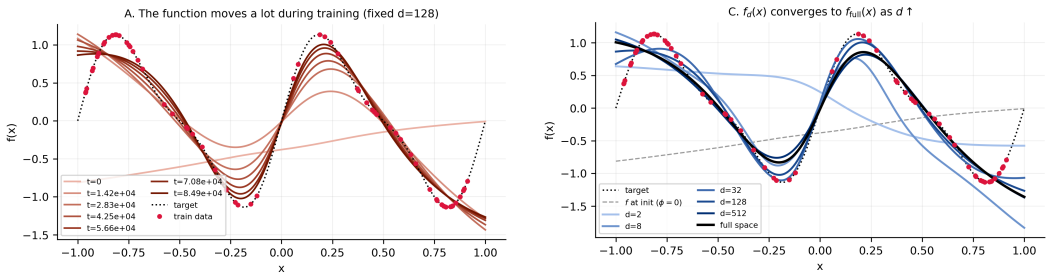


Figure 3: **Subspace training recovers the full-space function** (1-D regression, $m = 1024$, $N = 64$). *Left:* at fixed $d = 128$, the learned function evolves from initialization to the target during training. *Right:* the converged function f_d approaches the full-space solution as d grows, matching it already at $d = 128$.

remains effectively frozen along the training path, as predicted by Theorem 4.2. Additional diagnostics on function evolution and the closed-form NTK dynamics are provided in Appendix D.2.

5.3 Function-Space Recovery

A frozen kernel does not imply a frozen function. Figure 3 illustrates this on a one-dimensional regression target $y = \sin(2\pi x) + 0.3\sin(4\pi x)$ with $m = 1024$ and $N = 64$. At fixed $d = 128$, the learned function evolves substantially during training, despite the near-constant tangent kernel observed in Figure 2. Moreover, the converged subspace solution approaches the full-space solution as d increases: while $d = 2$ underfits, the solutions at $d \geq 128$ are visually indistinguishable from full-space training. Thus moderate subspaces preserve the relevant function-space solution despite containing only a small fraction of the parameters. Additional diagnostics, including kernel spectra, rank deficiency, target-alignment experiments, convergence measurements, and comparisons with the theoretical intrinsic-dimension bounds, are provided in Appendix D.2.

6 Conclusion

We showed that, in the infinite-width regime, the NTK survives restriction to a random low-dimensional subspace when $d^* \sim \varepsilon^{-2} \log N$, independently of the ambient parameter dimension D. Our lower bounds establish the optimal dimension-reduction scaling in the worst case, providing a theoretical explanation of the intrinsic dimension phenomenon: random subspaces suffice because they preserve the kernel governing training dynamics. We also identified a distinct obstruction at $d < N$: the projected kernel becomes rank-deficient, potentially leaving a component of the training error permanently frozen. Important open questions include improving the L^3 dependence, obtaining finite-width guarantees, and explaining why empirical intrinsic dimensions can be much smaller than the worst-case $\varepsilon^{-2} \log N$ prediction.

AI Use Statement

Generative AI tools were used to assist with experimental code, plotting scripts, and manuscript language editing. All theoretical results, derivations, and proofs were developed by the authors, but polished and technically reviewed by AI tools. AI-assisted code and text were reviewed and verified by the authors, who take responsibility for the final content.

References

Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252. PMLR, 2019.

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.

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.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2018.

Frithjof Gressmann, Zach Eaton-Rosen, and Carlo Luschi. Improving neural network training in low dimensional random bases. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12140–12150, 2020.

Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Edward J Hu, Yelong Shen, Phil Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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

William B Johnson, Joram Lindenstrauss, et al. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

Kasper Green Larsen and Jelani Nelson. Optimality of the johnson-lindenstrauss lemma. In *2017 IEEE 58th annual symposium on foundations of computer science (FOCS)*, pp. 633–638. IEEE, 2017.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018.

Zhe Li, Bicheng Ying, Zidong Liu, and Haibo Yang. Learning dynamics of zeroth-order optimization: A kernel perspective. In *International Conference on Machine Learning (ICML)*, 2026.

Sadhika Malladi, Alexander Wettig, Dingli Yu, Danqi Chen, and Sanjeev Arora. A kernel-based view of language model fine-tuning. In *International Conference on Machine Learning*, pp. 23610–23641. PMLR, 2023.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Ziyu Wang, Masrour Zoghi, Frank Hutter, David Matheson, and Nando De Freitas. Bayesian optimization in high dimensions via random embeddings. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

Anru R Zhang and Yuchen Zhou. On the non-asymptotic and sharp lower tail bounds of random variables. *Stat*, 9(1):e314, 2020.

A Theoretical Analysis: Projected Dimension Upper-bound

A.1 Notation and Setup

This section fixes all notation. We state every definition for general output dimension $n_L \geq 1$.

A.1.1 THE NETWORK AND ITS AMBIENT PARAMETER SPACE

We adopt the network model of Jacot et al. (2018). Fix a depth $L \geq 1$ and layer widths $n_0, n_1, \ldots, n_L$, where n_0 is the input dimension and n_L is the output dimension. The network parameters are

$$\theta = \big(W^{(1)}, b^{(1)}, \ldots, W^{(L)}, b^{(L)}\big), \qquad W^{(\ell)} \in \mathbb{R}^{n_{\ell-1} \times n_\ell}, \;\; b^{(\ell)} \in \mathbb{R}^{n_\ell},$$

initialized i.i.d. $\theta_{0,i} \sim \mathcal{N}(0,1)$; we write θ_0 for this initial value and $\theta(t)$ for the parameter trajectory under whatever training dynamics is in force (full-space or projected, introduced below). The nonlinearity $\sigma : \mathbb{R} \to \mathbb{R}$ is assumed Lipschitz, twice continuously differentiable, with bounded second derivative, throughout. We write $f_\theta : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ for the resulting network function, and fix once and for all a dataset $\mathcal{X} = \{x_1, \ldots, x_N\}$ of N distinct points, on which all kernels and gradients below are evaluated. The number of parameters in layer ℓ (weights and bias together) is

$$N_\ell \triangleq (n_{\ell-1} + 1)\, n_\ell,$$

and the total (ambient) parameter dimension is $D \triangleq \sum_{\ell=1}^{L} N_\ell$, so that $\theta \in \mathbb{R}^D$. We write $\mathrm{vec}(W^{(\ell)}, b^{(\ell)}) \in \mathbb{R}^{N_\ell}$ for the layer-ℓ parameter block flattened into a single vector, so that $\theta = \big(\mathrm{vec}(W^{(1)}, b^{(1)}), \ldots, \mathrm{vec}(W^{(L)}, b^{(L)})\big) \in \mathbb{R}^D$ under the obvious identification of $\mathbb{R}^D$ with $\mathbb{R}^{N_1} \times \cdots \times \mathbb{R}^{N_L}$.

A.1.2 GRADIENTS AND THE FULL-SPACE NTK

For $x \in \mathbb{R}^{n_0}$, the full parameter-space Jacobian of f_θ at x is

$$\nabla_\theta f_\theta(x) \in \mathbb{R}^{D \times n_L},$$

whose restriction to the layer-ℓ block we write as $\nabla_{W^{(\ell)}, b^{(\ell)}} f_\theta(x) \in \mathbb{R}^{N_\ell \times n_L}$, so that $\nabla_\theta f_\theta(x)$ is the vertical concatenation of these L blocks. The (matrix-valued) empirical Neural Tangent Kernel of f_θ on $\mathcal{X}$ is

$$\Theta(x, x') \triangleq \nabla_\theta f_\theta(x)^\top \nabla_\theta f_\theta(x') \in \mathbb{R}^{n_L \times n_L}, \qquad x, x' \in \mathcal{X},$$

following the convention of Jacot et al. (2018); we write $\Theta^{(m)}, K^{(m)}, K_\infty$, etc., as in the rest of this paper, for this same object evaluated at a given width configuration m or in the infinite-width limit, with the understanding that for $n_L = 1$ it reduces to the scalar $N \times N$ Gram matrix $K_{ij} = \langle g_i, g_j \rangle$ used in Theorem A.3 and elsewhere (Remark A.1).

A.1.3 BLOCK-STRUCTURED RANDOM SUBSPACE PROJECTION

Rather than projecting the full parameter vector θ through a single matrix, we assign each layer $\ell = 1, \ldots, L$ its own independent random projection of its own parameter block. Let

$$P^{(\ell)} \in \mathbb{R}^{N_\ell \times d_\ell}, \qquad \big(P^{(\ell)}\big)_{ab} \overset{i.i.d.}{\sim} \mathcal{N}\Big(0, \tfrac{1}{d_\ell}\Big),$$

independently across $\ell = 1, \ldots, L$ and independently of θ_0, for fixed positive integers $d_1, \ldots, d_L$. Set $d \triangleq \sum_{\ell=1}^{L} d_\ell$, the total subspace dimension. The trainable subspace parameter is $\phi = (\phi^{(1)}, \ldots, \phi^{(L)}) \in \mathbb{R}^d$, $\phi^{(\ell)} \in \mathbb{R}^{d_\ell}$, and the reparametrization is applied block-by-block:

$$\mathrm{vec}\big(W^{(\ell)}, b^{(\ell)}\big)(\phi^{(\ell)}) = P^{(\ell)} \phi^{(\ell)} + \mathrm{vec}\big(W_0^{(\ell)}, b_0^{(\ell)}\big), \qquad \phi^{(\ell)}(0) = 0, \quad \ell = 1, \ldots, L.$$

Throughout, $\theta(\phi)$ denotes the full parameter vector assembled from all L blocks via this map, and $\theta_0 = \theta(0)$ its value at $\phi = 0$; since each block is initialized at $\phi^{(\ell)}(0) = 0$, this θ_0 coincides with the i.i.d. $\mathcal{N}(0,1)$ initialization of Section 1.1, as required.

A.1.4 PROJECTED GRADIENT FLOW AND LAYER-WISE LEARNING RATES

Given the squared-error loss $\mathcal{L}(\theta) = \frac{1}{2} \sum_{i=1}^{N} \big(f_\theta(x_i) - y_i\big)^2$ (general cost functionals are treated in Remark A.16), the subspace parameters evolve under the projected gradient flow

$$\partial_t \phi^{(\ell)}(t) = -\big(P^{(\ell)}\big)^\top \nabla_{W^{(\ell)}, b^{(\ell)}} \mathcal{L}\big(\theta(\phi(t))\big), \qquad \ell = 1, \ldots, L, \tag{1}$$

equivalently, in terms of the ambient parameters restricted to layer ℓ,

$$\partial_t \operatorname{vec}\big(W^{(\ell)}, b^{(\ell)}\big)(t) = -\widetilde{P}^{(\ell)} \nabla_{W^{(\ell)},b^{(\ell)}} \mathcal{L}\big(\theta(t)\big), \qquad \widetilde{P}^{(\ell)} \triangleq P^{(\ell)}\big(P^{(\ell)}\big)^\top \in \mathbb{R}^{N_\ell \times N_\ell}.$$

We define the *block-projected NTK* at time t by

$$\Theta^P(t)\big|_{x,x'} = \sum_{\ell=1}^{L} \Big(P^{(\ell)\top} \nabla_{W^{(\ell)},b^{(\ell)}} f_{\theta(t)}(x)\Big)^\top \Big(P^{(\ell)\top} \nabla_{W^{(\ell)},b^{(\ell)}} f_{\theta(t)}(x')\Big) \in \mathbb{R}^{n_L \times n_L},$$

and, specializing to the dataset $\mathcal{X}$, write $\Theta^P_{ij}(t)$ for this kernel evaluated at (x_i, x_j); for $n_L = 1$ this is the scalar

$$\Theta^P_{ij}(t) = \sum_{\ell=1}^{L} \Big\langle \big(P^{(\ell)}\big)^\top \nabla_{W^{(\ell)},b^{(\ell)}} f_{\theta(t)}(x_i),\ \big(P^{(\ell)}\big)^\top \nabla_{W^{(\ell)},b^{(\ell)}} f_{\theta(t)}(x_j) \Big\rangle$$

used throughout Section A.3 and Section A.4 below. In either case, Θ^P is the inner product of the projected per-layer gradient blocks, summed over layers.

Remark A.1 (Scalar kernel as the case $n_L = 1$). Theorem A.3 and the lemmas of Section A.3 are stated with a scalar kernel $K_{ij} = \langle g_i, g_j \rangle$. By Remark A.16, this is exactly the $n_L = 1$ case of the matrix-valued kernels $\Theta(x, x')$ and $\Theta^P(t)$ defined above; nothing in the proofs of Sections A.3 and A.4 depends on this specialization, and the general-n_L statements are recovered by replacing every scalar product $\langle \cdot, \cdot \rangle$ above with the corresponding $n_L \times n_L$ matrix product, applied identically to each of the n_L output coordinates.

A.2 Stage I: Closeness at Initialization

In this section we establish that the block-projected network agrees with its unprojected counterpart at initialization, in a precise probabilistic sense. The argument rests on the Johnson–Lindenstrauss lemma, applied not to the ambient gradient vector as a whole but separately to each layer's gradient block, exploiting the block-diagonal structure of P: since the projected kernel Θ^P is by construction a sum of L per-layer inner products, the Johnson–Lindenstrauss distortion incurred at each layer accumulates additively rather than compounding, and a union bound over the L independent projections converts the L separate high-probability guarantees into a single joint one.

The resulting statement, Theorem A.3 below, shows that conditional on any fixed realization of the initial parameters θ_0, the projected and unprojected kernels agree up to an additive error of order εM_m with probability at least $1 - \delta$ over the randomness of the projections alone, provided each block dimension d_ℓ exceeds the corresponding Johnson–Lindenstrauss threshold; letting the network width grow then yields the same guarantee, in the sense of $\liminf$ probability, for the projected kernel against the infinite-width limit K_∞ of the standard (unprojected) Neural Tangent Kernel.

Lemma A.2 (Inner-Product Johnson–Lindenstrauss)**.** *Let $V \subset \mathbb{R}^N$ be a finite set of n vectors. Let $Q \in \mathbb{R}^{N \times k}$ have independent entries $Q_{ab} \sim \mathcal{N}(0, 1/k)$. For any $\varepsilon \in (0,1)$, $\delta \in (0,1)$, if $k \geq C_{\mathrm{JL}}\, \varepsilon^{-2} \log(n/\delta)$ for a universal constant $C_{\mathrm{JL}} > 0$, then with probability at least $1 - \delta$, simultaneously for all $u, v \in V$,*

$$\big|\langle Q^\top u, Q^\top v \rangle - \langle u, v \rangle\big| \leq \varepsilon \max\big(\|u\|_2^2, \|v\|_2^2\big).$$

Proof. We deduce the inner-product form from the distance form of Lemma 4.1. Apply Lemma 4.1 to the finite point set $V \cup \{0\} \cup \{u + v : u, v \in V\}$, which has at most $1 + n + n^2 = O(n^2)$ points; choosing the target dimension as $k \geq C_{\mathrm{JL}} \varepsilon^{-2} \log(n^2/\delta) = 2C_{\mathrm{JL}} \varepsilon^{-2} \log(n/\delta)$ (absorbing the factor of 2 into the universal constant C_{JL}) guarantees, with probability at least $1 - \delta$, that the distance bound of Lemma 4.1 holds simultaneously for every pair of points drawn from this enlarged set. In particular, taking Q in place of P, on this event we have simultaneously for all $u, v \in V$:

$$(1-\varepsilon)\|u\|_2^2 \leq \|Qu\|_2^2 \leq (1+\varepsilon)\|u\|_2^2, \qquad (1-\varepsilon)\|v\|_2^2 \leq \|Qv\|_2^2 \leq (1+\varepsilon)\|v\|_2^2,$$

(applying the distance bound to the pair $(u, 0)$ and $(v, 0)$ respectively), and

$$(1-\varepsilon)\|u+v\|_2^2 \leq \|Qu + Qv\|_2^2 \leq (1+\varepsilon)\|u+v\|_2^2$$

(applying it to the pair $(u+v, 0)$). By the polarization identity, $\langle u, v\rangle = \frac{1}{2}\big(\|u+v\|_2^2 - \|u\|_2^2 - \|v\|_2^2\big)$, and likewise $\langle Qu, Qv\rangle = \frac{1}{2}\big(\|Qu+Qv\|_2^2 - \|Qu\|_2^2 - \|Qv\|_2^2\big)$. Subtracting,

$$\langle Qu, Qv\rangle - \langle u, v\rangle = \tfrac{1}{2}\Big[\big(\|Qu+Qv\|_2^2 - \|u+v\|_2^2\big) - \big(\|Qu\|_2^2 - \|u\|_2^2\big) - \big(\|Qv\|_2^2 - \|v\|_2^2\big)\Big].$$

Each of the three bracketed terms is bounded in absolute value by ε times the squared norm of the corresponding point, by the three displayed inequalities above: $\varepsilon\|u+v\|_2^2$, $\varepsilon\|u\|_2^2$, and $\varepsilon\|v\|_2^2$ respectively. Using $\|u+v\|_2^2 \le 2\|u\|_2^2 + 2\|v\|_2^2 \le 4\max(\|u\|_2^2, \|v\|_2^2)$, the triangle inequality gives

$$\big|\langle Qu, Qv\rangle - \langle u, v\rangle\big| \;\le\; \tfrac{1}{2}\varepsilon\big(4\max(\|u\|_2^2, \|v\|_2^2) + \|u\|_2^2 + \|v\|_2^2\big) \;\le\; 3\varepsilon\,\max\big(\|u\|_2^2, \|v\|_2^2\big).$$

Replacing ε by $\varepsilon/3$ throughout (equivalently, absorbing the constant factor 3 into C_{JL}, since the required dimension $k = O(\varepsilon^{-2}\log(n/\delta))$ only changes by this same constant factor under $\varepsilon \mapsto \varepsilon/3$) yields the claimed bound $\big|\langle Qu, Qv\rangle - \langle u, v\rangle\big| \le \varepsilon\max(\|u\|_2^2, \|v\|_2^2)$ for all $u, v \in V$ simultaneously, with probability at least $1-\delta$, completing the proof. □

Theorem A.3 (Block-Projected NTK Approximation at Initialization)**.** *Let $\mathcal{X} = \{x_1, \dots, x_N\}$ be fixed. For each network width configuration m (indexing $n_1, \dots, n_{L-1} \to \infty$), let $g_i^{(m),(\ell)} \triangleq \nabla_{W^{(\ell)}, b^{(\ell)}} f_{\theta_0}^{(m)}(x_i) \in \mathbb{R}^{N_\ell^{(m)}}$ denote the layer-ℓ gradient block at initialization, and*

$$g_i^{(m)} \triangleq \Big(g_i^{(m),(1)}, \dots, g_i^{(m),(L)}\Big)$$

the concatenated gradient, with empirical NTK $K_{ij}^{(m)} = \langle g_i^{(m)}, g_j^{(m)}\rangle$. Let $P_m^{(1)}, \dots, P_m^{(L)}$ be independent block projections as above, each with its own fixed dimension d_ℓ, independent of θ_0. Define the projected kernel $K_{P,ij}^{(m)} \triangleq \sum_{\ell=1}^{L} \big\langle (P_m^{(\ell)})^\top g_i^{(m),(\ell)}, (P_m^{(\ell)})^\top g_j^{(m),(\ell)}\big\rangle$. Fix $\varepsilon, \delta \in (0,1)$ and an allocation $\varepsilon_1, \dots, \varepsilon_L > 0$ with $\sum_{\ell=1}^{L} \varepsilon_\ell = \varepsilon$, and $\delta_1, \dots, \delta_L > 0$ with $\sum_{\ell=1}^{L} \delta_\ell = \delta$. Suppose

$$d_\ell \;\ge\; C_{\mathrm{JL}}\,\varepsilon_\ell^{-2}\log(N/\delta_\ell), \qquad \ell = 1, \dots, L. \tag{2}$$

Then, for every m,

$$\mathbb{P}\Big(\max_{1\le i,j\le N} \big|(K_P^{(m)})_{ij} - (K^{(m)})_{ij}\big| \le \varepsilon\, M_m \;\Big|\; \theta_0\Big) \;\ge\; 1-\delta,$$

where $M_m \triangleq \max_{1\le i\le N} \big\|g_i^{(m)}\big\|_2^2$ and the probability is over the randomness of $(P_m^{(1)}, \dots, P_m^{(L)})$ only. Consequently, if $K^{(m)} \xrightarrow[m\to\infty]{\mathbb{P}} K_\infty$ entrywise (so that $M_m \to M_\infty$ in probability, with $M_\infty < \infty$), then for every $\eta > 0$,

$$\liminf_{m\to\infty}\; \mathbb{P}\Big(\max_{1\le i,j\le N} \big|(K_P^{(m)})_{ij} - (K_\infty)_{ij}\big| \le \varepsilon M_\infty + \eta\Big) \;\ge\; 1-\delta. \tag{3}$$

Proof. Condition on θ_0, so that $g_1^{(m)}, \dots, g_N^{(m)}$ and their layer blocks are deterministic. Fix ℓ and apply Lemma A.2 to $V_\ell = \{g_1^{(m),(\ell)}, \dots, g_N^{(m),(\ell)}\} \subset \mathbb{R}^{N_\ell^{(m)}}$ with $Q = P_m^{(\ell)}$, $k = d_\ell$, $\varepsilon = \varepsilon_\ell$, $\delta = \delta_\ell$: since (2) holds, with probability at least $1-\delta_\ell$ over $P_m^{(\ell)}$ alone,

$$\begin{aligned}\max_{1\le i,j\le N} \Big|\big\langle (P_m^{(\ell)})^\top g_i^{(m),(\ell)}, (P_m^{(\ell)})^\top g_j^{(m),(\ell)}\big\rangle - \big\langle g_i^{(m),(\ell)}, g_j^{(m),(\ell)}\big\rangle\Big| \;&\le\; \varepsilon_\ell \max_{1\le i\le N} \big\|g_i^{(m),(\ell)}\big\|_2^2 \\ &\le\; \varepsilon_\ell M_m,\end{aligned}$$

using $\big\|g_i^{(m),(\ell)}\big\|_2 \le \big\|g_i^{(m)}\big\|_2$. By independence of $P_m^{(1)}, \dots, P_m^{(L)}$, a union bound over $\ell = 1, \dots, L$ shows that, with probability at least $1 - \sum_{\ell=1}^{L}\delta_\ell = 1-\delta$ jointly, all L of these bounds hold simultaneously. On this event, for every i, j,

$$\big|(K_P^{(m)})_{ij} - (K^{(m)})_{ij}\big| = \Big|\sum_{\ell=1}^{L}\Big(\big\langle (P_m^{(\ell)})^\top g_i^{(m),(\ell)}, (P_m^{(\ell)})^\top g_j^{(m),(\ell)}\big\rangle - \big\langle g_i^{(m),(\ell)}, g_j^{(m),(\ell)}\big\rangle\Big)\Big|$$

$$\leq \sum_{\ell=1}^{L} \varepsilon_\ell M_m$$
$$= \varepsilon M_m,$$

which proves the first claim. We now establish (3). Fix $\eta > 0$. For every m and every pair i, j, the triangle inequality gives

$$\big|(K_P^{(m)})_{ij} - (K_\infty)_{ij}\big| \;\leq\; \big|(K_P^{(m)})_{ij} - (K^{(m)})_{ij}\big| \;+\; \big|(K^{(m)})_{ij} - (K_\infty)_{ij}\big|,$$

and hence, taking the maximum over i, j,

$$\max_{1\leq i,j\leq N} \big|(K_P^{(m)})_{ij} - (K_\infty)_{ij}\big| \;\leq\; \max_{1\leq i,j\leq N} \big|(K_P^{(m)})_{ij} - (K^{(m)})_{ij}\big| + \max_{1\leq i,j\leq N} \big|(K^{(m)})_{ij} - (K_\infty)_{ij}\big|.$$

Denote the two terms on the right by A_m and B_m respectively. By the first part of the proof, $\mathbb{P}(A_m \leq \varepsilon M_m \mid \theta_0) \geq 1 - \delta$ for every m, hence also unconditionally $\mathbb{P}(A_m \leq \varepsilon M_m) \geq 1 - \delta$ by the tower rule (averaging over θ_0). By hypothesis, $K^{(m)} \to K_\infty$ entrywise in probability as $m \to \infty$, and since the maximum in B_m ranges over the fixed finite index set $\{1, \ldots, N\}^2$, this entrywise convergence upgrades to $B_m \xrightarrow{\mathbb{P}} 0$ (a finite union of vanishing-probability bad events is itself vanishing). Likewise $M_m \to M_\infty$ in probability by hypothesis, so $\varepsilon M_m \to \varepsilon M_\infty$ in probability, and in particular $\mathbb{P}(\varepsilon M_m \leq \varepsilon M_\infty + \eta/2) \to 1$.

Fix $\eta > 0$ and consider the event

$$G_m := \big\{A_m \leq \varepsilon M_m\big\} \cap \big\{\varepsilon M_m \leq \varepsilon M_\infty + \eta/2\big\} \cap \big\{B_m \leq \eta/2\big\}.$$

On G_m, combining the displayed triangle inequality with $A_m \leq \varepsilon M_m \leq \varepsilon M_\infty + \eta/2$ and $B_m \leq \eta/2$ gives

$$\max_{i,j} \big|(K_P^{(m)})_{ij} - (K_\infty)_{ij}\big| \;\leq\; (\varepsilon M_\infty + \eta/2) + \eta/2 = \varepsilon M_\infty + \eta,$$

so G_m is contained in the event appearing in (3). It remains to lower-bound $\mathbb{P}(G_m)$. By a union bound on the complements,

$$\mathbb{P}(G_m) \;\geq\; 1 - \mathbb{P}\big(A_m > \varepsilon M_m\big) - \mathbb{P}\big(\varepsilon M_m > \varepsilon M_\infty + \eta/2\big) - \mathbb{P}\big(B_m > \eta/2\big).$$

The first term is at most δ for every m, as established above. The second and third terms tend to 0 as $m \to \infty$, by the convergence in probability of M_m and B_m noted above. Taking $\liminf_{m\to\infty}$ of both sides therefore gives

$$\liminf_{m\to\infty} \mathbb{P}(G_m) \;\geq\; 1 - \delta - 0 - 0 = 1 - \delta,$$

and since G_m is contained in the event of interest for every m,

$$\liminf_{m\to\infty} \mathbb{P}\Big(\max_{1\leq i,j\leq N} \big|(K_P^{(m)})_{ij} - (K_\infty)_{ij}\big| \leq \varepsilon M_\infty + \eta\Big) \;\geq\; \liminf_{m\to\infty} \mathbb{P}(G_m) \;\geq\; 1 - \delta,$$

which is exactly (3). As $\eta > 0$ was arbitrary, this completes the proof. □

Remark A.4. The bound (2) is the source of the dimension requirement on each block. Summing over ℓ with the symmetric choice $\varepsilon_\ell = \varepsilon/L$, $\delta_\ell = \delta/L$ gives the aggregate requirement $d = \sum_{\ell=1}^{L} d_\ell \geq C_{\mathrm{JL}}\, L^3\, \varepsilon^{-2} \log(NL/\delta)$, used in Theorem A.14 below. Any other allocation summing to (ε, δ) is equally valid; the symmetric choice is not optimized, it is chosen for legibility.

A.3 Stage II: Uniform-in-Time Stability of the Projected Kernel

We establish that $\Theta_{ij}^P(t)$ remains close to $K_\infty(i, j)$ uniformly over all $t \geq 0$, in the same width-growth regime as Stage I: along the sequence of width configurations m indexing $n_1^{(m)}, \ldots, n_{L-1}^{(m)} \to \infty$, write $\underline{n}^{(m)} \triangleq \min_{1\leq \ell\leq L-1} n_\ell^{(m)}$ for the bottleneck hidden width at configuration m, with the subspace dimensions $d_1, \ldots, d_L$ held fixed throughout. Unlike the ambient parameters $\theta(t)$, whose displacement from θ_0 is not controlled under this projection (Remark A.5), the subspace coordinates $\phi(t)$ are shown below to remain, with probability tending to one as $m \to \infty$, inside a fixed bounded ball $\{\|\phi\|_2 \leq R^*\}$ whose radius R^* depends only on quantities at $t = 0$ and not on m

or t; it is this confinement, together with the $O((\underline{n}^{(m)})^{-1/2})$ uniform Hessian bound of the NTK parametrization, that plays the role of the lazy-training regime here. The argument proceeds in two steps: (i) we establish that the gradient vectors $h_i^{(\ell)}(t)$ stay close to their initial values $h_i^{(\ell)}(0)$ at rate $O((\underline{n}^{(m)})^{-1/2})$, uniformly for as long as $\phi(t)$ remains in this ball; and (ii) an exact algebraic decomposition separates the total kernel error into the JL initialization term (controlled by d, already handled in Stage I) and gradient-drift terms (controlled by $\underline{n}^{(m)}$, handled here).

As discussed earlier, the residual dynamics take the form

$$\dot{r}(t) = -\Theta^P(t)\, r(t), \qquad r(t) := \big(f_{\theta_P(t)}(x_i) - y_i\big)_{i\in[N]} \in \mathbb{R}^N,$$

and the subspace coordinates evolve as

$$\dot{\phi}(t) = -P^\top J(\theta_P(t))\, r(t) = -\sum_{i=1}^{N} r_i(t)\, h_i(t), \tag{4}$$

where $J(\theta) \triangleq \big[g_1(\theta) \;\cdots\; g_N(\theta)\big] \in \mathbb{R}^{D\times N}$ is the ambient Jacobian with i-th column $g_i(\theta) \triangleq \nabla_\theta f_\theta(x_i)$ (the $n_L = 1$ case), $P \triangleq \mathrm{blkdiag}\big(P^{(1)}, \dots, P^{(L)}\big) \in \mathbb{R}^{D\times d}$ is the block-diagonal assembly of the per-layer projections, and $h_i(t) \triangleq P^\top g_i(\theta_P(t)) = \big(h_i^{(1)}(t), \dots, h_i^{(L)}(t)\big) \in \mathbb{R}^d$, with $h_i^{(\ell)}(t) \triangleq (P^{(\ell)})^\top g_i^{(\ell)}(\theta_P(t)) \in \mathbb{R}^{d_\ell}$ the layer-ℓ block of the stacked projected gradient, is the projected gradient appearing in other sections.

Remark A.5 (Ambient displacement is not controlled). Because $P^{(\ell)}$ has entries $\mathcal{N}(0, 1/d_\ell)$, the map $\phi^{(\ell)} \mapsto P^{(\ell)}\phi^{(\ell)}$ satisfies $\mathbb{E}\|P^{(\ell)}\phi^{(\ell)}\|_2^2 = (N_\ell/d_\ell)\|\phi^{(\ell)}\|_2^2$, which is unbounded as $N_\ell \to \infty$ with d_ℓ fixed. Hence $\|\theta_P(t) - \theta_0\|_2$ need not stay $O(1)$, in contrast to ordinary (unprojected) lazy training. The confinement result of Lemma A.10 below is stated for the *subspace* coordinate $\phi(t)$, not the ambient parameter, and this is what the Hessian bound of Lemma A.7 is stated relative to.

A.3.1 Bootstrap Radius and Stopping Time

Fix $\lambda_0 > 0$ such that $\lambda_{\min}(K_\infty) \geq 2\lambda_0$, and set

$$H_0 \triangleq \max_{1\leq i\leq N,\ 1\leq \ell\leq L} \|h_i^{(\ell)}(0)\|_2, \qquad R^* \triangleq \frac{8\sqrt{N}\,\|r(0)\|_2\, H_0}{\lambda_0}.$$

Both H_0 and R^* are determined entirely by the (random, but m- and t-independent) data θ_0, P at initialization, via $r(0) = (f_{\theta_0}(x_i) - y_i)_i$ and $h_i^{(\ell)}(0) = (P^{(\ell)})^\top g_i^{(\ell)}(\theta_0)$; in particular R^* is fixed before the dynamics (4) are run, and does not depend on any property of the trajectory $\phi(\cdot)$ for $t > 0$.

Definition A.6 (Joint Stopping Time)**.**

$$T_* \triangleq \inf\Big\{t \geq 0 \,:\, \lambda_{\min}\big(\Theta^P(t)\big) < \tfrac{\lambda_0}{2} \;\text{ or }\; \|\phi(t)\|_2 > R^*\Big\}.$$

By definition of the infimum, both $\lambda_{\min}(\Theta^P(t)) \geq \lambda_0/2$ and $\|\phi(t)\|_2 \leq R^*$ hold for every $t \in [0, T_*)$; this is the *bootstrap region* referenced in the Hessian bounds below, and Lemma A.10 shows $T_* = \infty$ with probability tending to one as $m \to \infty$.

A.3.2 Uniform Hessian operator norm in NTK parametrization

The following structural lemma is the key fact from the NTK parametrization that makes the entire argument work. Unlike the standard small-motion estimates (which bound parameter changes), this lemma bounds the network's second derivative uniformly over the fixed bootstrap ball $\{\|\phi\|_2 \leq R^*\}$ fixed in Section A.3.1.

Lemma A.7 (Uniform Hessian Operator Norm)**.** *Under the NTK parametrization of Jacot et al. (2018), for every layer $\ell \in [L]$, every data point $x_i \in \mathcal{X}$, every width configuration m, and every $\phi \in \mathbb{R}^d$ with $\|\phi\|_2 \leq R^*$, writing $\theta_P \triangleq \theta_0 + P\phi$, with probability tending to one as $m \to \infty$,*

$$\Big\|H_i^{(\ell)}(\theta_P)\Big\|_{\mathrm{op}} \;\leq\; \frac{C_H}{\sqrt{\underline{n}^{(m)}}},$$

where $H_i^{(\ell)}(\theta) := \nabla^2_{\mathrm{vec}(W^{(\ell)},b^{(\ell)})} f_\theta(x_i) \in \mathbb{R}^{N_\ell \times N_\ell}$ *is the layer-*ℓ *diagonal Hessian block and* $C_H > 0$ *is an absolute constant depending only on* L, R^*, *and the Lipschitz and second-derivative bounds on* σ.

Proof. Fix ℓ, i, and ϕ with $\|\phi\|_2 \le R^*$, and write $\theta_P = \theta_0 + P\phi$. Since $\alpha^{(\ell-1)}(x_i;\theta_P)$ does not depend on $(W^{(\ell)}, b^{(\ell)})$, and $z^{(\ell)}(x_i;\theta_P) = n_{\ell-1}^{-1/2} W^{(\ell)\top}\alpha^{(\ell-1)}(x_i;\theta_P) + \beta b^{(\ell)}$ is affine in $(W^{(\ell)}, b^{(\ell)})$, the chain rule gives

$$H_i^{(\ell)}(\theta_P) = J_z^\top M^{(\ell)} J_z, \qquad J_z \triangleq \frac{\partial z^{(\ell)}(x_i;\theta_P)}{\partial \,\mathrm{vec}(W^{(\ell)}, b^{(\ell)})} \in \mathbb{R}^{N_\ell \times n_\ell}, \quad M^{(\ell)} \triangleq \nabla^2_{z^{(\ell)}} f_{\theta_P}(x_i) \in \mathbb{R}^{n_\ell \times n_\ell},$$

where $M^{(\ell)}$ is the Hessian of the scalar output with respect to the layer-ℓ pre-activation, holding layers $1,\dots,\ell-1$ fixed. By construction J_z is block-diagonal with n_ℓ identical blocks, each equal to $\big(n_{\ell-1}^{-1/2}\alpha^{(\ell-1)}(x_i;\theta_P),\, \beta\big) \in \mathbb{R}^{n_{\ell-1}+1}$, so

$$\|J_z\|_{\mathrm{op}}^2 = n_{\ell-1}^{-1}\|\alpha^{(\ell-1)}(x_i;\theta_P)\|_2^2 + \beta^2.$$

Since $\|\phi\|_2 \le R^*$ is a fixed, m-independent constraint, a standard uniform concentration argument over the (compact) ball $\{\|\phi\|_2 \le R^*\}$ — a covering-number/Lipschitz argument in ϕ, using that $\alpha^{(\ell-1)}(x_i;\cdot)$ is Lipschitz in θ_P with constant $O(1)$ by the assumed Lipschitz bound on σ — shows $n_{\ell-1}^{-1}\|\alpha^{(\ell-1)}(x_i;\theta_P)\|_2^2 = O(1)$ uniformly over this ball, with probability tending to one as $m \to \infty$; hence $\|J_z\|_{\mathrm{op}}^2 = O(1)$ on this event. It remains to bound $\|M^{(\ell)}\|_{\mathrm{op}}$. Each entry of $M^{(\ell)}$ is a sum, over paths from layer ℓ to the output, of products of: one factor of σ'' evaluated at a pre-activation (bounded by $\|\sigma''\|_\infty$), and, for each of the $L-\ell$ remaining layers traversed, a factor of $n_{\ell'}^{-1/2}$ from the NTK-parametrization scaling together with an $O(1)$-normalized weight/backprop term, controlled as above uniformly on the same ball. Collecting these factors, $\|M^{(\ell)}\|_{\mathrm{op}} = O\big((\underline{n}^{(m)})^{-1/2}\big)$ on this event. Combining the two bounds gives $\|H_i^{(\ell)}(\theta_P)\|_{\mathrm{op}} \le \|J_z\|_{\mathrm{op}}^2\|M^{(\ell)}\|_{\mathrm{op}} = O\big((\underline{n}^{(m)})^{-1/2}\big)$, as claimed. □

Remark A.8. The bound $C_H/\sqrt{\underline{n}^{(m)}}$ is a consequence of the $1/\sqrt{n_\ell}$ weight scaling in the NTK parametrization together with confinement of ϕ to a *fixed, m-independent* ball, not of proximity to initialization in the ambient Euclidean norm (cf. Remark A.5).

Lemma A.9 (Projected Hessian Bound)**.** *For all* $i \in [N]$, $\ell \in [L]$, *and all* $\phi \in \mathbb{R}^d$ *with* $\|\phi\|_2 \le R^*$, *writing* $\theta_P \triangleq \theta_0 + P\phi$, *the sandwiched projected Hessian satisfies*

$$S_i^{(\ell)} := (P^{(\ell)})^\top H_i^{(\ell)}(\theta_P)\, P^{(\ell)} \in \mathbb{R}^{d_\ell \times d_\ell}, \qquad \big\|S_i^{(\ell)}\big\|_{\mathrm{op}} \le \frac{C_H}{\sqrt{\underline{n}^{(m)}}},$$

with probability tending to one as $m \to \infty$, *where* C_H *is the same constant as in Lemma A.7.*

Proof. For any unit vector $u \in \mathbb{R}^{d_\ell}$,

$$u^\top S_i^{(\ell)} u = (P^{(\ell)}u)^\top H_i^{(\ell)}(\theta_P)\,(P^{(\ell)}u).$$

By Lemma A.7, $|u^\top S_i^{(\ell)} u| \le \|H_i^{(\ell)}(\theta_P)\|_{\mathrm{op}} \cdot \|P^{(\ell)}u\|_2^2 \le \frac{C_H}{\sqrt{\underline{n}^{(m)}}} \cdot \|P^{(\ell)}u\|_2^2$. Since $P^{(\ell)}u \sim \mathcal{N}(0, \|u\|_2^2/d_\ell \cdot I_{N_\ell})$ for a unit vector u, a standard Gaussian concentration argument gives $\|P^{(\ell)}u\|_2^2 \le 2$ with probability $1 - e^{-cd_\ell}$ for an absolute constant $c > 0$. Taking a union bound over a $1/4$-net of $S^{d_\ell - 1}$ (of cardinality 9^{d_ℓ}, independent of m) and using the net-to-operator-norm comparison gives $\|S_i^{(\ell)}\|_{\mathrm{op}} \le C_H/\sqrt{\underline{n}^{(m)}}$ with probability $1 - 9^{d_\ell}e^{-cd_\ell} \to 1$ as $m \to \infty$, since d_ℓ is fixed. □

A.3.3 Joint Bootstrap Lemma

Lemma A.10 (Joint Bootstrap: Confinement, Kernel Stability, and $T_* = \infty$)**.** *There exists a width threshold* $m_0 = m_0(\lambda_0, L, N, d, C_H, \|r(0)\|_2, H_0)$ *such that for all* $m \ge m_0$, *with probability tending to one as* $m \to \infty$:

(a) $T_* = \infty$;

(b) $\|\phi(t)\|_2 \le R^*/2$ *for all* $t \ge 0$;

(c) $\sup_{t\ge 0} \left\|\Theta^P(t) - \Theta^P(0)\right\|_{\mathrm{op}} \le \dfrac{\lambda_0}{8}$;

(d) $\lambda_{\min}(\Theta^P(t)) \ge \frac{5\lambda_0}{8}$ *for all* $t \ge 0$.

Proof. Throughout, fix $t < T_*$, so that by Definition A.6, $\lambda_{\min}(\Theta^P(s)) \ge \lambda_0/2$ and $\|\phi(s)\|_2 \le R^*$ for all $s \in [0,t]$; in particular $\theta_P(s)$ lies in the bootstrap ball of Lemmas A.7–A.9 throughout $[0,t]$, so both lemmas apply at every $s \in [0,t]$.

Step 1: residual decay on $[0,T_*)$. The ODE $\dot r(s) = -\Theta^P(s) r(s)$ and $\lambda_{\min}(\Theta^P(s)) \ge \lambda_0/2$ on $[0,t]$ give

$$\frac{d}{ds}\tfrac{1}{2}\|r(s)\|_2^2 = -r(s)^\top \Theta^P(s) r(s) \le -\tfrac{\lambda_0}{2}\|r(s)\|_2^2,$$

so by Grönwall, $\|r(s)\|_2 \le \|r(0)\|_2 e^{-\lambda_0 s/2}$ for all $s \in [0,t]$.

Step 2: a priori bound on the gradient drift, via Grönwall. Set

$$\rho(s) \triangleq \max_{1\le i\le N,\, 1\le \ell\le L} \|h_i^{(\ell)}(s) - h_i^{(\ell)}(0)\|_2$$

and $\bar\rho(t) \triangleq \sup_{s\in[0,t]} \rho(s)$. From (4) and the chain rule,

$$\frac{d}{ds} h_i^{(\ell)}(s) = -\sum_{\ell'=1}^{L} S_i^{(\ell,\ell')}(s) \sum_{k=1}^{N} r_k(s)\, h_k^{(\ell')}(s), \qquad S_i^{(\ell,\ell')}(s) \triangleq (P^{(\ell)})^\top H_i^{(\ell,\ell')}(\theta_P(s)) P^{(\ell')},$$

where $H_i^{(\ell,\ell')}$ is the (off-diagonal, for $\ell \ne \ell'$) cross-layer Hessian block, which satisfies the same $C_H/\sqrt{\underline{n}^{(m)}}$ operator-norm bound as Lemma A.9 by the identical argument applied to $H_i^{(\ell,\ell')}$ in place of $H_i^{(\ell)}$. Hence, for $s \in [0,t]$, using Step 1 and $\sum_k |r_k(s)| \le \sqrt{N}\|r(s)\|_2$,

$$\left\| \frac{d}{ds} h_i^{(\ell)}(s) \right\|_2 \le \frac{LC_H}{\sqrt{\underline{n}^{(m)}}} \sqrt{N}\, \|r(0)\|_2 e^{-\lambda_0 s/2} \big(H_0 + \bar\rho(t)\big).$$

Integrating over $[0,t]$ and taking the max over i,ℓ,

$$\rho(t) \;\le\; \frac{2LC_H\sqrt{N}\,\|r(0)\|_2}{\lambda_0\sqrt{\underline{n}^{(m)}}} \big(H_0 + \bar\rho(t)\big) \;=:\; \kappa_m \big(H_0 + \bar\rho(t)\big).$$

Since this holds for every $t < T_*$ with the same right-hand side (which depends on t only through $\bar\rho(t)$), taking sup over t gives $\bar\rho(t) \le \kappa_m(H_0 + \bar\rho(t))$, i.e., whenever $\kappa_m \le 1/2$,

$$\bar\rho(t) \;\le\; \frac{\kappa_m H_0}{1-\kappa_m} \;\le\; 2\kappa_m H_0 \;=:\; \frac{C_3}{\sqrt{\underline{n}^{(m)}}}, \qquad C_3 \triangleq \frac{4LC_H\sqrt{N}\,\|r(0)\|_2 H_0}{\lambda_0}.$$

Let m_2 be such that $\kappa_m \le 1/2$ and $C_3/\sqrt{\underline{n}^{(m)}} \le H_0$ for all $m \ge m_2$ (both hold for $\underline{n}^{(m)}$ large enough, since C_3 does not depend on m). Then for $m \ge m_2$ and all $t < T_*$: $\rho(t) \le C_3/\sqrt{\underline{n}^{(m)}} \le H_0$.

Step 3: strict confinement of ϕ. Integrating (4) and using Step 1–2,

$$\begin{aligned}
\|\phi(t)\|_2 &\le \int_0^t \sqrt{N}\|r(s)\|_2 \max_{k,\ell} \|h_k^{(\ell)}(s)\|_2\, ds \\
&\le \sqrt{N}\|r(0)\|_2 (H_0 + \rho(t)) \int_0^\infty e^{-\lambda_0 s/2} ds \\
&\le \frac{2\sqrt{N}\|r(0)\|_2}{\lambda_0} \cdot 2H_0 = \frac{R^*}{2},
\end{aligned}$$

for $m \geq m_2$, using $\rho(t) \leq H_0$ from Step 2. This holds for every $t < T_*$.

Step 4: strict kernel stability. For each i, j, ℓ, by Step 2 (for $m \geq m_2$, so $\rho(t) \leq H_0$),

$$\left|\langle h_i^{(\ell)}(t), h_j^{(\ell)}(t)\rangle - \langle h_i^{(\ell)}(0), h_j^{(\ell)}(0)\rangle\right| \leq \rho(t)\big(2H_0 + \rho(t)\big) \leq \frac{C_3}{\sqrt{\underline{n}^{(m)}}} \cdot 3H_0 \triangleq \frac{C_4}{\sqrt{\underline{n}^{(m)}}}.$$

Summing over $\ell \in [L]$ and bounding the operator norm by N times the entrywise max (Frobenius bound),

$$\left\|\Theta^P(t) - \Theta^P(0)\right\|_{\mathrm{op}} \leq N \max_{i,j} \left| \sum_{\ell=1}^{L} \left(\left| \langle h_i^{(\ell)}(t), h_j^{(\ell)}(t)\rangle - \langle h_i^{(\ell)}(0), h_j^{(\ell)}(0)\rangle \right| \right) \right| \leq \frac{LC_4 N}{\sqrt{\underline{n}^{(m)}}} =: \frac{L_1}{\sqrt{\underline{n}^{(m)}}},$$

for every $t < T_*$. Let $m_1 \geq m_2$ be such that $L_1/\sqrt{\underline{n}^{(m)}} \leq \lambda_0/8$ for $m \geq m_1$, and set $m_0 \triangleq m_1$.

Step 5: closing the loop by continuity. For $m \geq m_0$, Steps 3–4 give, for every $t < T_*$: $\|\phi(t)\|_2 \leq R^*/2$ and (using $\lambda_{\min}(\Theta^P(0)) \geq 3\lambda_0/4$ from Stage I together with Weyl's inequality and Step 4) $\lambda_{\min}(\Theta^P(t)) \geq 3\lambda_0/4 - \lambda_0/8 = 5\lambda_0/8$. Suppose for contradiction $T_* < \infty$. The maps $t \mapsto \phi(t)$ and $t \mapsto \lambda_{\min}(\Theta^P(t))$ are continuous (the right-hand side of (4) is locally Lipschitz in ϕ), so the bounds just derived on $[0, T_*)$ pass to the limit at $t = T_*$: $\|\phi(T_*)\|_2 \leq R^*/2 < R^*$ and $\lambda_{\min}(\Theta^P(T_*)) \geq 5\lambda_0/8 > \lambda_0/2$. But $T_* < \infty$ forces, by definition of the infimum, that at least one of $\lambda_{\min}(\Theta^P(T_*)) \leq \lambda_0/2$ or $\|\phi(T_*)\|_2 \geq R^*$ holds — contradicting both strict bounds just obtained. Hence $T_* = \infty$, proving (a); and (b)–(d) then hold for all $t \geq 0$ by Steps 3–4 applied with $t \to \infty$. □

A.3.4 Algebraic Decomposition of Kernel Drift

With the stopping-time argument closed, we now bound the total error $\Theta_{ij}^P(t) - K_\infty(i,j)$ via an exact decomposition.

Lemma A.11 (Exact Kernel Error Decomposition). *Let $u_i^{(\ell)}(t) := g_i^{(\ell)}(\theta_P(t))$ and $v_i^{(\ell)} := g_i^{(\ell)}(\theta_0)$. The error decomposes as*

$$\Theta_{ij}^P(t) - K_\infty(i,j) = \underbrace{\sum_{\ell=1}^{L} v_i^{(\ell)\top}(P^{(\ell)}P^{(\ell)\top} - I_{D_\ell})v_j^{(\ell)}}_{A_{ij}} + \underbrace{\sum_{\ell=1}^{L}\left[\langle h_i^{(\ell)}(t), h_j^{(\ell)}(t)\rangle - \langle h_i^{(\ell)}(0), h_j^{(\ell)}(0)\rangle\right]}_{B_{ij}(t)}$$
$$+ \underbrace{\left(\sum_{\ell=1}^{L}\langle v_i^{(\ell)}, v_j^{(\ell)}\rangle - K_\infty(i,j)\right)}_{D_{ij}}, \tag{5}$$

where $h_i^{(\ell)}(t) = (P^{(\ell)})^\top u_i^{(\ell)}(t)$ and $h_i^{(\ell)}(0) = (P^{(\ell)})^\top v_i^{(\ell)}$.

Proof. Write

$$\Theta_{ij}^P(t) = \sum_\ell \langle h_i^{(\ell)}(t), h_j^{(\ell)}(t)\rangle = \sum_\ell \langle h_i^{(\ell)}(0), h_j^{(\ell)}(0)\rangle + B_{ij}(t).$$

Also

$$\sum_\ell \langle h_i^{(\ell)}(0), h_j^{(\ell)}(0)\rangle = \sum_\ell v_i^{(\ell)\top} P^{(\ell)} P^{(\ell)\top} v_j^{(\ell)} = \sum_\ell v_i^{(\ell)\top} v_j^{(\ell)} + A_{ij}.$$

And $\sum_\ell \langle v_i^{(\ell)}, v_j^{(\ell)}\rangle = K^{(m)}(i,j) = K_\infty(i,j) + D_{ij}$, where $D_{ij} = K^{(m)}(i,j) - K_\infty(i,j) \to 0$ in probability as $m \to \infty$ by the classical NTK convergence result Jacot et al. (2018). □

Theorem A.12 (Uniform-in-Time Stability of the Block-Projected NTK). *Fix $\varepsilon, \delta \in (0,1)$ and $T \geq 0$. Under the assumptions of Theorem A.3, and provided*

$$d_\ell \geq C_{\mathrm{II}}\, \varepsilon^{-2} \log\left(\frac{2N^2}{\delta_\ell}\right), \qquad \ell = 1, \ldots, L, \tag{6}$$

where $\delta_\ell = \delta/(2L)$ and $C_{\mathrm{II}} > 0$ is an absolute constant, then for all m sufficiently large (depending on $\lambda_0, L, N, d, \delta$),

$$\mathbb{P}\Big(\sup_{0\le t\le T}\max_{1\le i,j\le N}\big|\Theta^P_{ij}(t) - K_\infty(i,j)\big| \;\le\; \varepsilon \max_{1\le i\le N} K_\infty(i,i)\Big) \;\ge\; 1-\delta. \tag{7}$$

Proof. We bound each term in the decomposition (5).

Term A_{ij}: JL initialization error. By Stage I, with probability $\ge 1-\delta/2$, simultaneously for all i, j, ℓ,

$$\max_{i,j}|A_{ij}| \;\le\; \sum_{\ell=1}^{L}\epsilon^{(\ell)}_{\mathrm{JL}} \;\le\; \frac{\varepsilon}{2}\max_i K_\infty(i,i)$$

provided (6) holds (absorbing $M_m \to M_\infty$ as in Theorem A.3).

Term $B_{ij}(t)$: temporal kernel drift. By Lemma A.10, for

$$m \ge m_0(\lambda_0, L, N, d, C_H, \|r(0)\|_2, H_0),$$

$\sup_{t\ge 0}\big\|\Theta^P(t) - \Theta^P(0)\big\|_{\mathrm{op}} \le \lambda_0/8$ with probability tending to one as $m\to\infty$; choosing m large enough within this regime makes this $\le \varepsilon M_\infty/4$.

Term D_{ij}: finite-width NTK error. Again, from Stage I, $\big\|K^{(m)} - K_\infty\big\|_{\mathrm{op}} \le \lambda_0/2 \le \varepsilon M_\infty/4$ for $m \ge m_0$ with probability $\ge 1-\delta/4$.

Combining the three bounds via a union bound over the three events (probability $\ge 1-\delta/2-\delta/4-\delta/4 = 1-\delta$) and the triangle inequality on (5) gives $\sup_{t\ge 0}\max_{i,j}|\Theta^P_{ij}(t) - K_\infty(i,j)| \le \varepsilon M_\infty$ for all m sufficiently large. □

Corollary A.13 (Optimal Layerwise Dimension Allocation)**.** *Fix total subspace dimension $d = \sum_\ell d_\ell$. The optimal allocation minimizing the dominant JL error in Theorem A.12 is*

$$d^*_\ell = d\cdot\frac{G_\ell^{4/3}}{\sum_{k=1}^{L} G_k^{4/3}}, \qquad G_\ell := \limsup_{m\to\infty}\max_{1\le i\le N}\Big\|g_i^{(\ell)}(\theta_0)\Big\|_2,$$

yielding an overall error rate of $\mathcal{O}(d^{-1/2})$.

Proof. The dominant term in Theorem A.12 is $\sum_\ell C_{\mathrm{JL}} G_\ell^2 d_\ell^{-1/2}$ (dropping the log factors for this optimization). Minimizing over $\{d_\ell\}$ subject to $\sum_\ell d_\ell = d$ via the Lagrangian gives $d^*_\ell \propto G_\ell^{4/3}$. Substituting back gives total error $\propto d^{-1/2}(\sum_\ell G_\ell^{4/3})^{3/2} = O(d^{-1/2})$. □

A.4 Stage III: Upper-bound on the Subspace Dimension

We now combine Theorems A.3 and A.12. Recall that Theorem A.3 is conditional on θ_0 with probability over P alone, while Theorem A.12 is a joint statement over network randomness and P together; both are valid for *every* realization of the conditioning variable, so a union bound over the two corresponding bad events applies unconditionally.

Theorem A.14 (Main Result: Block-Projected NTK)**.** *Fix $T > 0$, $\varepsilon, \delta \in (0,1)$. Suppose σ is Lipschitz, twice continuously differentiable, with bounded second derivative, and that the standard (unprojected) NTK at initialization converges entrywise in probability to a deterministic kernel K_∞ as $\min_\ell n_\ell \to \infty$, with $M_\infty = \max_i K_\infty(i,i) < \infty$. Let $d_1, \dots, d_L$ be fixed positive integers satisfying, for the symmetric allocation $\varepsilon_\ell = \varepsilon/L$, $\delta_\ell = \delta'_\ell = \delta/(2L)$,*

$$d_\ell \;\ge\; C_{\mathrm{JL}}\, L^2\, \varepsilon^{-2}\log\Big(\frac{2NL}{\delta}\Big), \qquad \ell = 1, \dots, L, \tag{8}$$

equivalently $d = \sum_{\ell=1}^{L} d_\ell \ge C_{\mathrm{JL}}\, L^3\, \varepsilon^{-2}\log(2NL/\delta)$. Then, for every $\eta > 0$,

$$\liminf_{\min_\ell n_\ell\to\infty}\mathbb{P}\Big(\sup_{0\le t\le T}\max_{1\le i,j\le N}\big|\Theta^P_{ij}(t) - K_\infty(i,j)\big| \le \varepsilon M_\infty + \eta\Big) \;\ge\; 1-\delta.$$

Proof. Write the event of interest as the intersection of two events:

$$F_1 \triangleq \left\{ \max_{i,j} \left| \Theta^P_{ij}(0) - K_\infty(i,j) \right| \le \varepsilon M_\infty + \eta/2 \right\},$$
$$F_2 \triangleq \left\{ \sup_{0\le t\le T} \max_{i,j} \left| \Theta^P_{ij}(t) - \Theta^P_{ij}(0) \right| \le \eta/2 \right\}.$$

On $F_1 \cap F_2$, the triangle inequality $|\Theta^P_{ij}(t) - K_\infty(i,j)| \le |\Theta^P_{ij}(t) - \Theta^P_{ij}(0)| + |\Theta^P_{ij}(0) - K_\infty(i,j)|$ gives the bound in the statement for every $t \in [0,T]$ and every i,j, hence $F_1 \cap F_2$ is contained in the event whose probability we wish to lower-bound. It therefore suffices to lower-bound $\mathbb{P}(F_1 \cap F_2) \ge 1 - \mathbb{P}(F_1^c) - \mathbb{P}(F_2^c)$.

By Theorem A.3 with the stated allocation (valid since (8) implies (2) for $\varepsilon_\ell = \varepsilon/L$, $\delta_\ell = \delta/(2L)$), $\Theta^P_{ij}(0) = K^{(m)}_{P,ij}$ in the notation of that theorem, so $\liminf_{m\to\infty} \mathbb{P}(F_1) \ge 1 - \delta/2$, i.e. $\limsup_{m\to\infty} \mathbb{P}(F_1^c) \le \delta/2$. By Theorem A.12 with $\delta' = \delta/2$ (using $\delta'_\ell = \delta/(2L)$), $\liminf_{\min_\ell n_\ell\to\infty} \mathbb{P}(F_2) \ge 1 - \delta/2$, i.e. $\limsup_{\min_\ell n_\ell\to\infty} \mathbb{P}(F_2^c) \le \delta/2$. Hence

$$\liminf_{\min_\ell n_\ell\to\infty} \mathbb{P}(F_1 \cap F_2) \ge 1 - \limsup_{\min_\ell n_\ell\to\infty} \mathbb{P}(F_1^c) - \limsup_{\min_\ell n_\ell\to\infty} \mathbb{P}(F_2^c) \ge 1 - \delta/2 - \delta/2 = 1 - \delta,$$

which completes the proof. □

Remark A.15 (On the dimension lower bound). The bound (8) is a *sufficient* condition for the conclusion of Theorem A.14: it guarantees that block dimensions of this size suffice for the projected NTK to track K_∞ within $\varepsilon M_\infty + \eta$, uniformly over $[0,T]$, with probability at least $1-\delta$ (for all pre-fixed $\delta, \varepsilon, T > 0$). It makes no claims regarding the *necessity* or a lower-bound on the dimension of the projected subspace.

Remark A.16 (General output dimension and general cost functionals). This appendix already states $\Theta(x,x')$ and $\Theta^P(t)$ for general n_L, with the scalar kernels used in Theorem A.3 and Section A.3 identified as the $n_L = 1$ case in Remark A.1. We confirm here that this specialization costs nothing: for general n_L, every step of Jacot et al.'s proof of their Theorem 2 is carried out for the block kernel $\Theta^{(L)} \otimes I_{n_L}$, and our substitution argument operates identically on each of the n_L output coordinates. We have, however, stated Theorem A.14 for squared-error loss specifically; this restriction can be removed at no extra cost. For a general training direction $d_t \in F$ (in the sense of (Jacot et al., 2018), Section 5), Theorem 2 of that paper — and hence our Theorem A.12 — requires only that

$$\int_0^T \|d_t\|_{p^{\text{in}}}\, dt$$

stay stochastically bounded as the widths grow; this is the only place the choice of cost functional enters the proof. For squared-error loss this hypothesis is immediate, since

$$\|d_t\|_{p^{\text{in}}} = \left\| f_{\theta(t)} - f^* \right\|_{p^{\text{in}}}$$

is non-increasing along the (kernel) gradient flow once the kernel is approximately constant, hence bounded by its value at $t = 0$; for other convex cost functionals satisfying the same stochastic boundedness hypothesis, the theorem as stated applies verbatim with no further argument required, exactly as in the original NTK paper.

B A Dimension Lower-bound for Random Projections

The achievability result of Theorem A.3 shows that a Gaussian projection of dimension $d \gtrsim \varepsilon^{-2} \log N$ suffices to preserve the projected NTK to within ε at initialization, with high probability. In this section we show that the same family of random projections cannot succeed below this threshold: for any fixed pair of gradient vectors, a Gaussian projection at insufficient dimension fails to preserve their inner product with probability bounded away from zero, regardless of the ambient dimension D or the number of remaining vectors $N - 2$. No special structure on the vectors is assumed; the obstruction comes entirely from the randomness of the projection.

B.1 A Sharp Tail Bound for the Chi-Squared Distribution

The key technical input is a matching lower bound on the lower tail of the chi-squared distribution.

Lemma B.1 (Sharp lower tail for χ^2_k; (Zhang & Zhou, 2020), Corollary 3)**.** *Let $Y \sim \chi^2_k$ for an integer $k \geq 1$, and let $X = Y - k$. There exist absolute constants $c, C_\varepsilon > 0$, with C_ε depending only on $\varepsilon \in (0, 1)$, such that*

$$\mathbb{P}\big(X \leq -x\big) \;\geq\; c\,\exp\!\left(-C_\varepsilon \frac{x^2}{k}\right), \qquad 0 < x < (1-\varepsilon)k. \tag{9}$$

Proposition B.2 (Lower-tail failure of a single random projection)**.** *Fix $u \in \mathbb{R}^D \setminus \{0\}$ and let $P \in \mathbb{R}^{d\times D}$ have i.i.d. entries $P_{ij} \sim \mathcal{N}(0, 1/d)$. For every $\varepsilon \in (0, 1)$ and every $d \geq 1$,*

$$\mathbb{P}\Big(\|Pu\|_2^2 \leq (1-\varepsilon)\|u\|_2^2\Big) \;\geq\; c\,\exp\big(-C_\varepsilon\, d\,\varepsilon^2\big), \tag{10}$$

where $c, C_\varepsilon > 0$ are the constants of Lemma B.1.

Proof. Write $Pu = \big(\langle P_{1,:}, u\rangle, \ldots, \langle P_{d,:}, u\rangle\big)$. Since the rows of P are i.i.d. Gaussian vectors with entries $\mathcal{N}(0, 1/d)$, each inner product $\langle P_{k,:}, u\rangle \sim \mathcal{N}(0, \|u\|_2^2/d)$, independently across $k = 1, \ldots, d$. Hence $Y := d\|Pu\|_2^2/\|u\|_2^2 \sim \chi^2_d$. Applying Lemma B.1 with $k = d$ and $x = d\varepsilon$ — valid since $0 < d\varepsilon < (1-\varepsilon)d$ for $\varepsilon \in (0, 1/2)$, and extendable to $\varepsilon \in [1/2, 1)$ by enlarging C_ε by a direct check —

$$\mathbb{P}\big(Y \leq d(1-\varepsilon)\big) \;\geq\; c\exp\!\left(-C_\varepsilon \frac{(d\varepsilon)^2}{d}\right) = c\exp(-C_\varepsilon\, d\,\varepsilon^2).$$

Since $\{Y \leq d(1-\varepsilon)\} = \{\|Pu\|_2^2 \leq (1-\varepsilon)\|u\|_2^2\}$, the claim follows. □

B.2 Impossibility Below the Critical Dimension

Theorem B.3 (Random-projection lower bound)**.** *Fix $N \geq 2$ and let $u_1, u_2 \in \mathbb{R}^D \setminus \{0\}$ be any two vectors — no special structure is assumed on u_1, u_2 or on any remaining vectors. Let $P \in \mathbb{R}^{d\times D}$ have i.i.d. entries $P_{ij} \sim \mathcal{N}(0, 1/d)$, and fix $\varepsilon \in (0, 1)$. If*

$$d \;\leq\; \frac{\log(1/c)}{C_\varepsilon\,\varepsilon^2}, \tag{11}$$

for the constants c, C_ε of Lemma B.1, then

$$\mathbb{P}\Big(\|Pu_1\|_2^2 \leq (1-\varepsilon)\|u_1\|_2^2\Big) \;\geq\; c\,\exp(-C_\varepsilon\, d\,\varepsilon^2) \;\geq\; c^2 \;>\; 0, \tag{12}$$

a probability bounded below by an absolute constant independent of D, d, and N. Consequently, by the polarization identity used in the proof of Lemma A.2, with probability at least c^2 the projection P fails to preserve the inner product $\langle u_1, u_2\rangle$ to within accuracy $\mathcal{O}(\varepsilon)$, violating condition (2) of Theorem A.3 for this pair.

Proof. Condition (11) ensures $C_\varepsilon d\varepsilon^2 \leq \log(1/c)$, so $\exp(-C_\varepsilon d\varepsilon^2) \geq c$. Substituting into (10) gives (12). The polarization step follows the same algebraic identity as in the proof of Lemma A.2, applied in the reverse direction: failure to preserve $\|Pu_1\|_2^2$ within ε forces failure to preserve at least one of $\|Pu_1\|_2^2$, $\langle Pu_1, Pu_2\rangle$, $\|P(u_1+u_2)\|_2^2$ within a comparable accuracy. □

Remark B.4. Theorem B.3 holds for any fixed pair of vectors with no special structure, since the obstruction comes entirely from the randomness of P rather than from the geometry of the vectors. It applies directly to the family of projections used in Theorem A.3, confirming that the ε^{-2} rate in the upper bound is not improvable within this family. It does not, however, rule out the possibility that a non-Gaussian, structured, or data-dependent projection could achieve a smaller critical dimension for the same vectors; that question is addressed in Section C.

C Impossibility Results for Deterministic Projections

Theorem B.3 rules out the specific family of Gaussian random projections below the ε^{-2} threshold. We now show that the threshold is unavoidable for *any* linear map, not merely random ones, for a specific but explicitly constructible configuration of gradient vectors. This stronger statement is established via the following optimality result for Johnson–Lindenstrauss embeddings.

C.1 The Larsen–Nelson Lower Bound

Lemma C.1 (Optimality of Johnson–Lindenstrauss for linear maps; (Larsen & Nelson, 2017))**.** *For every integer $n > 1$ and every $\varepsilon \in (0, 1/2)$, there exists a set $V \subset \mathbb{R}^n$ of cardinality $N = \Theta(n^3)$ such that any matrix $A \in \mathbb{R}^{m\times n}$ satisfying*

$$\|Av\|_2^2 \in (1 \pm \varepsilon)\|v\|_2^2 \qquad \text{for every } v \in V \tag{13}$$

must have

$$m \;=\; \Omega\big(\min\{n,\, \varepsilon^{-2}\log n\}\big) \;=\; \Omega\big(\min\{n,\, \varepsilon^{-2}\log N\}\big). \tag{14}$$

Corollary C.2 (Inner-product form of the Larsen–Nelson bound)**.** *For every integer $n > 1$ and every $\varepsilon \in (0, 1/2)$, there exists a set $U \subset \mathbb{R}^n$ of cardinality $N = \Theta(n^3)$ such that any matrix $A \in \mathbb{R}^{m\times n}$ satisfying*

$$\big|\langle Au, Au'\rangle - \langle u, u'\rangle\big| \;\le\; \varepsilon \max\big(\|u\|_2^2, \|u'\|_2^2\big) \qquad \text{for every } u, u' \in U \tag{15}$$

must have $m = \Omega(\min\{n, \varepsilon^{-2}\log N\})$.

Proof. Take $U = V \cup \{0\}$ for V as in Lemma C.1, with ε replaced by $\varepsilon/3$ to absorb constants as in the proof of Lemma A.2. Applying condition (15) to the pair $(v, 0)$ for $v \in V$ gives $\big|\|Av\|_2^2 - \|v\|_2^2\big| \le \varepsilon\|v\|_2^2$, which is exactly hypothesis (13) of Lemma C.1. The conclusion follows immediately. □

C.2 Impossibility Below the Critical Rate for All Linear Maps

Theorem C.3 (Worst-case lower bound for all linear maps)**.** *For every $D > 1$ and every $\varepsilon \in (0, 1/2)$, there exists a configuration of $N = \Theta(D^3)$ vectors $g_1, \ldots, g_N \in \mathbb{R}^D$ such that any matrix $P \in \mathbb{R}^{D\times d}$, block-structured or otherwise, satisfying*

$$\big|\langle P^\top g_i, P^\top g_j\rangle - \langle g_i, g_j\rangle\big| \;\le\; \varepsilon \max\big(\|g_i\|_2^2, \|g_j\|_2^2\big) \qquad \text{for all } 1 \le i, j \le N \tag{16}$$

must satisfy

$$d \;=\; \Omega\big(\min\{D,\, \varepsilon^{-2}\log N\}\big). \tag{17}$$

In particular, once D grows large enough that $D/\log N$ exceeds ε^{-2}, any P satisfying (16) requires $d = \Omega(\varepsilon^{-2}\log N)$, matching the sufficient condition of Theorem A.3 up to constants and the dependence on L.

Proof. Apply Corollary C.2 with $n = D$, obtaining a set $U = \{u_1, \ldots, u_N\} \subset \mathbb{R}^D$, $N = \Theta(D^3)$, for which any inner-product-preserving linear map must have output dimension $\Omega(\min\{D, \varepsilon^{-2}\log N\})$. Taking $g_i := u_i$ and $A := P^\top$, condition (16) is exactly (15), so the dimension lower bound (17) follows immediately from Corollary C.2. □

Remark C.4. Theorem C.3 is an existential statement: it exhibits a specific configuration of $N = \Theta(D^3)$ vectors for which no linear map can beat the $\varepsilon^{-2}\log N$ rate, establishing that the upper bound of Theorem A.3 is tight as a statement about linear projections in general. The hard configuration is the one constructed by Larsen & Nelson (2017) and is not claimed to resemble the gradient vectors that arise from a randomly initialized network; whether typical NTK gradient vectors are hard or easy instances for dimension reduction is a separate question that neither this theorem nor Theorem B.3 addresses. Together, the two lower bounds say: the ε^{-2} rate is unavoidable for random Gaussian projections applied to any vectors (Theorem B.3), and unavoidable for all linear maps applied to some vectors (Theorem C.3); what remains open is whether typical NTK gradient vectors admit a smaller critical dimension under some non-Gaussian or structured projection.

Exp.	data	N	width m	ambient D	subspace grid d	prec.
1	MNIST 0/1	512	1024	804,865	$4-10^3$ (log)	fp32
2	synth. teacher	64	1024	—	$\{8, 32, 128, 512\}$	fp64
3	1-D regression	64	1024	—	$\{2, 8, 32, 128, 512\}$	fp64
4a	sphere	48	2048	—	24 (rank 24)	fp64
4b	sphere	60	1024	—	$\{8, \ldots, 120\}$, net d=30	fp64
5	synth. class.	512	1024	34,817	$2-10^3$ (log)	fp64

Table 1: Per-experiment configuration. All use $L = 2$, sigmoid activations, the NTK parameterisation, and five projection seeds for the swept quantities. Experiment 1 uses single precision for memory at $D \approx 8 \times 10^5$; all others use double precision.

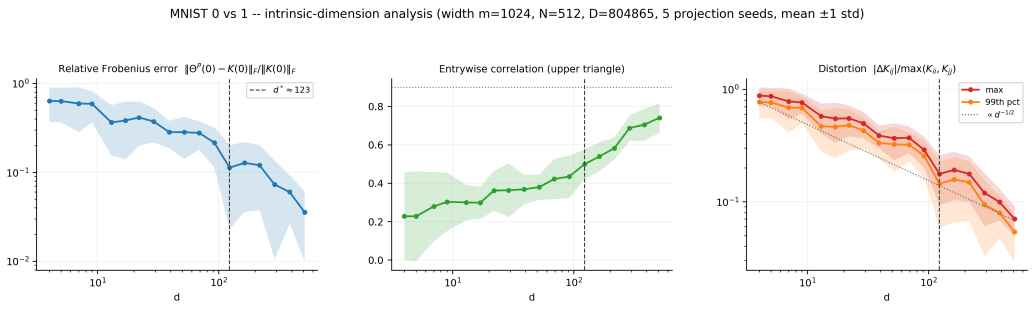


Figure 4: **Kernel approximation on MNIST** 0 **vs.** 1 ($m = 1024$, $N = 512$, $D = 804{,}865$; mean ± 1 std over five projection seeds). The relative Frobenius error decreases with d, entrywise correlation increases toward one, and the maximum and 99th-percentile distortions follow the predicted $d^{-1/2}$ scaling. The dashed line marks the kernel critical dimension $d^* \approx 123$.

D Additional Experimental Details and Results

This appendix records the full experimental configuration and the results deferred from Section 5: the kernel approximation measurements underlying the MNIST result, the kernel spectra underlying the critical dimension, an entrywise view of kernel constancy, the quantitative convergence of subspace training to the full-space solution, the frozen residual that appears when $d < N$, a controlled test of the alignment condition, and a direct comparison of the measured intrinsic dimension against the theoretical bounds.

D.1 Configuration

All runs use one-hidden-layer sigmoid networks in the NTK parameterisation, squared loss, and full-batch gradient descent, with the block-Gaussian subspace construction and step-size rule of Section 5. Curves that sweep d share a single initialisation θ_0 and use *nested* projections across the grid (common random numbers), which removes sampling jitter while leaving each individual d with exactly the i.i.d. projection law the theory assumes; no smoothing or curve-fitting is applied, and shaded bands are the seed-to-seed spread. Table 1 lists the per-experiment settings.

D.2 Additional results and remarks

Kernel approximation on MNIST. Figure 4 provides the kernel-approximation diagnostics underlying the MNIST result in Figure 1. For $m = 1024$, $N = 512$, and $D = 804{,}865$, the relative Frobenius error between the projected kernel $\Theta^P(0)$ and the true NTK $K(0)$ decreases with d, while the entrywise correlation increases toward one. The maximum and 99th-percentile entrywise distortions follow the predicted $d^{-1/2}$ scaling from Lemma A.2. Using a 15% distortion threshold gives the kernel critical dimension $d^* \approx 123$. These measurements quantify the kernel-approximation property used in the main-text generalisation experiment.

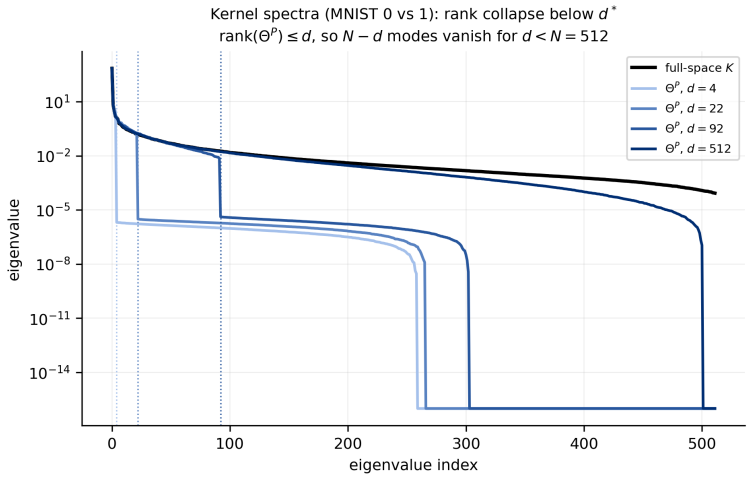


Figure 5: **Kernel spectra (MNIST** 0 **vs.** 1**).** The projected kernel Θ^P matches the top of the full-space spectrum and truncates the rest: $\mathrm{rank}(\Theta^P) \le d$, so $N - d$ modes vanish for $d < N = 512$. At $d = 10^3 > N$ the projection is full-rank and recovers K.

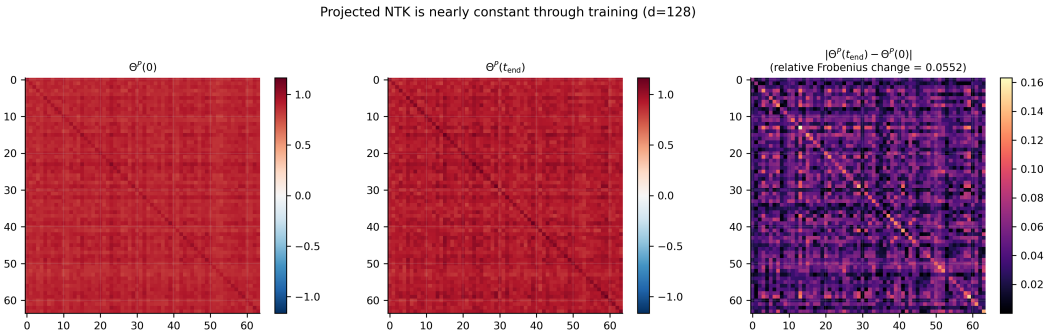


Figure 6: **Projected NTK is nearly constant through training** ($d = 128$). $\Theta^P(0)$ and $\Theta^P(t_{\mathrm{end}})$ are indistinguishable by eye; their difference has relative Frobenius norm 0.0552 and no structure.

Kernel spectra: rank collapse. Figure 5 shows why a critical dimension exists. Because $\Theta^P = \sum_\ell F_\ell F_\ell^\top$ with $F_\ell \in \mathbb{R}^{N \times d_\ell}$, its rank is at most d, so for $d < N = 512$ exactly $N - d$ eigenvalues are numerically zero. The projected spectrum tracks the top eigenvalues of the full-space kernel K and truncates the remainder at the (single-precision) floor; the cliff sits at the retained rank (dotted lines at $d = 4, 28, 142$), and once $d = 10^3 > N$ the projected kernel is full-rank and coincides with K across all 512 modes. The Frobenius critical dimension of Section 5.1 is precisely the point at which the retained top of the spectrum captures enough of K's energy for the relative error to fall below threshold.

Entrywise kernel constancy. Figure 6 complements the sup-norm and closed-form diagnostics of Figure 2 with an entrywise picture at $d = 128$. The kernel at initialisation and at the end of training are visually identical; over the entire run the relative Frobenius change is only 5.5%, and the difference map carries no low-rank or structured signal — the drift is a small, diffuse perturbation rather than a systematic deformation of the kernel.

Function evolution and closed-form NTK dynamics. The panels of Figure 7 provide complementary evidence for the kernel-regime interpretation. Although the projected kernel remains nearly constant, the network function changes by an $O(1)$ amount: $\|f_t(X) - f_0(X)\|_2/\sqrt{N}$ reaches approximately 1.6. Moreover, the measured training residual closely follows the closed-form kernel prediction $e^{-tK_{\mathrm{eff}}(0)/N}\vec{\Delta}_0$, with a deviation of at most approximately 2%. Thus the stability ob-

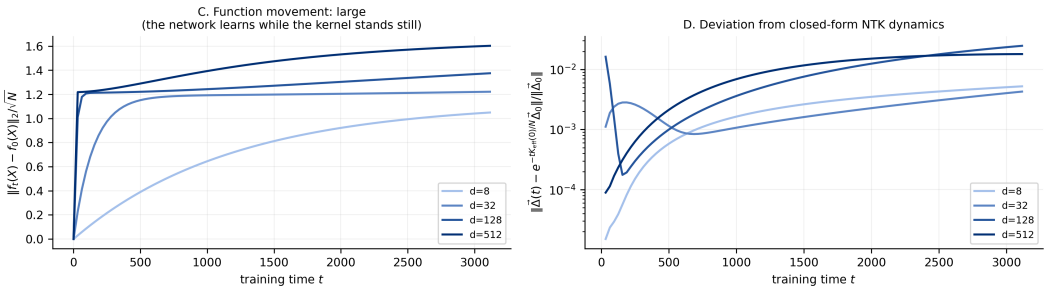


Figure 7: **Function evolution and closed-form NTK dynamics** (synthetic teacher, $m = 1024$, $N = 64$). *C:* the network function changes by an $O(1)$ amount during training despite the near-constant projected kernel. *D:* the measured residual closely follows the closed-form NTK dynamics, with deviation below approximately 2%.

served in the tangent kernel does not imply a static network function, and the observed dynamics are consistent with the predicted frozen-kernel evolution.

Convergence to the full-space solution. Figure 8 quantifies the convergence shown qualitatively in the main text. In the top panel the measured training residual $\|f_t(X) - y\|_2/\sqrt{N}$ (solid) follows the closed-form NTK prediction $e^{-tK_{\text{eff}}(0)/N}\vec{\Delta}_0$ (dashed) and settles at the analytic floor (dotted) set by the projection of the target onto $\mathcal{N}(K_{\text{eff}})$; the floor decreases with d (from ≈ 0.65 at $d = 2$ to ≈ 0.13 at $d = 512$). The bottom panel measures the function-space gap to full training directly: $\|f_d - f_{\text{full}}\|_2^2/n_{\text{grid}}$ decreases as a clean power law in d, from ≈ 0.5 at $d = 2$ to $\approx 1.3 \times 10^{-3}$ at $d = 512$, making precise the visual convergence of the main-text panel.

Frozen residual when $d < N$. When $d < N$ the effective kernel is rank-deficient, so any component of the residual lying in $\mathcal{N}(K_{\text{eff}})$ never decays: the error retains an irreducible floor (Section 4.3). Figure 9 isolates this with $N = 60$. The measured floor decreases with d and vanishes exactly at $d = N = 60$ (top right); it sits slightly below the isotropic prediction $\sqrt{(N-d)/N}$ because the target is not isotropic in the kernel eigenbasis. The spectrum shows exactly $N - d$ eigenvalues collapsing to zero (bottom left, cliffs at $d = 16, 45$; full rank at $d = 60$). Most importantly, a genuinely trained network at $d = 30$ (bottom right) descends toward but never crosses the analytic floor 0.581: the floor governs real gradient descent, not merely the closed-form idealisation.

Learnability and $\text{Col}(K_{\text{eff}})$. Figure 10 tests the alignment condition of Section 4.3 directly. With $N = 48$ and $d = 24$ (so $\operatorname{rank} K_{\text{eff}} = 24$), we synthesise targets that place a controlled fraction α of the initial residual's energy in $\mathcal{N}(K_{\text{eff}})$. When the residual is orthogonal to the null space ($\alpha = 0$) the loss decays to $\approx 5 \times 10^{-9}$ and the error vanishes; when it lies entirely in the null space ($\alpha = 1$) there is no progress at all; and at $\alpha = 0.5$ the loss plateaus exactly at the predicted height $\|\vec{\Delta}_{\text{null}}\|^2/2N$ (left panel). The middle panel resolves this into components — the range part decays while the null part is frozen — and the right panel shows the spectrum with exactly $N - d = 24$ zero modes. This is the sharp statement of Section 4.3: subspace training reduces error to zero if and only if the target is orthogonal to the frozen null space of K_{eff}.

The intrinsic dimension and its bounds. Finally, Figure 11 compares the measured intrinsic dimension against the bounds our theory proves. On a synthetic classification dataset ($N = 512$, ambient $D = 34{,}817$) we define the intrinsic dimension at accuracy ε as

$$d^*(\varepsilon) \;=\; \min\left\{\, d : \max_{ij}|\Theta^P_{ij}(0) - K_{ij}(0)| \big/ \max_i \|g_i\|^2 \le \varepsilon \,\right\},$$

which is exactly the accuracy controlled by Theorem A.3, and overlay the sufficient (upper) bound $C_{JL}\, L^3 \varepsilon^{-2} \log(NL/\delta)$ of Theorem A.14 and the necessary (lower) bound $c_{LB} \min\{D, \varepsilon^{-2} \log N\}$ of Theorem 4.3 (via Corollary C.2). The measured $d^*(\varepsilon)$ has log-log slope -2.13, matching the theoretical ε^{-2} rate, and lies inside the band; the absolute constants are fitted to $O(1)$ values while the band's multiplicative width $L^3 \log(NL/\delta)/\log N$ is the one the theory dictates (left panel). The middle panel confirms the underlying distortion $\varepsilon(d) \propto d^{-1/2}$ from which d^* is read off.

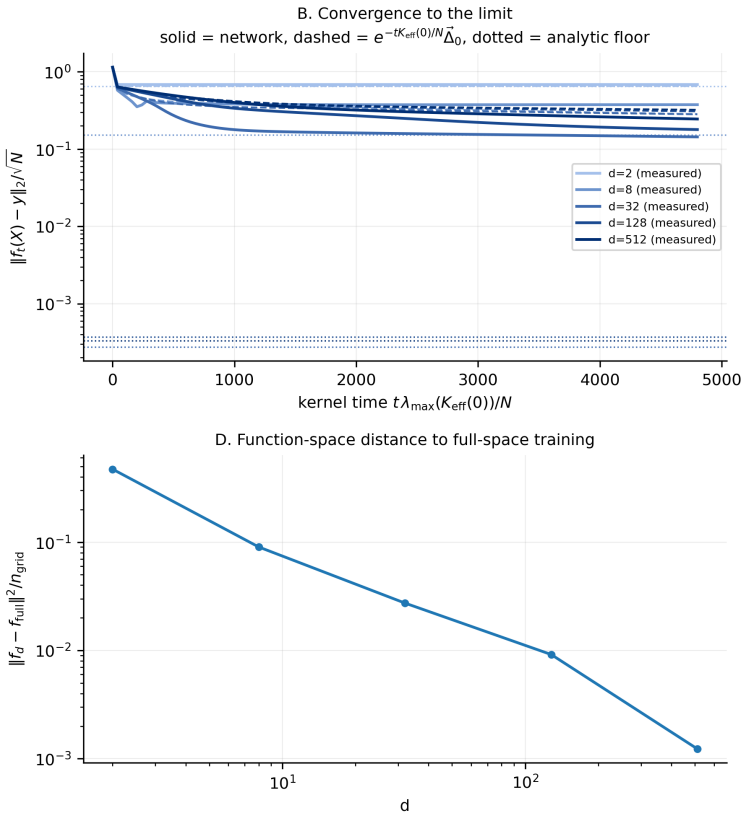


Figure 8: **Convergence of subspace training** (1-D regression, $m = 1024$, $N = 64$). *Top:* the residual follows the closed-form NTK dynamics and settles at the analytic floor set by the target's overlap with $\mathcal{N}(K_{\text{eff}})$; larger d gives a lower floor. *Bottom:* the function-space distance to full training decays polynomially in d.

The right panel adds train and test error after subspace training. Unlike MNIST 0 vs. 1 — where the task is essentially linearly separable and is solved far below d^* (Figure 1) — this synthetic target is genuinely hard, so the error decreases only gradually and the task-critical dimension is comparable to the kernel critical dimension $d^*(\varepsilon_0) = 101$ at the operating accuracy $\varepsilon_0 = 0.3$. The contrast underlines a distinction worth stating plainly: $d^*(\varepsilon)$ is a task-independent property of the kernel geometry, whereas the dimension at which a *particular* label sequence is learned depends on how that target aligns with the surviving eigenspace of K_{eff}.

Remarks on the measurements. Three points support the reading of the figures. *(i) Two kernels.* Our approximation results bound the unweighted Θ^P, for which $\mathbb{E}[\Theta^P] = K$, whereas the training dynamics are driven by the reweighted $K_{\text{eff}} = \sum_\ell \eta_\ell \Theta^{P,(\ell)}$; the two coincide in full space ($\eta_\ell = 1$). Because K_{eff} scales with d, we normalise every step size by $\lambda_{\max}(K_{\text{eff}}(0))$, so that the d-sweeps isolate geometry rather than optimisation speed. *(ii) Precision.* Experiment 1 runs in single precision to fit the $D \approx 8 \times 10^5$ kernel in memory; this places the truncated tail of the projected spectrum at the single-precision floor and makes the subspace solve unreliable for $d \gtrsim N$ (the error spike in Figure 1), which is why we read that figure only in the informative range $d < N$. All other experiments use double precision and resolve the exact-zero eigenvalues cleanly (e.g. the sharp rank cliffs in Figures 9 and 10). *(iii) Smoothness.* The d-sweeps are smooth because a single θ_0 and nested projections are shared across the grid (common random numbers); this is a variance-reduction device, not smoothing — each point retains the exact i.i.d. projection law the theory requires, and the reported bands are genuine seed spread.

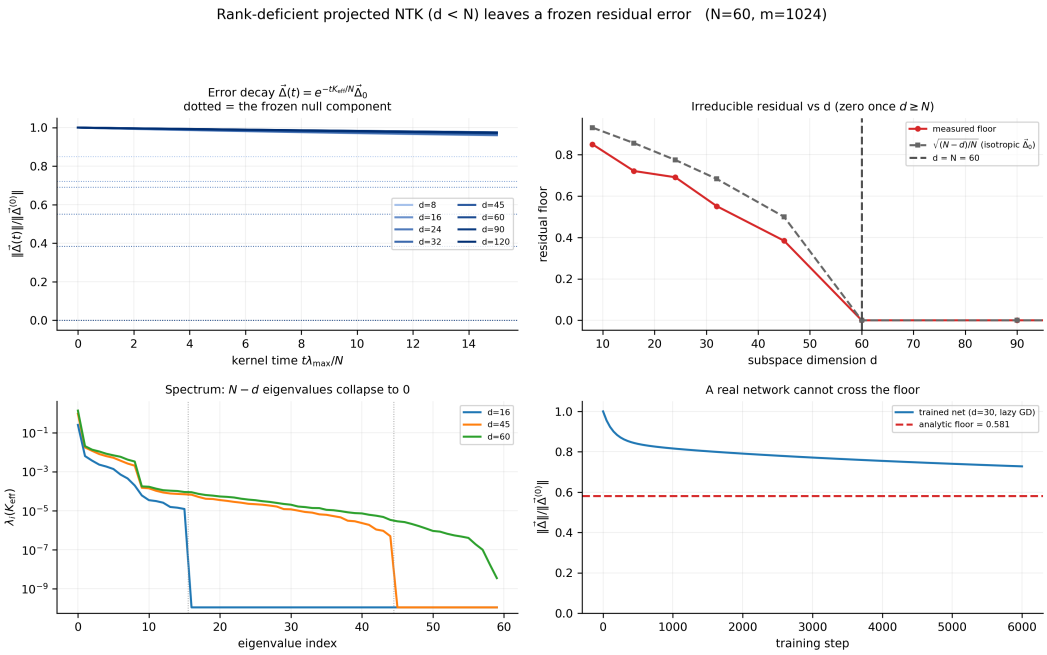


Figure 9: **Rank-deficient projected NTK ($d < N$) leaves a frozen residual** ($N = 60$, $m = 1024$). *Top left:* closed-form decay with the frozen null component (dotted). *Top right:* the residual floor vanishes at $d = N$ and lies just below the isotropic $\sqrt{(N-d)/N}$. *Bottom left:* $N - d$ eigenvalues collapse to zero. *Bottom right:* a trained network cannot cross the analytic floor.

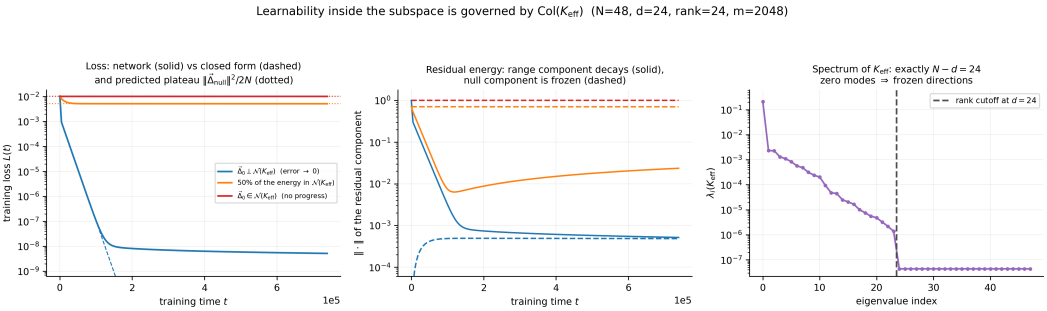


Figure 10: **Learnability inside the subspace is governed by** $\mathrm{Col}(K_{\text{eff}})$ ($N = 48$, $d = 24$, rank 24). *Left:* loss for a residual orthogonal to (α=0), half in (α=0.5), and entirely in (α=1) the null space; network (solid) matches the closed form (dashed) and the predicted plateau (dotted). *Middle:* the range component decays, the null component is frozen. *Right:* K_{eff} has exactly $N - d$ zero modes.

The intrinsic dimension and network depth The sufficient bound of Theorem A.14 scales as L^3. This dependence on depth enters entirely through the block-structured projection: the Johnson–Lindenstrauss budget is split across the L layers, so preserving each of them to accuracy ε/L costs $(L/\varepsilon)^2$ dimensions per layer and L^3/ε^2 in total (Appendix A.1.3). It is natural to ask whether this is a genuine property of the intrinsic dimension or an artefact of the bounding technique. We probe it empirically.

Setup. We fix a single synthetic classification task and a total parameter budget D, and build fully-connected networks of several depths L that all have (approximately) the same D. Because D is held fixed, deeper networks are proportionally narrower; we take D large enough ($D \approx 5 \times 10^5$) that even the deepest network stays wide (hidden width in the several hundreds), so every network is in the kernel regime. All networks are trained inside the same block-structured random subspace used

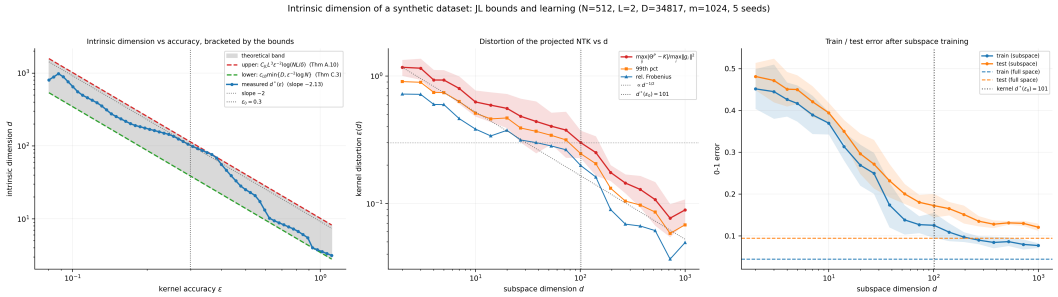


Figure 11: **Intrinsic dimension of a synthetic dataset vs. the bounds** ($N = 512$, $L = 2$, $D = 34{,}817$, $m = 1024$). *Left:* $d^*(\varepsilon)$ (slope -2.13) sits between the sufficient bound (Theorem A.14) and the necessary bound (Theorem C.3), both $\Theta(\varepsilon^{-2})$. *Middle:* the distortion $\varepsilon(d) \propto d^{-1/2}$ underlying d^*. *Right:* train/test error after subspace training; here the task dimension is comparable to the kernel $d^*(\varepsilon_0) = 101$.

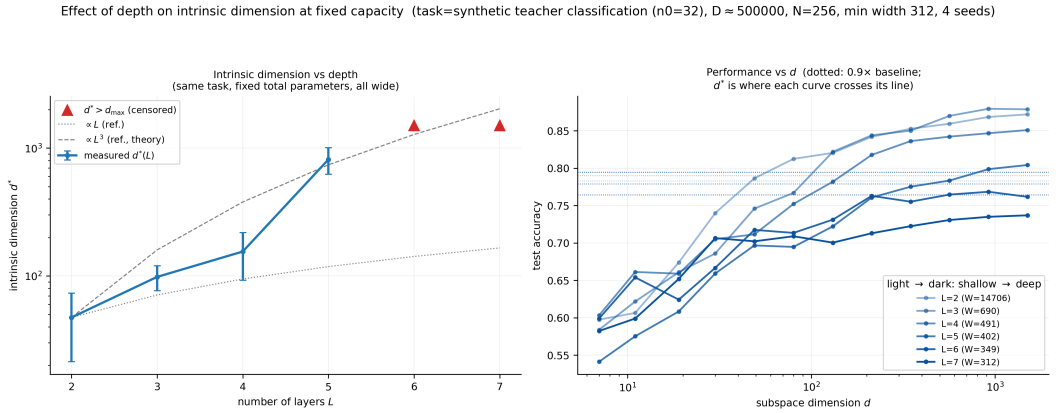


Figure 12: **Intrinsic dimension vs. depth at fixed capacity.** Networks of increasing depth L share the same parameter count $D \approx 5 \times 10^5$ (so deeper $\Rightarrow$ narrower; widths shown in the legend) and the same task, and are trained in the block-structured random subspace. *Left:* the intrinsic dimension $d^*(L)$ grows steeply, tracking the L^3 reference (dashed) rather than the L reference (dotted); red triangles mark depths whose d^* exceeds the largest tested subspace ($d > d_{\max} = 1500$, i.e. lower bounds). *Right:* the performance-vs-d curves from which d^* is read (dotted lines: the per-depth targets); the two deepest networks never reach their target within the tested range, hence the lower bounds.

throughout the paper, with the layer-wise learning rate $\eta_\ell = d_\ell / N_\ell$. We use the $\tanh$ nonlinearity here — smooth with a bounded second derivative, hence covered by the assumptions of Appendix A — because very deep saturating (sigmoid) networks fail to train from this initialisation. Following Li et al. (2018), the intrinsic dimension $d^*(L)$ is the smallest subspace dimension at which subspace training recovers a target fraction of the full-space performance; operationally, we shrink d until performance collapses. When even the largest subspace we test ($d \le d_{\max}$) fails to reach the target, we report d^* as a lower bound $d^* > d_{\max}$.

The intrinsic dimension grows steeply with depth. Figure 12 shows that $d^*(L)$ increases sharply with the number of layers. Under a stringent recovery criterion (matching close to the full-space accuracy) the measured intrinsic dimension tracks the L^3 reference over the range we can resolve: for $L = 2, 3, 4, 5$ we find $d^* \approx 47,\ 98,\ 155,\ 815$, and the two deepest networks ($L = 6, 7$) require more dimensions than the largest subspace tested ($d^* > 1500$, plotted as lower bounds). Over the resolved depths a log-log fit gives an exponent of roughly 2–3; the estimate is sensitive to the deepest, noisiest points, but the qualitative message is unambiguous — the dependence on depth is strong and comparable to L^3, not flat.

Caveats. Two caveats accompany this reading. First, holding D fixed couples depth with width (deeper $\Rightarrow$ narrower), so $d^*(L)$ measures the effect of added depth *at fixed capacity* rather than of depth in isolation; the width of every network is reported alongside the curve. Second, L^3 is a *sufficient* bound (Theorem A.14): that the measured exponent approaches 3 under a stringent criterion indicates the bound is close to tight in that regime, while the much milder growth under loose criteria shows that the intrinsic dimension of a typical, moderately-solved task is far smaller than the worst case.