Convergence of Near-Linear Width ReLU Networks with Unbalanced Initialization

Zhenghao Xu
Georgia Institute of Technology
zhenghaoxu@gatech.edu

Rachel Ward
University of Texas at Austin
rward@math.utexas.edu

Molei Tao
Georgia Institute of Technology
mtao@gatech.edu

Tuo Zhao
Georgia Institute of Technology
tourzhao@gatech.edu

Abstract

Linear-width convergence for two-layer ReLU networks is known in random-feature or conjugate kernel (CK) regimes whose analyses exploit nearly fixed hidden features. However, the corresponding neural tangent kernel (NTK) regime, where hidden-layer training is active, still suffers from substantially larger width requirements. The obstacle is the ReLU-induced kernel shift: activation-pattern changes can induce empirical NTK drift that breaks the contraction argument. We show that an output-dominant unbalanced Gaussian initialization creates an NTK-dominated training regime in which this kernel shift remains controllable. This single mechanism yields near-linear width, Nesterov acceleration, and low-rank adaptivity for two-layer ReLU networks with vector-valued outputs and a shared first layer. We prove that gradient descent (GD) achieves linear convergence for networks with $\widetilde{\Omega}(Nn/\lambda)$ neurons, where N is the sample size, n is the output dimension, and λ is the standard smallest eigenvalue of the limiting NTK. Within the same framework, Nesterov's accelerated gradient (NAG) attains a provable speedup without sacrificing near-linear width, improving the iteration complexity from $O(n\kappa \log \frac{1}{\epsilon})$ to $O(\sqrt{n\kappa} \log \frac{1}{\epsilon})$, where κ is the limiting NTK condition number. Finally, our analysis establishes low-rank adaptivity: by introducing a sketching step at initialization and a subspace analysis, the width requirement reduces to $\widetilde{\Omega}(Nr\kappa^2(\mathbf{Y})/\lambda)$ for responses $\mathbf{Y}$ of rank $r \ll n$, matching the ambient-dimensional result in the leading polynomial dependence, up to polylogarithmic factors, with n replaced by r when $\kappa(\mathbf{Y}) = O(1)$. In contrast to CK/random-feature analyses that exploit nearly fixed hidden features, our proof controls hidden-layer NTK contributions, ReLU activation-pattern changes, and kernel shift; this control enables both acceleration and low-rank adaptivity in the NTK-dominated regime.

1 Introduction

The study of neural network training dynamics has seen substantial progress in recent years, particularly in the overparameterized regime where networks can perfectly fit the training data [Du et al., 2018, Allen-Zhu et al., 2019]. A central question is how wide a network must be for first-order methods to provably converge to a global minimum. It is widely conjectured that a width linear in the sample size N should suffice, and this has been confirmed for several models with smooth activations [Liu et al., 2022a, Bombari et al., 2022]. For the widely used non-smooth ReLU activation, however, the theory still has a substantial gap. For two-layer ReLU networks, classical

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

neural tangent kernel (NTK, Jacot et al. 2018) analyses require widths that scale polynomially with N and with data-dependent quantities such as the smallest eigenvalue λ of the limiting NTK. In this line, the best-known width requirement has improved from $\Omega(N^6\lambda^{-4})$ [Du et al., 2018] to $\widetilde{\Omega}(N^2\lambda^{-2})$ [Munteanu et al., 2022], but it remains far from the conjectured linear dependence on N. A complementary result of Nguyen [2021] obtains linear width under a different unbalanced initialization in the random-feature regime, where the dynamics are driven by a conjugate kernel (CK) with nearly fixed hidden features.

This paper asks whether near-linear width can be achieved while staying in an NTK-dominated ReLU regime, rather than moving to a random-feature/CK-driven regime with provably suppressed hidden-layer movement. By NTK-dominated, we mean that the linearized residual dynamics are led by a controllable hidden-layer NTK contribution, so ReLU sign changes and the resulting kernel shift must be controlled along training. Writing the empirical total kernel at step t as $\mathbf{H}_t = \mathbf{K}_t + \mathbf{G}_t$, with $\mathbf{K}_t$ the hidden-layer NTK component and $\mathbf{G}_t$ the output-layer CK component, our output-dominant scaling makes $\mathbf{K}_t$ the leading component while keeping its shift controlled. Accordingly, the theorem statements are phrased in terms of the spectrum of $\mathbf{K}_0$. While $\mathbf{G}_t$ still enters the residual dynamics, it is controlled as a lower-order CK perturbation in this regime.

We prove that this is possible for two-layer ReLU networks trained with gradient descent (GD) and Nesterov's accelerated gradient (NAG). The vector-valued shared-first-layer setting makes the empirical hidden-layer NTK component an $Nn \times Nn$ Kronecker-structured matrix, which creates additional concentration and kernel-shift challenges. Our key technique is an output-dominant unbalanced Gaussian initialization: the scaling makes the hidden-layer NTK component dominate while still permitting controlled hidden-layer movement, and the Gaussian structure supplies the concentration and anti-concentration estimates needed to bound the evolving ReLU kernel. This mechanism yields three consequences.

• **Near-linear width in the NTK regime (Theorem 3.2):** GD achieves global linear convergence for vector-valued two-layer ReLU networks with width $\widetilde{\Omega}(Nn\lambda^{-1})$. This improves the quadratic dependence in classical scalar-output NTK analyses while handling the coupling introduced by a shared first layer in the vector-output setting.

• **Provable acceleration at near-linear width (Theorem 3.3):** In the same NTK-dominated vector-output setting, NAG achieves acceleration without increasing the width threshold, improving the rate from $O(n\kappa \log \frac{1}{\epsilon})$ to $O(\sqrt{n\kappa} \log \frac{1}{\epsilon})$.

• **Low-rank adaptivity (Theorem 4.1):** For rank-r responses with $r \ll n$, a sketching step embeds the response subspace into the output layer but makes the empirical kernel rank-deficient. We prove contraction of the reference dynamics on the subspace, show global convergence, and reduce the width to $\widetilde{\Omega}(Nr\kappa^2(\mathbf{Y})\lambda^{-1})$, matching the full-rank leading polynomial with n replaced by r when $\kappa(\mathbf{Y}) = O(1)$.

Collectively, these results, summarized in Table 1, show how near-linear width, acceleration, and low-rank adaptivity can coexist in an NTK-dominated ReLU regime. Full proofs and numerical experiments are provided in the appendix.

Relation to random-feature analyses. Table 1 separates the NTK and RF regimes by the object being controlled. RF/CK analyses study CK-driven dynamics with suppressed hidden-layer movement, whereas our NTK-dominated analysis controls an evolving hidden-layer NTK component. The two regimes are complementary: our results require arguments for hidden-layer NTK shift and ReLU sign changes, not a claim that acceleration or low-rank structure is impossible in fixed-feature models.

1.1 Related Work

Overparameterization and training regimes. In the infinite-width limit, training suitably scaled neural networks by gradient descent reduces to kernel regression with a fixed NTK, explaining linearized ("lazy") dynamics [Jacot et al., 2018, Lee et al., 2019, Chizat et al., 2019]. Different scalings lead to random-feature/CK, NTK-dominated, or richer feature-learning regimes [Neal, 2012, Chizat et al., 2019]. Smooth activations admit linear widths [Bombari et al., 2022, Liu et al., 2022a]; for ReLU, classical NTK analyses progressed from $\Omega(N^6)$ [Du et al., 2018] to $\widetilde{\Omega}(N^4)$ [Song and Yang, 2019, Oymak and Soltanolkotabi, 2020] and then $\widetilde{\Omega}(N^2)$ [Munteanu et al., 2022]. Nguyen [2021]

obtain linear width in the random-feature regime; our scaling instead targets an NTK-dominated regime with an active hidden-layer contribution.

Width and NTK eigenvalues. Our near-linear width bounds scale with the smallest limiting NTK eigenvalue λ, as is standard in NTK analyses. Recent work lower-bounds and concentrates NTK eigenvalues under geometric conditions [Karhadkar et al., 2024, Yang et al., 2024]; our vector-valued setting additionally requires controlling *kernel shift* during training.

Vector-valued outputs with a shared first layer. Unlike training n independent scalar heads, vector outputs with a shared first layer induce a *matrix-valued* kernel with Kronecker structure, a standard object in vector-valued RKHS and multi-output Gaussian processes [Micchelli and Pontil, 2005, Alvarez et al., 2012]. This coupling is the source of the matrix concentration and kernel-shift estimates used in our proofs.

Acceleration and low-rank adaptivity. Momentum methods accelerate NTK-style ReLU training [Wang et al., 2021, Liu et al., 2022b, Liao and Kyrillidis, 2024], but existing proofs use much larger widths and scalar outputs. We obtain NAG's $\sqrt{\kappa}$ speedup at near-linear width in the vector/shared-feature setting. Our low-rank sketching result connects to randomized sketching [Halko et al., 2011, Pilanci and Wainwright, 2016] and tailored-initialization analyses in matrix factorization and linear networks [Ward and Kolda, 2023, Xu et al., 2024].

Table 1: Results for two-layer ReLU networks. N is the sample size, λ is the smallest limiting NTK eigenvalue, κ is the limiting NTK condition number, κ' is the conjugate-kernel condition number, and $\kappa_r = \kappa(\mathbf{Y})$ with response rank r. "RF" denotes the random-feature regime where the dynamics are driven by CK.

Output	Algorithm	Initialization	Width	Rate	Regime
Scalar ($n = 1$)	GD [Du et al., 2018]	Random	$\Omega(N^6\lambda^{-4})$	$O(N^2\lambda^{-2}\log\frac{1}{\epsilon})$	NTK
	GD [Song and Yang, 2019]	Random	$\widetilde{\Omega}(N^4\lambda^{-4})$	$O(N^2\lambda^{-2}\log\frac{1}{\epsilon})$	NTK
	HB [Wang et al., 2021]	Random	$\widetilde{\Omega}(N^4\kappa^2\lambda^{-4})$	$O(\sqrt{\kappa}\log\frac{1}{\epsilon})$	NTK
	NAG [Liu et al., 2022b]	Random	$\widetilde{\Omega}(N^4\kappa^2\lambda^{-2})$	$O(\sqrt{\kappa}\log\frac{1}{\epsilon})$	NTK
	GD [Munteanu et al., 2022]	Coupled	$\widetilde{\Omega}(N^2\lambda^{-2})$	$O(N^2\lambda^{-2}\log\frac{1}{\epsilon})$	NTK
	GD (Ours, Theorem 3.2)	Unbalanced	$\widetilde{\Omega}(N\lambda^{-1})$	$O(\kappa\log\frac{1}{\epsilon})$	NTK
	NAG (Ours, Theorem 3.3)	Unbalanced	$\widetilde{\Omega}(N\lambda^{-1})$	$O(\sqrt{\kappa}\log\frac{1}{\epsilon})$	NTK
n-dim. Vector	GD [Nguyen, 2021]	Unbalanced	$\Omega(N)$	$O(\kappa'\log\frac{1}{\epsilon})$	RF
	GD (Ours, Theorem 3.2)	Unbalanced	$\widetilde{\Omega}(Nn\lambda^{-1})$	$O(n\kappa\log\frac{1}{\epsilon})$	NTK
	NAG (Ours, Theorem 3.3)	Unbalanced	$\widetilde{\Omega}(Nn\lambda^{-1})$	$O(\sqrt{n\kappa}\log\frac{1}{\epsilon})$	NTK
	NAG (Ours, Theorem 4.1)	Unbalanced	$\widetilde{\Omega}(Nr\kappa_r^2\lambda^{-1})$	$O(\sqrt{r\kappa}\kappa_r\log\frac{1}{\epsilon})$	NTK

2 Preliminaries

Throughout this paper, $\|\cdot\|$ denotes the Euclidean norm of vectors or spectral norm of matrices, $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of matrices. For any matrix, $[\cdot]_{i,:}$ and $[\cdot]_{:,i}$ denote the i-th row and column, respectively, $\sigma_i(\cdot)$ denotes the i-th largest singular value. For a square matrix, $\lambda_i(\cdot)$ denotes its i-th largest eigenvalue. For a matrix $\mathbf{X}$, we use $\mathrm{col}(\mathbf{X})$, $\mathrm{row}(\mathbf{X})$ and $\ker(\mathbf{X})$ to denote its column space, row space and null space respectively, and denote its condition number as $\kappa(\mathbf{X})$. For a rank-r rectangular matrix, $\kappa(\mathbf{X})$ means $\sigma_1(\mathbf{X})/\sigma_r(\mathbf{X})$; for a positive semidefinite matrix with rank r, it means $\lambda_1(\mathbf{X})/\lambda_r(\mathbf{X})$ on its positive eigenspace. We use $\odot$ and $\otimes$ to denote the Hadamard and Kronecker product, $\oplus$ to denote the direct sum of vector spaces, $\mathrm{vec}(\cdot)$ to denote the column-first vectorization of a matrix, and $\mathrm{diag}(\cdot)$ to denote the diagonal matrix from a vector. For scalars a, b, we occasionally use $a \vee b$ and $a \wedge b$ to denote their maximum and minimum. We use $\lesssim$ and $\gtrsim$ to omit absolute constants. The notation $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ suppresses polylogarithmic factors in the displayed problem parameters, including $N, n, d, m, r, \kappa, \kappa(\mathbf{Y})$ and $1/\delta$; polynomial dependence on these quantities is displayed explicitly.

We consider training vector-valued two-layer ReLU networks for regression under the square loss:

$$\min_{\mathbf{V},\mathbf{W}} \mathcal{L}(\mathbf{V}, \mathbf{W}) = \frac{1}{2}\left\|\mathbf{Y} - \mathbf{V}\sigma(\mathbf{W}^\top\mathbf{X})\right\|_{\mathrm{F}}^2. \tag{1}$$

Here, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) \in \mathbb{R}^{d\times N}$ denotes the input data, where d is the input dimension and N is the sample size. The response matrix is $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N) \in \mathbb{R}^{n\times N}$, with each response $\mathbf{y}_j$ an n-dimensional vector. The activation $\sigma(\cdot) = \max(\cdot, 0)$ is applied element-wise, and $\mathbf{V} \in \mathbb{R}^{n\times 2m}$ and $\mathbf{W} \in \mathbb{R}^{d\times 2m}$ are the output and hidden layers of a network with $2m$ neurons. We assume the data is normalized as in Du et al. [2018, 2019], Munteanu et al. [2022].

Assumption 2.1. For any $j \in [N]$, $\|\mathbf{x}_j\| = 1$, $\|\mathbf{y}_j\| \leq 1$.

Remark 2.2 (Boundedness of the data). The normalization in Assumption 2.1 is used only to keep constants explicit. If the covariates and responses are bounded after rescaling, the same proof applies with the corresponding constants. For sub-Gaussian data, one may condition on the standard high-probability event that all samples have bounded norm, rescale on that event, and absorb the resulting factors into the logarithmic and constant terms.

2.1 Gradient Descent and Nesterov's Acceleration

Let $\mathbf{R}(\mathbf{V}, \mathbf{W}) = \mathbf{V}\sigma(\mathbf{W}^\top\mathbf{X}) - \mathbf{Y}$ denote the residual. The (sub)gradients of the loss function in (1) with respect to $\mathbf{V}$ and $\mathbf{W}$ are:

$$\nabla_V\mathcal{L}(\mathbf{V}, \mathbf{W}) = \mathbf{R}(\mathbf{V}, \mathbf{W})\sigma(\mathbf{X}^\top\mathbf{W}),\ \nabla_W\mathcal{L}(\mathbf{V}, \mathbf{W}) = \mathbf{X}((\mathbf{R}(\mathbf{V}, \mathbf{W})^\top\mathbf{V}) \odot \mathbb{1}[\mathbf{X}^\top\mathbf{W} \geq 0]),$$

where $\mathbb{1}[\cdot]$ denotes entry-wise indicator. At the t-th iteration, GD performs the following update:

$$\mathbf{V}_{t+1} = \mathbf{V}_t - \eta\nabla_V\mathcal{L}(\mathbf{V}_t, \mathbf{W}_t),\ \mathbf{W}_{t+1} = \mathbf{W}_t - \eta\nabla_W\mathcal{L}(\mathbf{V}_t, \mathbf{W}_t), \tag{2}$$

where η is the step size. For NAG, we take the form initially designed for smooth strongly convex loss [Nesterov, 2013] with a fixed momentum parameter β for all iterations:

$$\begin{aligned}\mathbf{V}_{t+1} &= (1+\beta)(\mathbf{V}_t - \eta\nabla_V\mathcal{L}(\mathbf{V}_t, \mathbf{W}_t)) - \beta(\mathbf{V}_{t-1} - \eta\nabla_V\mathcal{L}(\mathbf{V}_{t-1}, \mathbf{W}_{t-1})),\\ \mathbf{W}_{t+1} &= (1+\beta)(\mathbf{W}_t - \eta\nabla_W\mathcal{L}(\mathbf{V}_t, \mathbf{W}_t)) - \beta(\mathbf{W}_{t-1} - \eta\nabla_W\mathcal{L}(\mathbf{V}_{t-1}, \mathbf{W}_{t-1})),\end{aligned} \tag{3}$$

where we set $\mathbf{V}_{-1} = \mathbf{V}_0$ and $\mathbf{W}_{-1} = \mathbf{W}_0$. When $\beta = 0$, NAG (3) reduces to GD (2). Therefore, we can study the convergence rate of GD and NAG in a unified form.

2.2 Limiting and Empirical Kernels

Our analysis relies on the neural tangent kernel (NTK, Jacot et al. 2018). We make the following standard assumption on the limiting NTK $\overline{\mathbf{K}} \in \mathbb{R}^{N\times N}$ of two-layer ReLU networks, defined as $\overline{\mathbf{K}}_{j,k} = \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\mathbf{x}_j^\top\mathbf{x}_k \cdot \mathbb{1}[\mathbf{w}^\top\mathbf{x}_j \geq 0, \mathbf{w}^\top\mathbf{x}_k \geq 0]\right]$.

Assumption 2.3. Assume $\overline{\mathbf{K}}$ has smallest eigenvalue $\lambda := \lambda_N(\overline{\mathbf{K}}) > 0$, and thus its condition number $\kappa := \kappa(\overline{\mathbf{K}}) < \infty$.

Assumption 2.3 holds when no two data points are parallel [Du et al., 2018]. The value of λ depends on data collinearity (see Remark 3.6).

Since we consider training both layers, the limiting conjugate kernel (CK, Neal 2012) $\overline{\mathbf{G}} := \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\mathbf{I}_d)}[\sigma(\mathbf{X}^\top\mathbf{w})\sigma(\mathbf{w}^\top\mathbf{X})]$ is also involved. According to the original definition in Jacot et al. [2018], the full NTK for training both layers is $\overline{\mathbf{K}} + \overline{\mathbf{G}}$. We refer to $\overline{\mathbf{K}}$ as the hidden-layer NTK component, and often simply as the NTK, to align with prior works that only consider optimizing the hidden layer [Du et al., 2018, Munteanu et al., 2022].

In our finite-width setting, the empirical hidden-layer NTK component and CK matrices at iteration t are defined as $\mathbf{K}_t := \sum_{i=1}^{2m}\mathbf{B}_{i,t}^\top \otimes \mathbf{A}_{i,t}$ and $\mathbf{G}_t := (\sigma(\mathbf{X}^\top\mathbf{W}_t)\sigma(\mathbf{W}_t^\top\mathbf{X})) \otimes \mathbf{I}_n$ respectively, where $\mathbf{A}_{i,t} = [\mathbf{V}_t]_{:,i}[\mathbf{V}_t]_{:,i}^\top$ and $\mathbf{B}_{i,t} = \mathrm{diag}(\mathbb{1}[\mathbf{W}_t^\top\mathbf{X} \geq 0]_{i,:})\mathbf{X}^\top\mathbf{X}\mathrm{diag}(\mathbb{1}[\mathbf{W}_t^\top\mathbf{X} \geq 0]_{i,:})$. We write $\mathbf{H}_t = \mathbf{K}_t + \mathbf{G}_t$ for the empirical total linearized kernel when both layers are trained.

It is useful to separate two ingredients that are sometimes both called "NTK concentration." The first is concentration of the empirical kernel at initialization around its limiting counterpart, a standard ingredient in NTK analyses [Jacot et al., 2018, Du et al., 2018, Song and Yang, 2019]. The second is stability of the empirical kernel along the training trajectory. For ReLU networks this second step is more delicate because activation patterns can change discontinuously, and controlling this kernel shift in the shared-feature vector-output setting is one of the main technical tasks of our analysis.

3 Results for General Vector Responses

We first state the full-rank vector-output guarantees under the coupled Gaussian initialization. This is the baseline setting before the low-rank sketching modification in Section 4. The initialization is

$$\mathbf{V}_0 = (c_1\mathbf{\Phi}, -c_1\mathbf{\Phi}), \quad \mathbf{W}_0 = (c_2\mathbf{\Psi}, c_2\mathbf{\Psi}), \tag{4}$$

where $\mathbf{\Phi} \in \mathbb{R}^{n\times m}$ and $\mathbf{\Psi} \in \mathbb{R}^{d\times m}$ are independent standard Gaussian random matrices, and $c_1, c_2 > 0$ are scaling factors. The two halves of neurons are coupled so that the initial network output is zero. Thus the scaling factors can be chosen to shape the training kernel without changing the initial loss. The following definition records the structural principle that the concrete Gaussian initialization is designed to realize.

Definition 3.1 (Output-dominant unbalanced initialization)**.** An initialization is output-dominant and unbalanced if it satisfies the following three structural properties: (i) the initial network output is zero, so the initial loss is independent of the scaling choice; (ii) the output-layer scale is sufficiently larger than the hidden-layer scale, so the hidden-layer NTK component dominates the CK component instead of placing training in the random-feature regime; and (iii) the overall kernel scale is sufficiently large, so that weight movement and the resulting ReLU-induced kernel shift remain controlled during training.

The coupled Gaussian construction in (4) is our concrete implementation of Definition 3.1 once the scaling conditions below hold. The ratio c_1/c_2 controls the CK-to-NTK bias, while the product c_1c_2 controls the overall kernel scale and stability region. This objective differs from He or Xavier initialization, whose primary role is balanced variance propagation across layers rather than selecting which training kernel dominates the dynamics.

We now instantiate this principle with concrete sufficient scaling conditions. For GD, we require

$$\frac{c_1^3}{c_2} \geq \widetilde{\Omega}\left(\frac{\sqrt{Nd}(\sqrt{N} \vee \kappa^4(\mathbf{K}_0))}{\sqrt{n}\lambda\kappa^3(\mathbf{K}_0)}\right), \ \frac{c_1}{c_2} \geq \widetilde{\Omega}\left(\frac{\sqrt{d}}{\sqrt{n}\kappa(\mathbf{K}_0)} \vee \frac{\sqrt{Nd}}{\sqrt{\lambda}}\right), c_1c_2 \geq \widetilde{\Omega}\left(\frac{\sqrt{Nn}\kappa(\mathbf{K}_0)}{\lambda}\right). \tag{5}$$

Here $\kappa = \kappa(\overline{\mathbf{K}})$ is the limiting-kernel condition number, whereas $\kappa(\mathbf{K}_0)$ is the realized initial-kernel condition number. Positive c_2 leaves Gaussian activation signs unchanged, and c_1^2 multiplies all eigenvalues of $\mathbf{K}_0$ by the same factor. Since changing the scaling factors does not affect the condition number of $\mathbf{K}_0$, there is no recursive definition in (5): on one event of probability at least $1-\delta$ over the standardized draws, the GD guarantee holds for every pair of scales satisfying (5). With this unbalanced scaling, we establish the main result for GD.

Theorem 3.2 (General vector responses, GD)**.** *Fix $\delta \in (0,1)$. Consider a two-layer ReLU network with initialization* (4) *satisfying the scaling conditions in* (5)*. If the network width is near-linear,*

$$m \geq \Omega\left(\frac{Nn}{\lambda} \cdot \mathrm{poly}\left(\log\frac{Nnd\kappa}{\delta}\right)\right),$$

then GD with step size $\eta = 1/L$ achieves $\mathcal{L}(\mathbf{V}_T, \mathbf{W}_T) \leq \epsilon \cdot \mathcal{L}(\mathbf{V}_0, \mathbf{W}_0)$ in

$$T = O\left(n\kappa \cdot \log(1/\epsilon)\right)$$

iterations with probability at least $1-\delta$, where $L = \lambda_1(\mathbf{K}_0)$ is the largest eigenvalue of the initial empirical hidden-layer NTK component, λ and κ are the smallest eigenvalue and condition number of the limiting NTK (Assumption 2.3), respectively.

Theorem 3.2 shows that if the output layer is sufficiently large compared to the hidden layer, then GD exhibits global linear convergence at a near-linear width. Initialization alone requires $\widetilde{\Omega}(\|\mathbf{X}\|^2 n/\lambda)$ width (Lemma D.2); the trajectory bounds still depend on N.

For NAG, analogous scale-separation conditions yield acceleration while retaining the same near-linear width order:

$$\frac{c_1^3}{c_2} \geq \widetilde{\Omega}\left(\frac{N\sqrt{d}}{\sqrt{n}\lambda}\right), \quad \frac{c_1}{c_2} \geq \widetilde{\Omega}\left(\frac{\sqrt{d}}{\sqrt{n}} \vee \frac{\sqrt{Nd}}{\sqrt{\lambda}}\right), \quad c_1c_2 \geq \widetilde{\Omega}\left(\frac{\sqrt{Nn}}{\lambda}\right). \tag{6}$$

These conditions have the same scale-separation interpretation as in GD; the detailed knob-by-knob role is summarized below.

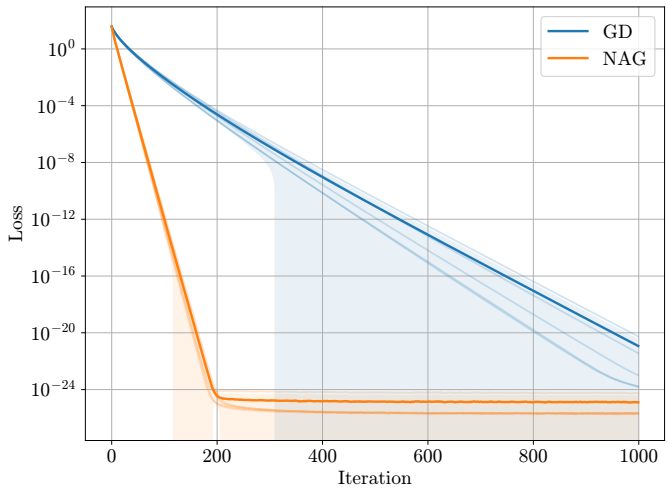


Figure 1: GD vs. NAG convergence. Bold curves show five-run mean loss, thin curves show individual trials, and bands show ± 1 standard deviation. NAG converges faster under the setting in Appendix B.

Theorem 3.3 (General vector responses, NAG)**.** *Fix $\delta \in (0,1)$. Consider a two-layer ReLU network with initialization* (4) *satisfying the scaling conditions in* (6)*. If the network width is near-linear,*

$$m \geq \Omega\Big(\frac{Nn}{\lambda} \cdot \mathrm{poly}\Big(\log \frac{Nnd\kappa}{\delta}\Big)\Big),$$

then NAG with step size $\eta = 1/L$ and momentum $\beta = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$ finds a solution satisfying $\mathcal{L}(\mathbf{V}_T, \mathbf{W}_T) \leq \epsilon \cdot \mathcal{L}(\mathbf{V}_0, \mathbf{W}_0)$ in

$$T = O\left(\sqrt{n\kappa} \cdot \log(1/\epsilon)\right)$$

iterations with probability at least $1 - \delta$, where $L = \lambda_1(\mathbf{K}_0)$ and $\mu = \lambda_{Nn}(\mathbf{K}_0)$ are the extreme eigenvalues of the initial empirical hidden-layer NTK component.

Theorem 3.3 shows that NAG accelerates the convergence rate from $O(n\kappa)$ to $O(\sqrt{n\kappa})$ within the same near-linear width regime. When $n = 1$, this recovers the usual $O(\kappa)$ versus $O(\sqrt{\kappa})$ dependence for scalar outputs. Numerical evidence of such an acceleration is shown in Figure 1, with experimental details in Appendix B.

How to read the scaling conditions. The headline statement of Theorems 3.2 and 3.3 is the near-linear width requirement $\widetilde{\Omega}(Nn/\lambda)$ together with the GD and NAG iteration bounds above.

The additional conditions on c_1, c_2 are sufficient scale-separation requirements used by the proof to realize the output-dominant NTK regime. The ratio c_1/c_2 suppresses CK-size perturbations, while the product $c_1 c_2$ sets the kernel scale and stability radius. Thus the two knobs should be read as "ratio controls perturbations" and "product sets scale and radius". The bookkeeping is centralized in Appendix A and verified in the theorem proofs; the conditions are not claimed to be optimized.

Remark 3.4 (Role of the width in the rate). The iteration bounds in Theorems 3.2 and 3.3 are written in terms of the realized initial kernel spectrum rather than an additional explicit factor of the width. Once m exceeds the threshold needed for concentration and kernel-shift control, the linearized rate is governed by $\kappa(\mathbf{K}_0) = \lambda_1(\mathbf{K}_0)/\lambda_{Nn}(\mathbf{K}_0)$. Indeed, Lemma D.2 gives $\lambda_{Nn}(\mathbf{K}_0) \gtrsim c_1^2 m\lambda$ and $\lambda_1(\mathbf{K}_0) \lesssim c_1^2 mn\lambda_1(\overline{\mathbf{K}})$, so the common factor m cancels in the condition number. Increasing m can still improve the high-probability concentration event and the realized conditioning, but above the required threshold it does not enter the theorem as a separate multiplicative rate factor. The factor n in the bound $\kappa(\mathbf{K}_0) = O(n\kappa)$ comes from $\mathbf{A}_{i,0} \preceq \|[\mathbf{V}_0]_{:,i}\|^2 \mathbf{I}_n$, whereas $\kappa(\mathbb{E}[\mathbf{K}_0]) = \kappa$ at fixed scales. At the same leading total width, n independent scalar heads have GD and NAG iteration bounds of $O(\kappa \log(1/\epsilon))$ and $O(\sqrt{\kappa}\log(1/\epsilon))$ per head. We do not establish that the extra shared-layer factors n and $\sqrt{n}$ are necessary.

Remark 3.5. In our proof, the unbalanced scaling permits a relative hidden-layer movement of at most $R_2/R_W = \widetilde{O}(\sqrt{N}/(m\sqrt{d}))$, or $\widetilde{O}(\lambda/(\sqrt{N}\,n\sqrt{d}))$ at the stated width. For comparison, the opposite scaling in Nguyen [2021] leads to the random feature regime, placing the dynamics in a CK-driven, nearly fixed-feature regime (see Figure 3 in Appendix B).

Remark 3.6. The smallest NTK eigenvalue $\lambda \coloneqq \lambda_N(\overline{\mathbf{K}})$ depends on the data separation $\Delta \coloneqq \min\{1/2, \min_{i\neq j} \min(\|\mathbf{x}_i - \mathbf{x}_j\|, \|\mathbf{x}_i + \mathbf{x}_j\|)\}$. For $d \geq 3$ and $\Delta > 0$, Lemma 7 in Karhadkar et al. [2024], together with the Schur product inequality, gives $\lambda \gtrsim \left(1 + \frac{d\log(1/\Delta)}{\log d}\right)^{-3}\Delta^2$. For strongly collinear data, small λ and large κ can make the sufficient width and iteration bounds quantitatively weak or vacuous. Kernel nondegeneracy is needed for uniform contraction of arbitrary residuals in this framework; we do not prove that the specific $1/\lambda$ width dependence is optimal.

3.1 Proof Overview for General Responses

The proof compares the nonlinear residual dynamics with fixed linear systems driven by the initial hidden-layer NTK component $\mathbf{K}_0$. For either initialization, write $R_V \coloneqq \max_{i\in[2m]} \|[\mathbf{V}_0]_{:,i}\|$, $R_W \coloneqq \max_{i\in[2m]} \|[\mathbf{W}_0]_{:,i}\|$, $\mathbf{R}_t \coloneqq \mathbf{R}(\mathbf{V}_t, \mathbf{W}_t)$, $\mathbf{r}_t \coloneqq \mathrm{vec}(\mathbf{R}_t)$, $\mathbf{P}_t \coloneqq \mathbf{V}_{t+1} - \mathbf{V}_t$, and $\mathbf{Q}_t \coloneqq \mathbf{W}_{t+1} - \mathbf{W}_t$. The exact NAG residual recursion has the form

$$\begin{pmatrix}\mathbf{r}_{t+1}\\ \mathbf{r}_t\end{pmatrix} = \mathbf{T}_{\mathrm{NAG}}\begin{pmatrix}\mathbf{r}_t\\ \mathbf{r}_{t-1}\end{pmatrix} + \begin{pmatrix}\boldsymbol{\xi}_t\\ 0\end{pmatrix}, \qquad \mathbf{T}_{\mathrm{NAG}} \coloneqq \begin{pmatrix}(1+\beta)(\mathbf{I}_{Nn} - \eta\mathbf{K}_0) & -\beta(\mathbf{I}_{Nn} - \eta\mathbf{K}_0)\\ \mathbf{I}_{Nn} & 0\end{pmatrix}, \tag{7}$$

where $\boldsymbol{\xi}_t$ contains the nonlinear, CK, and activation-pattern-change errors. Setting $\beta = 0$ gives the GD recursion $\mathbf{r}_{t+1} = (\mathbf{I}_{Nn} - \eta\mathbf{K}_0)\mathbf{r}_t + \boldsymbol{\xi}_t$.

The proof has five ingredients with full details in the appendix. First, coupled initialization gives $\|\mathbf{r}_0\| = \|\mathbf{Y}\|_{\mathrm{F}}$, radius bounds for R_V, R_W, and matrix-Chernoff concentration of $\mathbf{K}_0$, yielding $\kappa(\mathbf{K}_0) = O(n\kappa)$ at width $\widetilde{\Omega}(Nn/\lambda)$ (Lemmas D.1 and D.2, proofs in Appendices D.2 and D.3). Second, the exact nonlinear residual recursion is decomposed as (7), and the associated fixed systems contract: GD in Euclidean norm and NAG in a weighted quadratic norm (Proposition E.1 and lemma E.2, proofs in Appendices E.2 and E.3). Third, a complementary time-uniform kernel-shift event controls $\mathbf{K}_t - \mathbf{K}_0$ and the lower-order CK term $\mathbf{G}_t$ over the radius ball around initialization (Lemma F.2). Fourth, a two-sample sign-flip bound controls the ReLU activation changes without a union bound over time and yields a uniform Jacobian bound (Lemmas H.6 and H.7). Finally, the Jacobian bound and reference dynamics enter an induction that closes the residual-error and weight-stability estimates (Lemmas H.4 and H.8); the GD and NAG assemblies verify that (5)–(6) imply all proof requirements (Appendices H.6 and H.7, with the requirement ledger in Appendix A).

For orientation, the main estimates are as follows. With high probability,

$$\lambda_{Nn}(\mathbf{K}_0) \gtrsim c_1^2 m\lambda, \qquad \lambda_1(\mathbf{K}_0) \lesssim c_1^2 mn\lambda_1(\overline{\mathbf{K}}),$$

so the fixed-system condition number is $O(n\kappa)$. On any trajectory satisfying $\|[\mathbf{V}_t]_{:,i} - [\mathbf{V}_0]_{:,i}\| \leq R_1$ and $\|[\mathbf{W}_t]_{:,i} - [\mathbf{W}_0]_{:,i}\| \leq R_2 < c_2$, the ReLU-specific stability estimate (Lemma F.2) gives

$$\|\mathbf{K}_t - \mathbf{K}_0\| \lesssim mNR_V^2(R_2/c_2) + mNR_1(2R_V + R_1), \qquad \|\mathbf{G}_t\| \lesssim m\,\|\mathbf{X}\|^2\,(R_W + R_2)^2.$$

Together with the sign-flip bound $|S_j^{\perp}| \lesssim mR_2/c_2$, where $S_j^{\perp}$ denotes neurons whose activation on $\mathbf{x}_j$ can differ from initialization inside the radius ball, these inequalities control the terms in $\boldsymbol{\xi}_t$. Except for identical or nearly antipodal samples, the joint near-zero margin probability is $O((R_2/c_2)^2/\lambda)$ (Lemma H.6). This gives the sharper Jacobian bound in Lemma H.7, which allows the NAG induction to track the residual and scaled momentum together.

This organization is also why the scale conditions can be checked without circularity. The condition number of $\mathbf{K}_0$ is invariant under the common c_1^2 factor in the hidden-layer NTK component, so the width assumption first gives $\kappa(\mathbf{K}_0) = O(n\kappa)$ with high probability; after that, the ratio c_1/c_2 and product c_1c_2 are chosen to make the perturbation terms smaller than the contraction margin. In particular, the CK term is never used as the leading contraction mechanism. It enters only through the perturbation budget, while the leading reference dynamics remain governed by $\mathbf{K}_0$. The low-rank result below reuses exactly this perturbation mechanism, but replaces the ambient output space by the response subspace and uses the reference contraction only on the positive-eigenvalue subspace of the sketched kernel.

The Kronecker concentration, shared-feature Jacobian estimate and response-subspace argument are specific to this setting; the sign-flip and linear momentum analyses build on Wang et al. [2021], Liu et al. [2022b].

4 Results for Low-Rank Responses

The general vector-output bounds in Section 3 scale with the ambient output dimension n. When the response matrix has rank $r \ll n$, this dependence can be overly pessimistic. We exploit the response subspace by adding a sketching step to the output-layer initialization:

$$\mathbf{V}_0 = (c_1\mathbf{Y}\mathbf{\Phi}, -c_1\mathbf{Y}\mathbf{\Phi}), \quad \mathbf{W}_0 = (c_2\mathbf{\Psi}, c_2\mathbf{\Psi}), \tag{8}$$

where $\mathbf{\Phi} \in \mathbb{R}^{N\times m}$ and $\mathbf{\Psi} \in \mathbb{R}^{d\times m}$ are independent standard Gaussian random matrices. Compared to (4), the output weights now contain a random projection of $\mathbf{Y}$, so the low-dimensional response structure is built into the network at initialization. This label-dependent sketch uses the observed $\mathbf{Y}$ directly; no separate response basis is required.

When $r < n$, the price of this rank-adaptive initialization is that the initial empirical kernel $\mathbf{K}_0$ becomes rank-deficient. Hence the idealized linear dynamics cannot be a strict contraction on all of $\mathbb{R}^{Nn}$: on $\ker(\mathbf{K}_0)$, the lifted NAG block is $\left(\begin{smallmatrix} 1+\beta & -\beta \\ 1 & 0 \end{smallmatrix}\right)$ and fixes states $(\mathbf{u}^\top, \mathbf{u}^\top)^\top$. The right object is therefore the positive-eigenvalue subspace $\mathcal{E} \coloneqq \mathrm{col}(\mathbf{K}_0)$. Since $\mathbf{K}_0$ is symmetric PSD, $\mathbb{R}^{Nn} = \mathrm{col}(\mathbf{K}_0) \oplus \ker(\mathbf{K}_0)$ and $\mathrm{col}(\mathbf{K}_0) = \ker(\mathbf{K}_0)^\perp$; Lemma 4.6 later identifies this subspace as $\mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$.

Our analysis proves that the residual dynamics remain in this invariant subspace and converge to zero. The sufficient scaling conditions are

$$\frac{c_1^3}{c_2} \geq \widetilde{\Omega}\left(\frac{\sqrt{Nd}(\kappa(\mathbf{Y}) \vee \sqrt{N})}{\sqrt{r}\lambda\sigma_r^2(\mathbf{Y})\,\|\mathbf{Y}\|}\right),\ \frac{c_1}{c_2} \geq \widetilde{\Omega}\left(\frac{\sqrt{d}}{\sqrt{r}\,\|\mathbf{Y}\|} \vee \frac{\sqrt{Nd}}{\sqrt{\lambda}\sigma_r(\mathbf{Y})}\right),\ c_1c_2 \geq \widetilde{\Omega}\left(\frac{\sqrt{Nr}\kappa(\mathbf{Y})}{\lambda\sigma_r(\mathbf{Y})}\right). \tag{9}$$

As in Section 3, the ratio condition separates the hidden-layer NTK contribution from CK-size perturbations, while the product condition sets the overall kernel scale and stability radius. The stronger term in the ratio condition is a conservative sufficient condition used to control a CK-size contribution in the NAG perturbation bound; it is not claimed to be necessary or optimized.

Theorem 4.1 (Low-rank vector responses)**.** *Fix $\delta \in (0,1)$. Consider nonzero responses of rank $r \geq 1$ and a two-layer ReLU network with sketching initialization* (8) *satisfying the scaling conditions in* (9)*. If the network width is near-linear in both the sample size N and the rank r,*

$$m \geq \Omega\left(\frac{Nr\kappa^2(\mathbf{Y})}{\lambda} \cdot \mathrm{poly}\left(\log\frac{Nnrd\kappa\kappa(\mathbf{Y})}{\delta}\right)\right),$$

then NAG with step size $\eta = 1/L$ and momentum $\beta = (\sqrt{L}-\sqrt{\mu})/(\sqrt{L}+\sqrt{\mu})$ finds a solution satisfying $\mathcal{L}(\mathbf{V}_T, \mathbf{W}_T) \leq \epsilon \cdot \mathcal{L}(\mathbf{V}_0, \mathbf{W}_0)$ in

$$T = O\left(\sqrt{r\kappa}\kappa(\mathbf{Y}) \cdot \log(1/\epsilon)\right)$$

iterations with probability at least $1-\delta$, where $r = \mathrm{rank}(\mathbf{Y})$, $L = \lambda_1(\mathbf{K}_0)$, $\mu = \lambda_{Nr}(\mathbf{K}_0)$ is the smallest positive eigenvalue of $\mathbf{K}_0$, and κ is the limiting NTK condition number from Assumption 2.3.

Theorem 4.1 should be read as a rank-adaptive analogue of the full-rank NAG result. The leading polynomial width factor changes from Nn to $Nr\kappa^2(\mathbf{Y})$, and the accelerated iteration count changes from $\sqrt{n\kappa}$ to $\sqrt{r\kappa}\kappa(\mathbf{Y})$. Thus, when $\kappa(\mathbf{Y}) = O(1)$, the leading dependence matches Theorem 3.3 up to polylogarithmic factors with n replaced by r. This well-conditioned low-rank situation occurs, for example, when $r \ll \min(N,n)$ and the observed responses are r-dimensional Gaussian responses randomly embedded into n-dimensional Euclidean space [Vershynin, 2010]. The leading width improves only if $r\kappa^2(\mathbf{Y}) < n$, not when $\kappa^2(\mathbf{Y}) \geq n/r$; e.g., $n = 100$, $r = 5$, $\kappa(\mathbf{Y}) = 10$ give $500 > 100$.

The condition number κ in the rate is the limiting NTK condition number $\kappa(\overline{\mathbf{K}})$, while the proof uses the realized nonzero-spectrum condition number $\kappa(\mathbf{K}_0) = \lambda_1(\mathbf{K}_0)/\lambda_{Nr}(\mathbf{K}_0)$ on $\mathcal{E} = \mathrm{col}(\mathbf{K}_0)$. By Lemma 4.5, $\kappa(\mathbf{K}_0) = O(r\kappa\kappa^2(\mathbf{Y}))$ with high probability. The extra response-conditioning factors enter because the sketching initialization must preserve the rank-r response subspace and because contraction is proved only on the positive-eigenvalue subspace of $\mathbf{K}_0$.

Remark 4.2 (Approximately low-rank responses). Theorem 4.1 assumes exact low rank so that $\mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$ is an invariant subspace of the residual dynamics. If $\mathbf{Y} = \mathbf{Y}_r + \mathbf{E}$ with $\mathrm{rank}(\mathbf{Y}_r) = r$ and small residual component $\mathbf{E}$, the full noisy sketch need not preserve the rank-r subspace. With a truncated sketch based on $\mathbf{Y}_r$, one might expect accelerated convergence toward an error floor controlled by $\|\mathbf{E}\|_{\mathrm{F}}$, rather than exact convergence to zero loss. Proving such a perturbation result requires tracking leakage out of the rank-r subspace and is left for future work.

4.1 Proof Overview for Low-Rank Responses

The low-rank proof has one additional block before the general induction from Section 3.1: identify the right subspace and show that the residual recursion never leaves it. We first prove that the positive eigenspace of $\mathbf{K}_0$ is exactly $\mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$ (Lemma 4.6). We then show by induction that both the residual $\mathbf{r}_t$ and the perturbation $\boldsymbol{\xi}_t$ remain in this subspace for all $t \geq 0$ (Lemma 4.7). With this invariance in place, the NAG reference operator only needs to contract on $\mathcal{E} \times \mathcal{E}$ (Lemma 4.8); the remaining Jacobian and weight-change controls are inherited from the general proof in reduced coordinates.

When $r < n$, this subspace step is essential: $\mathbf{K}_0$ is singular on the ambient Nn-dimensional space, but the initial residual and all perturbations stay in $\mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$. Hence the proof needs stability only on the dynamically reachable positive-eigenvalue directions, which is what replaces the ambient dimension n by the intrinsic rank r.

The rest of this section lists the proof inputs used for this subspace argument and records where their full derivations appear. We first provide the initialization-radius and subspace-preservation bound for the sketched initialization (8).

Lemma 4.3. *Suppose $\mathbf{V}_0$ and $\mathbf{W}_0$ are initialized as* (8)*, $m \geq r$, and $\tau \geq 0$ (interpreting the bound at zero by continuity). For $\delta_0 \in (0,1)$, set $\ell_0 = \log(4m/\delta_0)$. Then with probability at least $1 - \delta_0/2 - 2e^{-\frac{9m(r\wedge d)}{64}} - 2e^{-\min\{(m-r+1)\log\frac{1}{c\tau}, c'm\}}$,*

$$\tfrac{1}{2}c_1\sigma_r(\mathbf{Y})\sqrt{r} \leq R_V \leq c_1\|\mathbf{Y}\|(\sqrt{r}+\sqrt{2\ell_0}), \qquad \tfrac{1}{2}c_2\sqrt{d} \leq R_W \leq c_2(\sqrt{d}+\sqrt{2\ell_0}),$$
$$\sigma_r(\mathbf{V}_0) \geq c_1\sqrt{2}\tau\left(\sqrt{m}-\sqrt{r-1}\right)\sigma_r(\mathbf{Y}).$$

If the failure probability is small for some $\tau > 0$ and $m \geq r$, then $\mathrm{col}(\mathbf{V}_0) = \mathrm{col}(\mathbf{Y})$.

Corollary 4.4 (τ-free initialization radius event)**.** *Under the sketching initialization* (8)*, for any $\delta_0 \in (0,1)$ the first two bounds in Lemma 4.3 on R_V and R_W hold with failure probability at most $\delta_0/2 + 2e^{-\frac{9m(r\wedge d)}{64}}$, with $\ell_0 = \log(4m/\delta_0)$ and independently of τ. The low-rank NAG proof consumes only this radius event from Lemma 4.3.*

Proof. The bounds on R_V and R_W in Lemma 4.3 follow from the concentration steps in Appendix D.4 before the smallest-singular-value estimate is invoked. Those two steps contribute only the failure terms $\delta_0/2$ and $2e^{-\frac{9m(r\wedge d)}{64}}$, and do not involve τ. □

The proof of Lemma 4.3 is provided in Appendix D.4. The bound on R_V in Lemma 4.3 depends on r, the response singular values, and $\log(4m/\delta_0)$; $\mathbf{V}_0$ keeps the column space of $\mathbf{Y}$ with high probability. Equipped with these results, we show the corresponding initial-kernel concentration bound, with complete proof in Appendix D.5.

Lemma 4.5 (Eigenvalues of $\mathbf{K}_0$)**.** *Suppose $\mathbf{V}_0$ and $\mathbf{W}_0$ are as* (8)*, $\delta \in (0,1)$, if the width satisfies*

$$m \geq \widetilde{\Omega}\left(\frac{\|\mathbf{X}\|^2 r\kappa^2(\mathbf{Y})}{\lambda}\log\frac{Nr}{\delta}\right),$$

then $\lambda_{Nr}(\mathbf{K}_0) \geq \frac{1}{2}c_1^2 m\lambda\sigma_r^2(\mathbf{Y})$ and $\lambda_1(\mathbf{K}_0) \leq 9c_1^2 mr\lambda_1(\overline{\mathbf{K}})\|\mathbf{Y}\|^2$ with probability at least $1-\delta$.

By Lemma 4.5 and the range identity below, $\mathrm{rank}(\mathbf{K}_0) = Nr$ and $\kappa(\mathbf{K}_0) = O(r\kappa\kappa^2(\mathbf{Y}))$ with high probability. We characterize the positive eigensubspace of $\mathbf{K}_0$ and show its connection to the column space of $\mathbf{Y}$.

Lemma 4.6 (Eigensubspace of $\mathbf{K}_0$)**.** *Assume the conclusions of Lemma 4.5 hold. Then*

$$\ker(\mathbf{K}_0) \perp \mathrm{row}(\mathbf{K}_0) = \mathrm{col}(\mathbf{K}_0) = \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y}).$$

We further show that $\mathbf{r}_t$ and $\boldsymbol{\xi}_t$ are in the positive eigensubspace of $\mathbf{K}_0$ for all $t \geq 0$.

Lemma 4.7. *Assume the conclusions of Lemma 4.5 hold. Then* $\{\mathbf{r}_t, \boldsymbol{\xi}_t\}_{t\geq 0} \subseteq \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$.

The proofs of Lemmas 4.6 and 4.7 are provided in Appendices G.1 and G.2. Consequently, with the standard zero-momentum lifted initialization convention $\mathbf{r}_{-1} = \mathbf{r}_0$, the lifted state $\mathbf{z}_t = (\mathbf{r}_t^\top, \mathbf{r}_{t-1}^\top)^\top$ and perturbation $(\boldsymbol{\xi}_t^\top, 0)^\top$ both lie in $\mathcal{E} \times \mathcal{E}$ for every t. With these two lemmas, the low-rank residual recursion is restricted to the subspace on which we prove contraction of the NAG reference operator.

Lemma 4.8 (Low-rank contraction)**.** *On the event in Lemma 4.5, denote* $L \coloneqq \lambda_1(\mathbf{K}_0)$, $\mu \coloneqq \lambda_{Nr}(\mathbf{K}_0) > 0$, $\kappa(\mathbf{K}_0) = \frac{L}{\mu}$. *Let* $\mathcal{E} \coloneqq \mathrm{col}(\mathbf{K}_0)$ *be the positive-eigenvalue subspace of* $\mathbf{K}_0$*; by Lemma 4.6, in the present low-rank setting* $\mathcal{E} = \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$. *For any fixed* $\alpha \in (0, 1)$*, by setting* $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ *and* $\eta = \frac{1}{L}$*, there exists a positive definite matrix* $\mathsf{P}_{\alpha,\mathcal{E}} \succ 0$ *on* $\mathcal{E} \times \mathcal{E}$ *such that* $\mathsf{P}_{\alpha,\mathcal{E}} \succeq \mathbf{I}_{\mathcal{E}\times\mathcal{E}}$ *and, for any* $\mathbf{u}, \mathbf{v} \in \mathcal{E}$*,*

$$\left\| \mathbf{T}_{\mathrm{NAG}} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} \right\|_{\mathsf{P}_{\alpha,\mathcal{E}}} \leq \left(1 - \frac{\alpha}{\sqrt{\kappa(\mathbf{K}_0)}} \right) \left\| \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} \right\|_{\mathsf{P}_{\alpha,\mathcal{E}}}.$$

Moreover, $\lambda_1(\mathsf{P}_{\alpha,\mathcal{E}}) \leq C_\alpha \kappa(\mathbf{K}_0)^{3/2}$ *for the same type of constant* C_α *depending only on* α.

The proof is provided in Appendix E.4. After the subspace issue is resolved, the rest of the proof follows the induction architecture from Section 3.1. The sketching version of the kernel-shift argument is given in Appendix F.3; the weight-change bounds in Lemma H.4 are applied in the reduced output coordinates. The complete assembly of these inputs, including the choices of c_1, c_2, m and the requirement ledger, is provided in Appendix H.8.

5 Conclusion and Future Directions

We establish convergence guarantees for two-layer ReLU networks that simultaneously achieve near-linear width and provable acceleration in an NTK-dominated regime. Our analysis, enabled by an output-dominant unbalanced Gaussian initialization, demonstrates that: (i) a near-linear width of $\widetilde{\Omega}(Nn/\lambda)$ is sufficient for GD to converge to zero training loss; (ii) NAG provably accelerates convergence to an $O(\sqrt{n\kappa})$ rate without requiring additional width; and (iii) for low-rank responses, a sketching technique adapts the width requirement to $\widetilde{\Omega}(Nr\kappa^2(\mathbf{Y})/\lambda)$, depending on the intrinsic rank $r \ll n$ and the response conditioning. These findings sharpen the theory for training non-smooth neural networks in the NTK regime while complementing existing linear-width results in random-feature regimes.

Our guarantees require exact coupling; they do not establish robustness to imperfect coupling or superiority over He or Xavier initialization.

Our work lays a rigorous foundation for some critical avenues of future research. One is to extend our optimization analysis to the study of generalization. While our focus is on optimization, the ability to guarantee convergence under mild overparameterization provides a crucial stepping stone. Understanding how the choice of unbalanced scaling impacts the learned features and subsequent generalization is an important open question that our framework is well-suited to explore.

Furthermore, our analysis considers full-batch gradient methods. Extending these near-linear width guarantees to the stochastic setting (SGD) is a highly challenging but vital next step for aligning theory with modern practice. Similarly, adapting our techniques from shallow networks under square loss to deeper architectures and other loss functions, such as cross-entropy, would broaden the applicability of these results. A direct extension to deep networks is not automatic: activation-pattern changes then couple across layers, and the empirical kernel evolves through products of layerwise features rather than the single shared first-layer structure used here. Understanding whether an output-dominant scaling can still control this multi-layer kernel evolution is an important open problem. We believe the tight control over kernel dynamics demonstrated in our work provides a robust analytical toolkit for tackling these future challenges. The proof is intentionally modular: the appendix ledgers separate initialization, contraction, kernel stability, and induction requirements, so future refinements of the scale conditions can target individual blocks without changing the main contraction architecture.

Acknowledgments and Disclosure of Funding

MT is grateful for partial supports by NSF Grant DMS-2513699, DOE Grant SC0026274, and Richard Duke Fellowship. MT also thanks Simons Institute of the Theory of Computing.

References

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

Mauricio A Alvarez, Lorenzo Rosasco, Neil D Lawrence, et al. Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266, 2012.

Simone Bombari, Mohammad Hossein Amani, and Marco Mondelli. Memorization and optimization in deep neural networks with minimum over-parameterization. *Advances in Neural Information Processing Systems*, 35:7628–7640, 2022.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 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*, pages 1675–1685. PMLR, 2019.

Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*, 2018.

Alex Gittens and Joel A Tropp. Tail bounds for all eigenvalues of a sum of random matrices. *arXiv preprint arXiv:1104.4513*, 2011.

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

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

Kedar Karhadkar, Michael Murray, and Guido Montúfar. Bounds for the smallest eigenvalue of the ntk for arbitrary spherical data of arbitrary dimension. *Advances in Neural Information Processing Systems*, 2024.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338, 2000.

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

Wenbo V Li and Ang Wei. Gaussian integrals involving absolute value functions. In *High dimensional probability V: the Luminy volume*, volume 5, pages 43–60. Institute of Mathematical Statistics, 2009.

Fangshuo Liao and Anastasios Kyrillidis. Provable accelerated convergence of nesterov's momentum for deep relu neural networks. In *International Conference on Algorithmic Learning Theory*, pages 732–784. PMLR, 2024.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59: 85–116, 2022a.

Xin Liu, Zhisong Pan, and Wei Tao. Provable convergence of nesterov's accelerated gradient method for over-parameterized neural networks. *Knowledge-Based Systems*, 251:109277, 2022b.

Charles A Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005.

Alexander Munteanu, Simon Omlor, Zhao Song, and David Woodruff. Bounding the width of neural networks via coupled initialization a worst case analysis. In *International Conference on Machine Learning*, pages 16083–16122. PMLR, 2022.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

Quynh Nguyen. On the proof of global convergence of gradient descent for deep relu networks with linear widths. In *International Conference on Machine Learning*, pages 8056–8062. PMLR, 2021.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1(1):84–105, 2020.

Mert Pilanci and Martin J Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, 17(53):1–38, 2016.

Prabhakar Raghavan. Probabilistic construction of deterministic algorithms: approximating packing integer programs. *Journal of Computer and System Sciences*, 37(2):130–143, 1988.

Mark Rudelson and Roman Vershynin. Smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 62(12):1707–1739, 2009.

Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix chernoff bound. *arXiv preprint arXiv:1906.03593*, 2019.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

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

Jun-Kun Wang, Chi-Heng Lin, and Jacob D Abernethy. A modular analysis of provable acceleration via polyak's momentum: Training a wide relu network and a deep linear network. In *International Conference on Machine Learning*, pages 10816–10827. PMLR, 2021.

Rachel Ward and Tamara Kolda. Convergence of alternating gradient descent for matrix factorization. *Advances in Neural Information Processing Systems*, 36:22369–22382, 2023.

Zhenghao Xu, Yuqing Wang, Tuo Zhao, Rachel Ward, and Molei Tao. Provable acceleration of nesterov's accelerated gradient for rectangular matrix factorization and linear neural networks. *Advances in Neural Information Processing Systems*, 2024.

Hongru Yang, Ziyu Jiang, Ruizhe Zhang, Yingbin Liang, and Zhangyang Wang. Neural networks with sparse activation induced by large bias: Tighter analysis with bias-generalized ntk. *Journal of Machine Learning Research*, 25(321):1–51, 2024.

A Assumptions

This section summarizes the assumptions and proof dependencies used by the three main convergence results. It separates theorem-level inputs from internal induction parameters so the theorem proofs can refer back to one ledger instead of repeating the same bookkeeping. For the low-rank result, the ledger also records the additional invariant-subspace block before the general induction architecture is invoked.

Conditioning glossary. Throughout the ledgers and proofs, $\kappa = \kappa(\overline{\mathbf{K}})$ always denotes the limiting NTK condition number from Assumption 2.3. The realized condition number $\kappa(\mathbf{K}_0)$ refers to the empirical initial kernel: in the full-rank results it means $\lambda_1(\mathbf{K}_0)/\lambda_{Nn}(\mathbf{K}_0)$ on $\mathbb{R}^{Nn}$, while in the low-rank result it means $\lambda_1(\mathbf{K}_0)/\lambda_{Nr}(\mathbf{K}_0)$ on the positive-eigenvalue subspace $\mathcal{E} = \mathrm{col}(\mathbf{K}_0)$. We keep $\kappa(\mathbf{K}_0)$ explicit, rather than abbreviating it, to avoid conflating realized and limiting conditioning. For the rank-r response matrix, κ_r in Table 1 is $\kappa_r = \kappa(\mathbf{Y}) = \sigma_1(\mathbf{Y})/\sigma_r(\mathbf{Y})$ and $\kappa(\mathbf{Y}\mathbf{Y}^\top) = \kappa^2(\mathbf{Y})$. The high-probability control of $\kappa(\mathbf{K}_0)$ is supplied by Lemma D.2 in the full-rank results and by Lemma 4.5 in the low-rank result, while Lemma 4.6 identifies the corresponding low-rank positive-eigenvalue subspace.

Confidence-dependent radius bounds. Use $\delta_0 = \delta/8$ in Lemma D.1 and corollary 4.4, so $\ell_\delta = \log(32m/\delta)$. For $m \geq C\log(32/\delta)$, the upper and lower radius failures each cost at most $\delta/16$. Allocate $\delta/4$ each to the initial-spectrum bound, the bound on $\|\mathbf{W}_0\|_{\mathrm{F}}$, and the two-sample margin event (Lemma H.6). The total failure probability is at most $7\delta/8$.

We also use the following log self-bounding convention in the probability ledgers. A condition $m \geq CD\log^4(e + 128mN^2n^2/\delta)$ is implied by $m \geq C'D\log^4(e + C'DN^2n^2/\delta)$ for sufficiently large C', since $m/\log^4(e + 128mN^2n^2/\delta)$ is eventually increasing. Here $D = Nn/\lambda$ or $Nr\kappa^2(\mathbf{Y})/\lambda$; $1/\lambda \leq 2\kappa$ since $\lambda_1(\overline{\mathbf{K}}) \geq 1/2$.

A.1 General Vector Responses, GD

Theorem 3.2 uses the following inputs.

- **Model and data.** Two-layer ReLU network under square loss (1), normalized data in Assumption 2.1, positive limiting NTK eigenvalue in Assumption 2.3, and unsketched coupled initialization (4).
- **Algorithm parameters.** GD with $\eta = 1/\lambda_1(\mathbf{K}_0)$, equivalently $\beta = 0$ in Proposition E.1.
- **Quantitative scaling.** Width and scaling conditions are those in Theorem 3.2 and (5). The auxiliary radii R_1, R_2 are chosen inside the proof and are not additional theorem assumptions.
- **Probability and log accounting.** The theorem-level polylogarithmic width absorbs the $\log N, \log n, \log m$, and $\log(1/\delta)$ factors in (17). The event in Lemma H.6 is uniform over radius-bounded iterates, so no time-union bound is required.
- **Kernel conditioning.** The initial kernel is full rank on $\mathbb{R}^{Nn}$ by Lemma D.2; $\kappa(\mathbf{K}_0) = \lambda_1(\mathbf{K}_0)/\lambda_{Nn}(\mathbf{K}_0)$.
- **Proof dependencies.** The proof uses Lemmas D.1, D.2, H.6 and H.7, the weight-movement budgets in (20), and the one-step GD loss bound in Appendix H.6.

A.2 General Vector Responses, NAG

Theorem 3.3 uses the same model, data, and unsketched initialization assumptions as Theorem 3.2, with the following changes.

- **Algorithm parameters.** NAG uses $\eta = 1/L$ and $\beta = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$, where $L = \lambda_1(\mathbf{K}_0)$ and $\mu = \lambda_{Nn}(\mathbf{K}_0)$.
- **Quantitative scaling.** Width and scaling conditions are those in Theorem 3.3 and (6); R_1, R_2 are internal induction radii.
- **Probability and log accounting.** The same initialization and time-uniform Jacobian events as in the GD proof are used; all auxiliary failure probabilities are constant fractions of δ.

- **Kernel conditioning and norm.** The weighted contraction in Lemma E.2 describes the reference residual dynamics. The nonlinear proof controls the residual and scaled momentum together in Lemma H.8.
- **Proof dependencies.** In addition to the common initialization and Jacobian bounds, Lemma H.8 controls NAG, and (20) closes the weight induction in Appendix H.7.

A.3 Low-Rank Vector Responses, NAG

Theorem 4.1 uses the following inputs; Table 2 summarizes their roles.

- **Model and data.** Two-layer ReLU network under square loss (1), normalized data in Assumption 2.1, positive limiting NTK eigenvalue in Assumption 2.3, rank-r response matrix $\mathbf{Y}$, and sketching initialization (8).
- **Algorithm parameters.** NAG uses $\eta = 1/L$ and $\beta = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$, where $L = \lambda_1(\mathbf{K}_0)$ and $\mu = \lambda_{Nr}(\mathbf{K}_0)$ is the smallest positive eigenvalue.
- **Quantitative scaling.** Width and scaling conditions are those in Theorem 4.1 and (9); R_1, R_2 are internal induction radii.
- **Probability and log accounting.** The sketching proof inherits the time-uniform radius-ball event from Lemma H.6, since the hidden-layer initialization has the same law as in the unsketched case. The factor $\kappa^2(\mathbf{Y})$ in the width comes from the initial spectrum.
- **Kernel subspace and conditioning.** The effective subspace is $\mathcal{E} = \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y}) = \mathrm{col}(\mathbf{K}_0)$ by Lemma 4.6. The condition number $\kappa(\mathbf{K}_0) = \lambda_1(\mathbf{K}_0)/\lambda_{Nr}(\mathbf{K}_0)$ is defined on this positive eigenspace; no strict contraction is claimed on $\ker(\mathbf{K}_0)$.
- **Proof dependencies.** In coordinates on $\mathrm{col}(\mathbf{Y})$ the general Jacobian and NAG estimates apply with the sketched radius and spectrum bounds, as detailed below.

Table 2: Dependencies in the proof of Theorem 4.1.

Dependency block	Supplies	Consumed in Appendix H.8
Initialization and spectrum	τ-free R_V, R_W bounds from Corollary 4.4; nonzero-spectrum bounds from Lemma 4.5	Initial radii, L, μ, and theorem-level probability budget
Effective subspace	$\mathcal{E} = \mathrm{col}(\mathbf{K}_0) = \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$ from Lemma 4.6; invariance of $\mathbf{r}_t, \boldsymbol{\xi}_t$ from Lemma 4.7	Dynamics in r output coordinates
Gate changes	Joint count of near-zero margins from Lemma H.6	Jacobian bound uniform over the weight balls in Lemma H.7
NAG stability	Response to perturbations of the Jacobian from Lemma H.8	Residual decay on the effective subspace
Induction closure	Columnwise movement bounds in (20)	Weight radii at the next step, before applying the Jacobian bound

A.4 From the internal budgets to the displayed scales

For general responses set $(k, a, b) = (n, 1, 1)$; for low-rank responses set $(k, a, b) = (r, \|\mathbf{Y}\|, \sigma_r(\mathbf{Y}))$. Write $q = c_1/c_2$, $p = c_1 c_2$, and $\ell = \log(e + 128mN^2n^2/\delta)$. The radii in (18) are $h = R_2/c_2 = \sqrt{N}\ell^2/m$ and $R_1 = hc_1 b\sqrt{k}$. Table 3 gives the corresponding width and scale choices.

For GD the product bound in (5) also includes $\kappa(\mathbf{K}_0) \geq 1$. For choosing scales, if the three lower bounds on c_1^3/c_2, c_1/c_2 and $c_1 c_2$ are A_0, B_0, C_0, respectively, take $p \geq C_0$ and $q \geq \max\{B_0, \sqrt{A_0/p}\}$, then $c_1 = \sqrt{pq}$ and $c_2 = \sqrt{p/q}$.

The empirical choices are $p = 1, q = 100$ for the convergence and low-rank experiments, and $p = 1, q \in \{0.1, 0.3, 1, 3, 10\}$ for the regime sweep. Since the asymptotic conditions contain unspecified constants, these empirical choices diagnose behavior but do not certify every theorem condition.

Table 3: Sufficient width and scale choices for the Jacobian and weight-movement bounds.

Internal requirement	Sufficient choice
Gate and output-weight changes in the hidden-weight Jacobian (Lemma H.7)	$m \geq CNk(a/b)^2\ell^4/\lambda$
CK contribution from the output-weight Jacobian	$q \geq C\sqrt{Nd}/(b\sqrt{\lambda})$
Hidden-weight movement in (20)	$p \geq Ca\sqrt{Nk}/(\lambda b^2)$, since $mh = \sqrt{N}\ell^2$
Output-weight movement in (20)	The same product and ratio bounds, using $c_1^3/c_2 = pq^2$

B Numerical Experiments

This section provides supporting numerical experiments.

Firstly, we compare the convergence rates of GD and NAG. We set input dimension $d = 20$, output dimension $n = 10$, sample size $N = 100$, and width $2m = 2000$. We generate inputs $\{\mathbf{x}_i\}_{i=1}^N$ and responses $\{\mathbf{y}_i\}_{i=1}^N$ independently from the uniform distributions on $\mathbb{S}^{d-1}$ and $\mathbb{S}^{n-1}$, respectively. We apply initialization (4) with $c_1 = 10$ and $c_2 = 0.1$. We choose step size $\eta = \frac{1}{L}$ for GD and NAG, and, for NAG, momentum parameter $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$, according to Theorems 3.2 and 3.3, with L and μ the extreme eigenvalues of $\mathbf{K}_0$. We run 5 trials independently and plot the GD and NAG loss curves in Figure 1. NAG reaches low training loss in fewer iterations than GD.

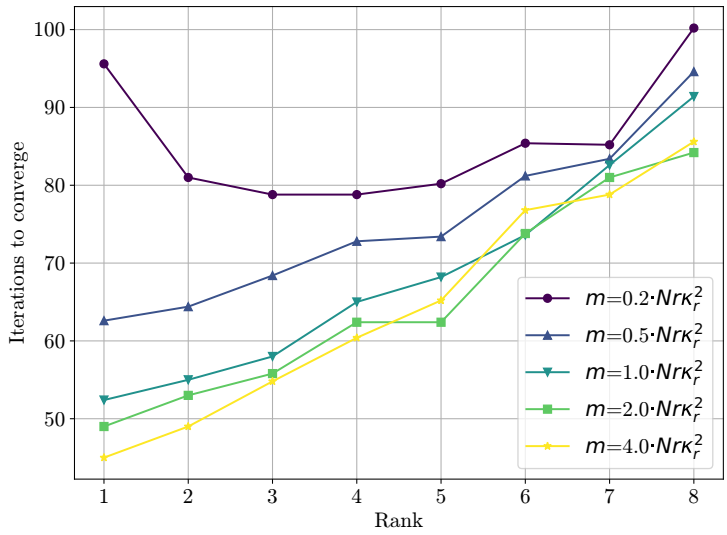


Figure 2: NAG iterations to training loss below 10^{-6} versus response rank. Each curve shows the mean over five trials at a fixed width multiplier α. The curves for $\alpha = 2$ and 4 are close over the ranks considered.

Secondly, we study low-rank adaptivity when applying initialization (8). We keep d, N, c_1 and c_2 as before, while changing the output dimension to $n = 50$. We sample input data from the uniform distribution on the unit sphere $\mathbb{S}^{d-1}$. For low-rank responses, we first generate independent standard Gaussian random matrices $\widetilde{\mathbf{Y}}_1 \in \mathbb{R}^{n\times r}$ and $\widetilde{\mathbf{Y}}_2 \in \mathbb{R}^{N\times r}$, then form $\widetilde{\mathbf{Y}} = \widetilde{\mathbf{Y}}_1\widetilde{\mathbf{Y}}_2^\top$, and finally normalize each column of $\widetilde{\mathbf{Y}}$ to get $\mathbf{Y}$ so that $\|\mathbf{y}_i\| = 1$, $\forall i \in [N]$. The rank r is chosen from 1 to 8. We scale the half-width m as $\alpha Nr\kappa_r^2$, where $\kappa_r = \kappa(\mathbf{Y})$ denotes the condition number on the nonzero response spectrum. The design uses $\alpha \in \{1/4, 1/2, 1, 2, 4\}$; the smallest multiplier is labeled 0.2 in Figure 2. We apply NAG with η and β specified in Theorem 4.1. We run 5 trials independently and plot the average number of iterations to reach training loss below 10^{-6} for each α and r in Figure 2. For the larger widths, the iteration count grows with rank, and increasing α from 2 to 4 changes the count little. For $\alpha = 1$, the network width $2m$ ranges from 200 to approximately

7000, compared with the ambient-dimensional scale $2Nn = 10000$. This comparison illustrates the rank dependence; it does not establish a finite-sample width threshold.

Thirdly, we vary the initialization ratio to compare CK-like (random-feature) and NTK-like training within the coupled family. For this part, we set $N = 100$, $d = 20$, $n = 10$, $m = Nn$, and apply NAG with initialization (4) for $T = 100$ iterations. For this comparison, we set $\eta = \frac{1}{L'}$ and $\beta = \frac{\sqrt{L'}-\sqrt{\mu'}}{\sqrt{L'}+\sqrt{\mu'}}$ where $L' = \lambda_1(\mathbf{H}_0)$, $\mu' = \lambda_{Nn}(\mathbf{H}_0)$, and $\mathbf{H}_0 = \mathbf{K}_0 + \mathbf{G}_0$ is the total kernel matrix at initialization. We use $\mathbf{H}_0$ for this diagnostic because the linearized two-layer residual update is governed by the total initial kernel; the decomposition $\mathbf{H}_t = \mathbf{K}_t + \mathbf{G}_t$ separates the hidden-layer NTK component from the output-layer/CK component when interpreting which mechanism dominates. The theorems are stated in terms of the hidden-layer NTK component $\mathbf{K}_0$ because the proof controls its kernel-shift and activation-pattern changes; on the NTK-like side of this diagnostic, $\mathbf{K}_0$ is the dominant component of $\mathbf{H}_0$, so $\mathbf{H}_0$-based tuning is a total-kernel analogue of the $\mathbf{K}_0$-based theoretical scaling. Since $\mathbf{G}_0 \succeq 0$, we have $\mathbf{H}_0 \succeq \mathbf{K}_0$ and hence $\lambda_1(\mathbf{H}_0) \geq \lambda_1(\mathbf{K}_0)$, so the step size $1/\lambda_1(\mathbf{H}_0)$ used in this diagnostic is conservative relative to the theorem-level $\mathbf{K}_0$ step-size scale; tuning from $\mathbf{K}_0$ alone can give too large a step on the CK-like side. The momentum parameter is chosen from the total-kernel spectrum because this experiment tracks the actual linearized two-layer dynamics. The inputs and responses are again sampled uniformly from $\mathbb{S}^{d-1}$ and $\mathbb{S}^{n-1}$, respectively. We vary $q = c_1/c_2$ over $\{0.1, 0.3, 1, 3, 10\}$ while fixing $c_1 c_2 = 1$. Within each of the five seeds, all ratios use the same data and unscaled Gaussian draws. In Figure 3, we plot the final relative training loss $\mathcal{L}_{100}/\mathcal{L}_0$, the condition number L'/μ' of $\mathbf{H}_0$, and the relative output- and hidden-layer movements with respect to the varying initialization scaling ratio c_1/c_2. The mean ratio $\|\mathbf{K}_0\|_{\mathrm{F}} / \|\mathbf{H}_0\|_{\mathrm{F}}$ increases from approximately 0.004 at $q = 0.1$ to 0.99 at $q = 10$. At the smaller ratio, the hidden layer moves less relative to initialization than the output layer; at the larger ratio, the hidden layer moves more and the final relative loss is smaller. We use "CK-like" and "NTK-like" as descriptive labels for this ratio sweep, not as theorem-level regime definitions. The comparison concerns only the coupled initialization family; it does not establish robustness to perturbations or to standard initializations.

At the NTK-like endpoint $q = 10$, we also compare tuning NAG from $\mathbf{H}_0$ with tuning it from $\mathbf{K}_0$. In a separate set of five trials with the same data and initialization for both rules, the mean ratios were $\lambda_1(\mathbf{H}_0)/\lambda_1(\mathbf{K}_0) = 1.011$ and $\|\mathbf{K}_0\|_{\mathrm{F}} / \|\mathbf{H}_0\|_{\mathrm{F}} = 0.988$. The two rules reached relative loss 10^{-6} in 40.2 and 39.8 iterations on average, respectively; both gave final relative loss below 10^{-14} after 100 steps.

Computation. The experiments in Figures 1 and 2 used an Intel Core i7-13700F CPU and an NVIDIA GeForce RTX 4060 Ti GPU. Each GPU experiment script took approximately 10–30 minutes; peak GPU memory was not measured. A dense 5000×5000 kernel matrix occupies about 0.1 GB in float32 or 0.2 GB in float64. For the ratio sweep and kernel diagnostics in Figures 3 and 4, the CPU implementation uses float64 arithmetic and five seeds $0, \ldots, 4$. The diagnostic script, including its paired tuning runs and implementation checks, took 25.33 seconds on an AMD EPYC 9V74 with two BLAS threads and used 140.91 MiB of peak process memory.

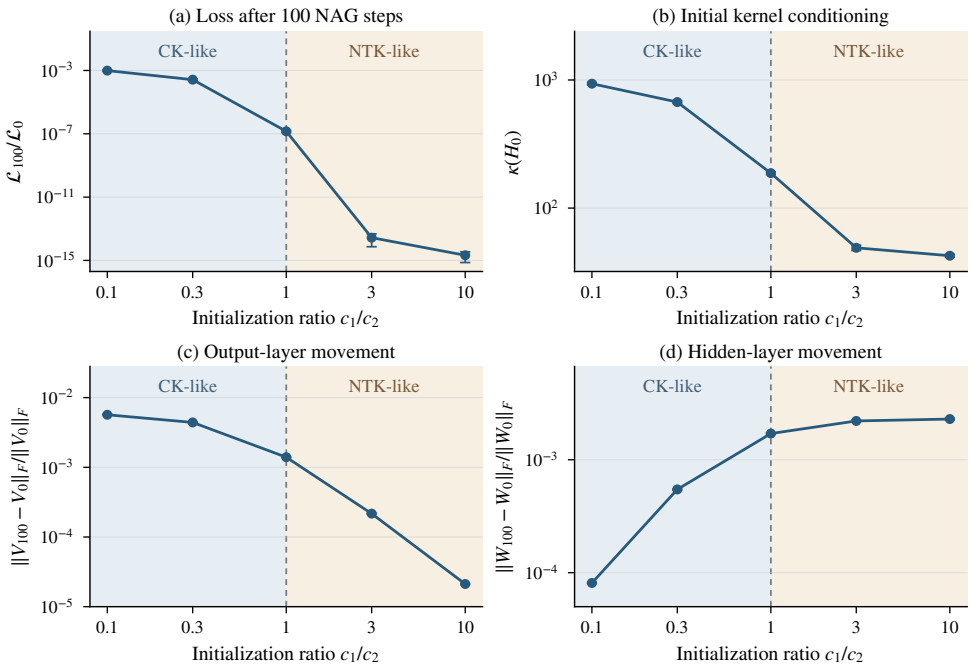


Figure 3: CK/NTK regime diagnostics as the initialization ratio c_1/c_2 varies with fixed overall scale $c_1 c_2 = 1$. Points show the mean over 5 seeds; error bars show ± 1 sample standard deviation. The top panels show final relative loss and the condition number of the empirical initial total kernel $\mathbf{H}_0 = \mathbf{K}_0 + \mathbf{G}_0$ used to set (η, β) in this experiment, where $\mathbf{K}_0$ is the empirical hidden-layer NTK component and $\mathbf{G}_0$ is the empirical output-layer/CK component defined in Section 2.2. The bottom panels show output-layer movement $\|\mathbf{V}_T - \mathbf{V}_0\|_{\mathrm{F}} / \|\mathbf{V}_0\|_{\mathrm{F}}$ and hidden-layer movement $\|\mathbf{W}_T - \mathbf{W}_0\|_{\mathrm{F}} / \|\mathbf{W}_0\|_{\mathrm{F}}$. The dashed line marks equal scales ($c_1/c_2 = 1$). Shading distinguishes the CK-like side ($c_1/c_2 < 1$) from the NTK-like side ($c_1/c_2 > 1$).

C Probability Tools

C.1 Concentration Inequalities

This section provides some basic concentration inequalities for random variables.

Proposition C.1 (Lemma 1 in Laurent and Massart [2000]). *Let $W \sim \chi^2(r)$ be the squared Euclidean norm of a standard Gaussian random vector in $\mathbb{R}^r$. Then for any $s > 0$,*

$$\mathbb{P}\left(W \geq r + 2\sqrt{rs} + 2s\right) \leq e^{-s},$$
$$\mathbb{P}\left(W \leq r - 2\sqrt{rs}\right) \leq e^{-s}.$$

Lemma C.2 (Bernstein's inequality, Theorem 2.8.1 in Vershynin [2018]). *Let $X_1, \ldots, X_m$ be independent zero-mean sub-exponential random variables. Then for every $t \geq 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^m X_i\right| \geq t\right) \leq 2\exp\left(-c\min\left\{\frac{t^2}{\sum_{i=1}^m \|X_i\|_{\psi_1}^2}, \frac{t}{\max_{1\leq i\leq m}\|X_i\|_{\psi_1}}\right\}\right),$$

where $c > 0$ is an absolute constant.

Lemma C.3 (Weighted Chernoff, Theorem 1 in Raghavan [1988]). *Suppose $a_i \geq 0$ for all i and $\max_{1\leq i\leq m} a_i > 0$, $X_1, X_2, \ldots, X_m$ are independent Bernoulli random variables with $\mathbb{E}[X_i] = p_i$, then for $\delta > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^m a_i X_i \geq (1+\delta)\sum_{i=1}^m a_i p_i\right) \leq \left[\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right]^{(\sum_{i=1}^m a_i p_i)/(\max_{1\leq i\leq m} a_i)}.$$

C.2 Singular Value of Random Matrix

This section provides a tail bound for the smallest singular value of sub-Gaussian random matrices.

Proposition C.4 (Rudelson and Vershynin [2009]). *Let $\mathbf{A}$ be an $N \times n$ random matrix, $N \geq n$, whose elements are i.i.d. zero mean sub-Gaussian random variables with unit variance. Then for $\tau \geq 0$, we have*

$$\mathbb{P}\big(\sigma_n(\mathbf{A}) \leq \tau(\sqrt{N} - \sqrt{n-1})\big) \leq (c_1\tau)^{N-n+1} + e^{-c_2 N}$$

where $c_1, c_2 > 0$ depend (polynomially) only on the sub-Gaussian moment.

C.3 Matrix Concentration

This section provides concentration inequalities for eigenvalues of the sum of random matrices.

Proposition C.5 (Matrix Chernoff, Theorem 4.1 in Gittens and Tropp [2011]). *Suppose $\{\mathbf{Z}_i\}_{i=1}^m$ is a set of independent random positive-semidefinite matrices with dimension n, $\|\mathbf{Z}_i\| \leq B$. For $k \in [n]$, let $\mu_k = \lambda_k(\sum_{i=1}^m \mathbb{E}\mathbf{Z}_i)$, then*

$$\mathbb{P}\left(\lambda_k\left(\sum_{i=1}^m \mathbf{Z}_i\right) \geq (1+\delta)\mu_k\right) \leq (n-k+1)\left[\frac{e^\delta}{(1+\delta)^{1+\delta}}\right]^{\mu_k/B}, \ \forall \delta > 0,$$
$$\mathbb{P}\left(\lambda_k\left(\sum_{i=1}^m \mathbf{Z}_i\right) \leq (1-\delta)\mu_k\right) \leq k\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_k/B}, \ \forall \delta \in [0,1).$$

D Initialization and Kernel Spectrum Proofs

D.1 General-Response Initialization

First, we have $\|\mathbf{r}_0\| \equiv \|\mathbf{Y}\|_{\mathrm{F}}$ by coupled initialization. Define $R_V \coloneqq \max_{i\in[2m]} \|[\mathbf{V}_0]_{:,i}\|$, $R_W \coloneqq \max_{i\in[2m]} \|[\mathbf{W}_0]_{:,i}\|$. We have the following bounds with proof in Appendix D.2.

Lemma D.1. *Suppose $\mathbf{V}_0$ and $\mathbf{W}_0$ are as* (4)*. For $\delta_0 \in (0,1)$, set $\ell_0 = \log(4m/\delta_0)$. Then with probability at least $1 - \delta_0/2 - 2e^{-\frac{9m(n\wedge d)}{64}}$,*

$$\tfrac{1}{2}c_1\sqrt{n} \leq R_V \leq c_1(\sqrt{n} + \sqrt{2\ell_0}), \qquad \tfrac{1}{2}c_2\sqrt{d} \leq R_W \leq c_2(\sqrt{d} + \sqrt{2\ell_0}).$$

By Lemma D.1, R_V and R_W are $\widetilde{O}(c_1\sqrt{n})$ and $\widetilde{O}(c_2\sqrt{d})$, respectively.

The convergence rate depends on the condition number of the initial empirical hidden-layer NTK component, $\mathbf{K}_0$. We show that a width of $\widetilde{\Omega}(Nn/\lambda)$ suffices to bound its extreme eigenvalues under our initialization.

Lemma D.2 (Eigenvalues of $\mathbf{K}_0$)**.** *Suppose $\mathbf{V}_0$ and $\mathbf{W}_0$ are as* (4)*, $\delta \in (0,1)$, if the width satisfies*

$$m \geq \widetilde{\Omega}\left(\frac{\|\mathbf{X}\|^2 n}{\lambda}\log\frac{Nn}{\delta}\right),$$

then $\lambda_{Nn}(\mathbf{K}_0) \geq \frac{1}{2}c_1^2 m\lambda$ and $\lambda_1(\mathbf{K}_0) \leq 9c_1^2 mn\lambda_1(\overline{\mathbf{K}})$ with probability at least $1-\delta$.

The proof is provided in Appendix D.3. The key point is that matrix Chernoff can exploit the Kronecker structure and positive-semidefiniteness of the summands, avoiding the entrywise deviation route that leads to quadratic width in earlier ReLU NTK analyses [Du et al., 2018, Song and Yang, 2019]. The lemma gives $\kappa(\mathbf{K}_0) = O(n\kappa)$ with high probability. The factor n comes from the largest eigenvalue bound, through the relaxation $\mathbf{A}_i \preceq \|[\mathbf{V}_0]_{:,i}\|^2 \mathbf{I}_n$; it is not a lower bound on shared-feature conditioning. When $\|\mathbf{X}\| \ll \|\mathbf{X}\|_{\mathrm{F}}$, the initialization concentration can be better than $\widetilde{\Omega}(Nn/\lambda)$. For independent uniformly distributed spherical inputs, $\|\mathbf{X}\|^2 = \widetilde{O}(1+N/d)$ [Vershynin, 2010], giving $m = \widetilde{\Omega}(Nn/(d\lambda))$ at initialization when $N \gtrsim d$. The trajectory bounds still depend on N.

We now give the proofs for the unsketched Gaussian initialization (4) and the sketched initialization (8).

D.2 Proof of Lemma D.1

Proof of Lemma D.1. Applying Proposition C.1 with $s = \ell_0 = \log(4m/\delta_0)$ for the upper part and taking the union bound, we get

$$\mathbb{P}\left(\exists i \in [m],\ \|[\mathbf{\Phi}]_{:,i}\| > \sqrt{n} + \sqrt{2\ell_0}\right) \leq \delta_0/4,$$
$$\mathbb{P}\left(\exists i \in [m],\ \|[\mathbf{\Psi}]_{:,i}\| > \sqrt{d} + \sqrt{2\ell_0}\right) \leq \delta_0/4,$$

On the other hand, since the columns of $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are independent, plugging $s = \frac{9n}{64}$ in Proposition C.1, we get

$$\mathbb{P}\left(\forall i \in [m],\ \|[\mathbf{\Phi}]_{:,i}\| \leq \frac{\sqrt{n}}{2}\right) \leq e^{-\frac{9nm}{64}}.$$

Similarly, we have

$$\mathbb{P}\left(\forall i \in [m],\ \|[\mathbf{\Psi}]_{:,i}\| \leq \frac{\sqrt{d}}{2}\right) \leq e^{-\frac{9dm}{64}}.$$

Combining the upper and lower bounds with coupled initialization and definition of R_V and R_W, we get the final results. □

D.3 Proof of Lemma D.2

Proof of Lemma D.2. In this proof, we omit the iteration index and denote $\mathbf{A}_i \coloneqq \mathbf{A}_{i,0}$, $\mathbf{B}_i \coloneqq \mathbf{B}_{i,0}$. Given coupled initialization (4), we have $\mathbf{K}_0 = 2\left(\sum_{i=1}^m \mathbf{B}_i^\top \otimes \mathbf{A}_i\right)$, hence it suffices to consider the sum of the first m neurons and multiply the final results by a factor of 2. For any $i \in [m]$, the expectation of $\mathbf{B}_i^\top \otimes \mathbf{A}_i$ is given by

$$\begin{aligned}\mathbb{E}[\mathbf{B}_i^\top \otimes \mathbf{A}_i] &= \mathbb{E}[\mathbf{B}_i] \otimes \mathbb{E}[\mathbf{A}_i] \\ &= \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\mathbf{I}_d)}\left[\mathrm{diag}(\mathbb{1}[\mathbf{w}^\top\mathbf{X} \geq 0])\mathbf{X}^\top\mathbf{X}\mathrm{diag}(\mathbb{1}[\mathbf{w}^\top\mathbf{X} \geq 0])\right] \otimes \left(c_1^2\mathbf{I}_n\right) \\ &= c_1^2\overline{\mathbf{K}} \otimes \mathbf{I}_n.\end{aligned}$$

Let $\overline{L} = c_1^2\lambda_1(\overline{\mathbf{K}})$, $\bar{\mu} = c_1^2\lambda$, then Assumption 2.3 implies

$$\lambda_1\left(\mathbb{E}[\mathbf{B}_i^\top \otimes \mathbf{A}_i]\right) = \overline{L} \geq \lambda_{Nn}\left(\mathbb{E}[\mathbf{B}_i^\top \otimes \mathbf{A}_i]\right) = \bar{\mu} > 0.$$

Step 1: lower-tail bound. We first derive a lower bound for $\lambda_{Nn}(\mathbf{K}_0)$. In both spectrum proofs, truncation acts on each summand separately and therefore preserves independence. Let $\mathbf{Z}_i = (\mathbf{B}_i^\top \otimes \mathbf{A}_i) \cdot \mathbb{1}[\|[\mathbf{V}_0]_{:,i}\| \leq M_V]$ where $M_V > 0$ is a threshold to be specified, we have

$$\begin{aligned}\mathbb{E}[\mathbf{Z}_i] &= \mathbb{E}\left[(\mathbf{B}_i^\top \otimes \mathbf{A}_i) \cdot \mathbb{1}[\|[\mathbf{V}_0]_{:,i}\| \leq M_V]\right] \\ &= c_1^2\overline{\mathbf{K}} \otimes \mathbf{I}_n - \mathbb{E}\left[(\mathbf{B}_i^\top \otimes \mathbf{A}_i) \cdot \mathbb{1}[\|[\mathbf{V}_0]_{:,i}\| > M_V]\right],\end{aligned}$$

and hence

$$\begin{aligned}\mu_{Nn} &\coloneqq \lambda_{Nn}\left(\sum_{i=1}^m \mathbb{E}[\mathbf{Z}_i]\right) \\ &= \lambda_{Nn}\left(m \cdot \mathbb{E}[\mathbf{B}_i^\top \otimes \mathbf{A}_i] - \sum_{i=1}^m \mathbb{E}\left[(\mathbf{B}_i^\top \otimes \mathbf{A}_i) \cdot \mathbb{1}[\|[\mathbf{V}_0]_{:,i}\| > M_V]\right]\right) \\ &\geq m\bar{\mu} - \sum_{i=1}^m \lambda_1\left(\mathbb{E}[\mathbf{B}_i] \otimes \mathbb{E}\left[\mathbf{A}_i \cdot \mathbb{1}[\|[\mathbf{V}_0]_{:,i}\| > M_V]\right]\right) \\ &= m\bar{\mu} - \sum_{i=1}^m \lambda_1\left(\overline{\mathbf{K}}\right)\lambda_1\left(\mathbb{E}\left[\mathbf{A}_i \cdot \mathbb{1}[\|[\mathbf{V}_0]_{:,i}\| > M_V]\right]\right),\end{aligned}$$

where the third line is from Weyl's inequality and the fourth line is the property of the eigenvalue of the Kronecker product. For any $i \in [m]$, let $\boldsymbol{\phi}_i \coloneqq [\boldsymbol{\Phi}]_{:,i}$, we have $\boldsymbol{\phi}_i \sim \mathcal{N}(0, \mathbf{I}_n)$ and

$$\begin{aligned}\lambda_1\left(\mathbb{E}\left[\mathbf{A}_i \cdot \mathbb{1}[\|[\mathbf{V}_0]_{:,i}\| > M_V]\right]\right) &\leq \mathbb{E}\left[\lambda_1\left(\mathbf{A}_i \cdot \mathbb{1}[\|[\mathbf{V}_0]_{:,i}\| > M_V]\right)\right] \\ &\leq \mathbb{E}\left[\lambda_1\left(c_1^2\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top \cdot \mathbb{1}\left[c_1\|\boldsymbol{\phi}_i\| > M_V\right]\right)\right] \\ &\leq \mathbb{E}_{\boldsymbol{\phi}\sim\mathcal{N}(0,\mathbf{I}_n)}\left[c_1^2\|\boldsymbol{\phi}\|^2 \cdot \mathbb{1}\left[c_1\|\boldsymbol{\phi}\| > M_V\right]\right] \\ &= c_1^2\mathbb{E}_{W\sim\chi^2(n)}\left[W \cdot \mathbb{1}\left[c_1\sqrt{W} > M_V\right]\right] \\ &\leq c_1^2\sqrt{\mathbb{E}_{W\sim\chi^2(n)}[W^2]}\sqrt{\mathbb{P}\left(W > \left(\frac{M_V}{c_1}\right)^2\right)} \\ &\leq c_1^2\sqrt{n(n+2)}e^{-\underline{s}/2}.\end{aligned}$$

Here, the first line is by the convexity of the top eigenvalue and Jensen's inequality. We then relax the indicator and apply Cauchy–Schwarz to the truncated $\chi^2(n)$ variable. The last line uses $\mathbb{E}[W^2] = n(n+2)$ for $W \sim \chi^2(n)$ and the tail bound in Proposition C.1 with $M_V = c_1\sqrt{n + 2\sqrt{n\underline{s}} + 2\underline{s}}$ for some $\underline{s} > 0$.

By choosing $\underline{s} = \max\{n, 2\log(6\sqrt{n(n+2)}\kappa(\overline{\mathbf{K}}))\}$, we have

$$\begin{aligned}\mu_{Nn} &\geq m\bar{\mu} - m\overline{L}\sqrt{n(n+2)}e^{-\underline{s}/2} \\ &\geq mc_1^2\lambda - mc_1^2\lambda_1(\overline{\mathbf{K}})(6\kappa(\overline{\mathbf{K}}))^{-1} \\ &\geq \frac{1}{2}m\bar{\mu},\end{aligned}$$

and correspondingly

$$\begin{aligned}M_V &= c_1 \cdot \max\left\{\sqrt{5n}, \sqrt{n + 2\sqrt{2n\log(6\sqrt{n(n+2)}\kappa(\overline{\mathbf{K}}))} + 4\log(6\sqrt{n(n+2)}\kappa(\overline{\mathbf{K}}))}\right\} \\ &\leq c_1\sqrt{5n + 6\log(6\sqrt{n(n+2)}\kappa(\overline{\mathbf{K}}))}.\end{aligned}$$

Moreover, for every $i \in [m]$ we have

$$\begin{aligned}\|\mathbf{Z}_i\| &= \|\mathbf{B}_i\| \, \|\mathbf{A}_i \cdot \mathbb{1}[\|[\mathbf{V}_0]_{:,i}\| \le M_V]\| \\ &\le \left\|\mathrm{diag}(\mathbb{1}[\mathbf{W}_0^\top \mathbf{X} \ge 0]_{i,:})\mathbf{X}^\top \mathbf{X} \mathrm{diag}(\mathbb{1}[\mathbf{W}_0^\top \mathbf{X} \ge 0]_{i,:})\right\| M_V^2 \\ &\le \|\mathbf{X}\|^2 M_V^2.\end{aligned}$$

By plugging $\mathbf{Z}_i$ into the matrix Chernoff bound (Proposition C.5) and using relative deviation $1/2$ for the lower eigenvalue bound, we get

$$\begin{aligned}&\mathbb{P}\left(\lambda_{Nn}\left(\sum_{i=1}^m \mathbf{B}_i^\top \otimes \mathbf{A}_i\right) \le \frac{1}{4} m\bar{\mu}\right) \\ \le &\mathbb{P}\left(\lambda_{Nn}\left(\sum_{i=1}^m (\mathbf{B}_i^\top \otimes \mathbf{A}_i) \cdot \mathbb{1}[\|[\mathbf{V}_0]_{:,i}\| \le M_V]\right) \le \frac{1}{2}\mu_{Nn}\right) \\ \le &\exp\left(-\frac{\mu_{Nn}}{10\|\mathbf{X}\|^2 M_V^2} + \log(Nn)\right) \\ \le &\exp\left(-\frac{m\lambda}{40\|\mathbf{X}\|^2 (5n + 6\log(6\sqrt{n(n+2)}\kappa(\overline{\mathbf{K}})))} + \log(Nn)\right),\end{aligned}$$

which gives the lower bound for $\lambda_{Nn}(\mathbf{K}_0)$ after restoring the coupling factor 2. The first inequality is because $\mathbf{B}_i$ and $\mathbf{A}_i$ are positive-semidefinite, thus removing summands will only make the eigenvalue smaller. Choosing $m \ge \frac{40\|\mathbf{X}\|^2(5n+6\log(6\sqrt{n(n+2)}\kappa))}{\lambda} \log\frac{4Nn}{\delta}$ makes this lower-tail failure probability at most $\delta/4$.

Step 2: upper-tail bound. We now derive an upper bound for $\lambda_1(\mathbf{K}_0)$. Conditioned on $\mathbf{\Phi}$, $\mathbf{A}_i \preceq \|[\mathbf{V}_0]_{:,i}\|^2 \mathbf{I}_n$ for all $i \in [m]$. Denote $a_i = \|[\mathbf{V}_0]_{:,i}\|^2$, then by Kronecker product and positive semidefinite summands, $\mathbf{B}_i^\top \otimes \mathbf{A}_i \preceq a_i \cdot \mathbf{B}_i \otimes \mathbf{I}_n \preceq a_i \|\mathbf{X}\|^2 \cdot \mathbf{I}_{Nn}$ and

$$\lambda_1\left(\sum_{i=1}^m \mathbf{B}_i^\top \otimes \mathbf{A}_i\right) \le \lambda_1\left(\sum_{i=1}^m a_i \cdot \mathbf{B}_i^\top \otimes \mathbf{I}_n\right) \le \lambda_1\left(\sum_{i=1}^m a_i \cdot \mathbf{B}_i\right).$$

Let $\mathbf{Z}_i = a_i \mathbf{B}_i$, then $\lambda_1\left(\sum_{i=1}^m \mathbb{E}[\mathbf{Z}_i \mid \mathbf{\Phi}]\right) = \left(\sum_{i=1}^m a_i\right)\lambda_1(\overline{\mathbf{K}})$. Given $\mathbf{\Phi}$, the hidden columns remain independent. Applying Proposition C.5 to the $N \times N$ matrices $\mathbf{Z}_i$ gives

$$\mathbb{P}\left(\lambda_1\left(\sum_{i=1}^m \mathbf{Z}_i\right) \ge \frac{3}{2}\left(\sum_{i=1}^m a_i\right)\lambda_1(\overline{\mathbf{K}}) \,\middle|\, \mathbf{\Phi}\right) \le \exp\left(-\frac{\left(\sum_{i=1}^m a_i\right)\lambda_1(\overline{\mathbf{K}})}{10\|\mathbf{X}\|^2 \max_{i\in[m]} a_i} + \log N\right).$$

By definition, $\sum_{i=1}^m a_i = c_1^2 \|\mathbf{\Phi}\|_{\mathrm{F}}^2 = c_1^2 W$ where $W \sim \chi^2(mn)$, $\max_{i\in[m]} a_i = R_V^2$. Using the lower tail of Proposition C.1 with $s = \frac{mn}{16}$ and the upper tail with $s = \frac{mn}{4}$, we have $\frac{mn}{2} \le W \le \frac{5mn}{2} < 3mn$ with probability at least $1 - e^{-\frac{mn}{16}} - e^{-\frac{mn}{4}}$. Put $\ell_\delta = \log(32m/\delta)$. The output-only upper-radius bound in Appendix D.2 with $\delta_0 = \delta/8$ gives $\max_i a_i \le 2c_1^2(n + 2\ell_\delta)$ except with probability $\delta/32$. These events depend only on $\mathbf{\Phi}$. Integrating the conditional bound and restoring the coupling factor gives

$$\begin{aligned}\mathbb{P}\left(\lambda_1(\mathbf{K}_0) > 9c_1^2 mn\lambda_1(\overline{\mathbf{K}})\right) \le &\frac{\delta}{32} + e^{-mn/16} + e^{-mn/4} \\ &+ \exp\left(-\frac{mn\lambda_1(\overline{\mathbf{K}})}{40\|\mathbf{X}\|^2 (n + 2\ell_\delta)} + \log N\right).\end{aligned}$$

The width condition $m = \widetilde{\Omega}((\|\mathbf{X}\|^2 n/\lambda)\log(Nn/\delta))$ makes the final term at most $\delta/4$ and each chi-square tail at most $\delta/16$, after enlarging absolute constants and polylogarithmic factors. Indeed, $\|\mathbf{X}\|^2 \ge 1$, $\lambda \le 1/2$, $n \ge 1$, and $\lambda_1(\overline{\mathbf{K}}) \ge \lambda$; the remaining $\log m$ is absorbed by the log self-bounding convention in Appendix A. Combining with the lower-tail budget $\delta/4$ gives total failure at most $21\delta/32 < \delta$. No separate polynomial dependence on $1/\delta$ is needed. □

D.4 Proof of Lemma 4.3

Proof of Lemma 4.3. We only need to show the R_V part, as R_W is the same as in Lemma D.1. Suppose $\mathbf{Y}$ has singular value decomposition $\mathbf{Y} = \mathbf{U}_Y \mathbf{\Sigma}_Y \mathbf{V}_Y^\top$, then the columns of $\mathbf{V}_Y^\top \mathbf{\Phi} \in \mathbb{R}^{r \times m}$ are independent Gaussian vectors with distribution $\mathcal{N}(0, \mathbf{V}_Y^\top \mathbf{V}_Y) = \mathcal{N}(0, \mathbf{I}_r)$.

Applying Proposition C.1 with $s = \ell_0 = \log(4m/\delta_0)$ for the upper part and taking the union bound, we get

$$\mathbb{P}\left(\exists i \in [m],\ \left\|[\mathbf{V}_Y^\top \mathbf{\Phi}]_{:,i}\right\| > \sqrt{r} + \sqrt{2\ell_0}\right) \leq \delta_0/4.$$

On the other hand, since the columns of $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are independent, plugging $s = \frac{9r}{64}$ in Proposition C.1, we get

$$\mathbb{P}\left(\forall i \in [m],\ \left\|[\mathbf{V}_Y^\top \mathbf{\Phi}]_{:,i}\right\| \leq \frac{\sqrt{r}}{2}\right) \leq e^{-\frac{9rm}{64}}.$$

Combining the upper and lower bounds and the singular values of $\mathbf{Y}$ yields the results for R_V and R_W.

Applying Proposition C.4 to $(\mathbf{V}_Y^\top \mathbf{\Phi})^\top$,

$$\mathbb{P}\left(\sigma_r(\mathbf{V}_Y^\top \mathbf{\Phi}) \leq \tau\left(\sqrt{m} - \sqrt{r-1}\right)\right) \leq e^{-(m-r+1)\log\frac{1}{c\tau}} + e^{-c'm}$$

for some absolute constants c and c' and any $\tau \geq 0$, interpreting the bound at zero by continuity. Combining this estimate with the radius event, the spectral bound follows from the fact that

$$c_1\sqrt{2} \cdot \sigma_r(\mathbf{V}_Y^\top \mathbf{\Phi})\sigma_r(\mathbf{Y}) \leq \sigma_r(\mathbf{V}_0)$$

where the factor $\sqrt{2}$ is due to the coupled initialization. If the displayed lower singular-value bound is positive, $\mathbf{V}_0$ has rank r and $\mathrm{col}(\mathbf{V}_0) \subseteq \mathrm{col}(\mathbf{Y})$, so the spaces are equal. □

D.5 Proof of Lemma 4.5

Proof of Lemma 4.5. Similar to the proof of Lemma D.2, we omit the iteration index and denote $\mathbf{A}_i \coloneqq \mathbf{A}_{i,0}$, $\mathbf{B}_i \coloneqq \mathbf{B}_{i,0}$, and consider the sum of the first m neurons and multiply the final results by a factor of 2. For any $i \in [m]$, the expectation of $\mathbf{B}_i^\top \otimes \mathbf{A}_i$ is given by

$$\begin{aligned}
\mathbb{E}[\mathbf{B}_i^\top \otimes \mathbf{A}_i] &= \mathbb{E}[\mathbf{B}_i] \otimes \mathbb{E}[\mathbf{A}_i] \\
&= \mathbb{E}_{\mathbf{w}\sim\mathcal{N}(0,\mathbf{I})}\left[\mathrm{diag}(\mathbb{1}[\mathbf{w}^\top \mathbf{X} \geq 0])\mathbf{X}^\top \mathbf{X}\mathrm{diag}(\mathbb{1}[\mathbf{w}^\top \mathbf{X} \geq 0])\right] \otimes \left(c_1^2 \mathbf{Y}\mathbf{Y}^\top\right) \\
&= \overline{\mathbf{K}} \otimes \left(c_1^2 \mathbf{Y}\mathbf{Y}^\top\right),
\end{aligned}$$

where we use $\mathbb{E}[\mathbf{A}_i] = \mathbb{E}[[\mathbf{V}_0]_{:,i}[\mathbf{V}_0]_{:,i}^\top] = c_1^2 \mathbb{E}_{\boldsymbol{\phi}_i \sim \mathcal{N}(0,\mathbf{I}_N)}[\mathbf{Y}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top \mathbf{Y}^\top] = c_1^2 \mathbf{Y}\mathbf{I}_N\mathbf{Y}^\top = c_1^2 \mathbf{Y}\mathbf{Y}^\top$.

Let $\overline{L} = c_1^2 \lambda_1(\overline{\mathbf{K}})\lambda_1(\mathbf{Y}\mathbf{Y}^\top)$, $\bar{\mu} = c_1^2 \lambda\lambda_r(\mathbf{Y}\mathbf{Y}^\top)$, then Assumption 2.3 and $\mathrm{rank}(\mathbf{Y}) = r$ implies

$$\lambda_1\left(\mathbb{E}[\mathbf{B}_i^\top \otimes \mathbf{A}_i]\right) = \overline{L} \geq \lambda_{Nr}\left(\mathbb{E}[\mathbf{B}_i^\top \otimes \mathbf{A}_i]\right) = \bar{\mu} > 0,$$

and any remaining eigenvalues vanish.

Step 1: lower-tail bound on the positive spectrum. We first derive a lower bound for $\lambda_{Nr}(\mathbf{K}_0)$. Let $\mathbf{Z}_i = (\mathbf{B}_i^\top \otimes \mathbf{A}_i) \cdot \mathbb{1}[\|[\mathbf{V}_0]_{:,i}\| \leq M_V]$ where $M_V > 0$ is a threshold to be specified, we have

$$\begin{aligned}
\mathbb{E}[\mathbf{Z}_i] &= \mathbb{E}\left[(\mathbf{B}_i^\top \otimes \mathbf{A}_i) \cdot \mathbb{1}[\|[\mathbf{V}_0]_{:,i}\| \leq M_V]\right] \\
&= \overline{\mathbf{K}} \otimes \left(c_1^2 \mathbf{Y}\mathbf{Y}^\top\right) - \mathbb{E}\left[(\mathbf{B}_i^\top \otimes \mathbf{A}_i) \cdot \mathbb{1}[\|[\mathbf{V}_0]_{:,i}\| > M_V]\right],
\end{aligned}$$

and hence

$$\begin{aligned}
\mu_{Nr} &\coloneqq \lambda_{Nr}\left(\sum_{i=1}^{m}\mathbb{E}[\mathbf{Z}_i]\right)\\
&= \lambda_{Nr}\left(m\cdot\mathbb{E}[\mathbf{B}_i^\top\otimes\mathbf{A}_i]-\sum_{i=1}^{m}\mathbb{E}\left[(\mathbf{B}_i^\top\otimes\mathbf{A}_i)\cdot\mathbb{1}[\|[\mathbf{V}_0]_{:,i}\|>M_V]\right]\right)\\
&\geq m\bar{\mu}-\sum_{i=1}^{m}\lambda_1\left(\mathbb{E}[\mathbf{B}_i]\otimes\mathbb{E}\left[\mathbf{A}_i\cdot\mathbb{1}[\|[\mathbf{V}_0]_{:,i}\|>M_V]\right]\right)\\
&= m\bar{\mu}-\sum_{i=1}^{m}\lambda_1\left(\overline{\mathbf{K}}\right)\lambda_1\left(\mathbb{E}\left[\mathbf{A}_i\cdot\mathbb{1}[\|[\mathbf{V}_0]_{:,i}\|>M_V]\right]\right),
\end{aligned}$$

where the third line is from Weyl's inequality and the fourth line is the property of the eigenvalue of the Kronecker product. Suppose $\mathbf{Y}$ has SVD $\mathbf{Y}=\mathbf{U}_Y\mathbf{\Sigma}_Y\mathbf{V}_Y^\top$. For any $i\in[m]$, we have $\mathbf{V}_Y^\top\boldsymbol{\phi}_i\sim\mathcal{N}(0,\mathbf{I}_r)$, thus

$$\begin{aligned}
\lambda_1\left(\mathbb{E}\left[\mathbf{A}_i\cdot\mathbb{1}[\|[\mathbf{V}_0]_{:,i}\|>M_V]\right]\right) &\leq \mathbb{E}\left[\lambda_1\left(\mathbf{A}_i\cdot\mathbb{1}[\|[\mathbf{V}_0]_{:,i}\|>M_V]\right)\right]\\
&\leq \mathbb{E}\left[c_1^2\|\mathbf{Y}\|^2\left\|\mathbf{V}_Y^\top\boldsymbol{\phi}_i\right\|^2\cdot\mathbb{1}\left[c_1\|\mathbf{Y}\|\left\|\mathbf{V}_Y^\top\boldsymbol{\phi}_i\right\|>M_V\right]\right]\\
&= \mathbb{E}_{\boldsymbol{\psi}\sim\mathcal{N}(0,\mathbf{I}_r)}\left[c_1^2\|\mathbf{Y}\|^2\|\boldsymbol{\psi}\|^2\cdot\mathbb{1}\left[c_1\|\mathbf{Y}\|\|\boldsymbol{\psi}\|>M_V\right]\right]\\
&= c_1^2\|\mathbf{Y}\|^2\,\mathbb{E}_{W\sim\chi^2(r)}\left[W\cdot\mathbb{1}\left[c_1\|\mathbf{Y}\|\sqrt{W}>M_V\right]\right]\\
&\leq c_1^2\|\mathbf{Y}\|^2\sqrt{\mathbb{E}_{W\sim\chi^2(r)}[W^2]}\sqrt{\mathbb{P}\left(W>\left(\frac{M_V}{c_1\|\mathbf{Y}\|}\right)^2\right)}\\
&\leq c_1^2\|\mathbf{Y}\|^2\sqrt{r(r+2)}e^{-\underline{s}/2}.
\end{aligned}$$

Here, the first line is by the convexity of the top eigenvalue and Jensen's inequality. We use $\lambda_1(\mathbf{U}_Y\mathbf{\Sigma}_Y\boldsymbol{z}\boldsymbol{z}^\top\mathbf{\Sigma}_Y\mathbf{U}_Y^\top)\leq\|\mathbf{Y}\|^2\|\boldsymbol{z}\|^2$ with $\boldsymbol{z}=\mathbf{V}_Y^\top\boldsymbol{\phi}_i$, relax the indicator, and apply Cauchy–Schwarz to the truncated $\chi^2(r)$ variable. The last line uses $\mathbb{E}[W^2]=r(r+2)$ for $W\sim\chi^2(r)$ and the tail bound in Proposition C.1 with $M_V=c_1\|\mathbf{Y}\|\sqrt{r+2\sqrt{r\underline{s}}+2\underline{s}}$ for some $\underline{s}>0$.

Let $\Gamma_r\coloneqq 6\sqrt{r(r+2)}\kappa(\overline{\mathbf{K}})\kappa(\mathbf{Y}\mathbf{Y}^\top)$. By choosing $\underline{s}=\max\{r,2\log\Gamma_r\}$, we have

$$\begin{aligned}
\mu_{Nr} &\geq m\bar{\mu}-m\bar{L}\sqrt{r(r+2)}e^{-\underline{s}/2}\\
&\geq mc_1^2\lambda\lambda_r(\mathbf{Y}\mathbf{Y}^\top)-mc_1^2\lambda_1(\overline{\mathbf{K}})\lambda_1(\mathbf{Y}\mathbf{Y}^\top)(6\kappa(\overline{\mathbf{K}})\kappa(\mathbf{Y}\mathbf{Y}^\top))^{-1}\\
&\geq \frac{1}{2}m\bar{\mu},
\end{aligned}$$

and correspondingly

$$\begin{aligned}
M_V &= c_1\|\mathbf{Y}\|\cdot\max\left\{\sqrt{5r},\sqrt{r+2\sqrt{2r\log\Gamma_r}+4\log\Gamma_r}\right\}\\
&\leq c_1\|\mathbf{Y}\|\sqrt{5r+6\log\Gamma_r}.
\end{aligned}$$

Moreover, for every $i\in[m]$ we have

$$\|\mathbf{Z}_i\|=\|\mathbf{B}_i\|\,\|\mathbf{A}_i\cdot\mathbb{1}[\|[\mathbf{V}_0]_{:,i}\|\leq M_V]\|\leq\|\mathbf{X}\|^2M_V^2.$$

By plugging $\mathbf{Z}_i$ into the matrix Chernoff bound (Proposition C.5) and using relative deviation $1/2$ for the lower eigenvalue bound, we get

$$\begin{aligned}
&\mathbb{P}\left(\lambda_{Nr}\left(\sum_{i=1}^{m}\mathbf{B}_i^\top\otimes\mathbf{A}_i\right)\le\frac{1}{4}m\bar{\mu}\right)\\
\le&\mathbb{P}\left(\lambda_{Nr}\left(\sum_{i=1}^{m}(\mathbf{B}_i^\top\otimes\mathbf{A}_i)\cdot\mathbb{1}[\|[\mathbf{V}_0]_{:,i}\|\le M_V]\right)\le\frac{1}{2}\mu_{Nr}\right)\\
\le&\exp\left(-\frac{\mu_{Nr}}{10\left\|\mathbf{X}\right\|^2M_V^2}+\log(Nr)\right)\\
\le&\exp\left(-\frac{m\lambda\lambda_r(\mathbf{Y}\mathbf{Y}^\top)}{40\left\|\mathbf{X}\right\|^2\left\|\mathbf{Y}\right\|^2(5r+6\log\Gamma_r)}+\log(Nr)\right)\\
\le&\exp\left(-\frac{m\lambda}{40\left\|\mathbf{X}\right\|^2\kappa(\mathbf{Y}\mathbf{Y}^\top)(5r+6\log\Gamma_r)}+\log(Nr)\right),
\end{aligned}$$

which gives the lower bound for $\lambda_{Nr}(\mathbf{K}_0)$ after restoring the coupling factor 2. The first inequality is because $\mathbf{B}_i$ and $\mathbf{A}_i$ are positive-semidefinite, thus removing summands will only make the eigenvalue smaller. Choosing $m\ge\frac{40\|\mathbf{X}\|^2\kappa^2(\mathbf{Y})(5r+6\log\Gamma_r)}{\lambda}\log\frac{4Nr}{\delta}$ makes this lower-tail failure probability at most $\delta/4$.

Step 2: upper-tail bound. We now derive an upper bound for $\lambda_1(\mathbf{K}_0)$. Similar to the proof of Lemma D.2 in Appendix D.3, we denote $a_i=\|[\mathbf{V}_0]_{:,i}\|^2$. Let $\mathbf{Z}_i=a_i\mathbf{B}_i$. Conditioned on $\mathbf{\Phi}$, applying Proposition C.5 to these $N\times N$ matrices gives

$$\mathbb{P}\left(\lambda_1\left(\sum_{i=1}^{m}\mathbf{Z}_i\right)\ge\frac{3}{2}\left(\sum_{i=1}^{m}a_i\right)\lambda_1(\overline{\mathbf{K}})\,\middle|\,\mathbf{\Phi}\right)\le\exp\left(-\frac{(\sum_{i=1}^{m}a_i)\,\lambda_1(\overline{\mathbf{K}})}{10\left\|\mathbf{X}\right\|^2\max_{i\in[m]}a_i}+\log N\right).$$

By definition, $\sum_{i=1}^{m}a_i=c_1^2\left\|\mathbf{U}_Y\mathbf{\Sigma}_Y\mathbf{V}_Y^\top\mathbf{\Phi}\right\|_{\mathrm{F}}^2$. Since the entries of $\mathbf{V}_Y^\top\mathbf{\Phi}$ are independent standard Gaussians, $W=\left\|\mathbf{V}_Y^\top\mathbf{\Phi}\right\|_{\mathrm{F}}^2\sim\chi^2(mr)$ and

$$c_1^2\sigma_r^2(\mathbf{Y})W\le\sum_{i=1}^{m}a_i\le c_1^2\left\|\mathbf{Y}\right\|^2W.$$

Using the lower tail of Proposition C.1 with $s=\frac{mr}{16}$ and the upper tail with $s=\frac{mr}{4}$, we have $\frac{mr}{2}\le W\le\frac{5mr}{2}<3mr$ with probability at least $1-e^{-\frac{mr}{16}}-e^{-\frac{mr}{4}}$. Put $\ell_\delta=\log(32m/\delta)$. The output-only upper-radius bound in Appendix D.4 with $\delta_0=\delta/8$ gives $\max_i a_i\le2c_1^2\left\|\mathbf{Y}\right\|^2(r+2\ell_\delta)$ except with probability $\delta/32$. These events depend only on $\mathbf{\Phi}$. Integrating the conditional bound and restoring the coupling factor gives

$$\begin{aligned}
\mathbb{P}\left(\lambda_1(\mathbf{K}_0)>9c_1^2mr\lambda_1(\overline{\mathbf{K}})\left\|\mathbf{Y}\right\|^2\right)\le&\frac{\delta}{32}+e^{-mr/16}+e^{-mr/4}\\
&+\exp\left(-\frac{mr\lambda_1(\overline{\mathbf{K}})}{40\left\|\mathbf{X}\right\|^2(r+2\ell_\delta)\kappa^2(\mathbf{Y})}+\log N\right).
\end{aligned}$$

The width condition $m=\widetilde{\Omega}((\|\mathbf{X}\|^2r\kappa^2(\mathbf{Y})/\lambda)\log(Nr/\delta))$ makes the final term at most $\delta/4$ and each chi-square tail at most $\delta/16$. The justification is the same log self-bounding argument as in Appendix D.3, with $r\ge1$ and $\kappa(\mathbf{Y})\ge1$. Together with the lower-tail budget $\delta/4$, the total failure probability is at most $21\delta/32<\delta$. □

E Residual Dynamics and Linear Contraction Proofs

This section proves the residual dynamics decomposition in Proposition E.1 and the contraction lemmas for the associated reference systems. We define several additional notations. Write $\mathbf{R}_t\coloneqq$

$\mathbf{R}(\mathbf{V}_t, \mathbf{W}_t)$, $\mathbf{r}_t \coloneqq \mathrm{vec}(\mathbf{R}_t)$, $\mathbf{P}_t \coloneqq \mathbf{V}_{t+1} - \mathbf{V}_t$, and $\mathbf{Q}_t \coloneqq \mathbf{W}_{t+1} - \mathbf{W}_t$. Let $\mathbf{H}_t = \mathbf{G}_t + \mathbf{K}_t$ where $\mathbf{G}_t$ and $\mathbf{K}_t$ are defined in Section 2.2. Let $\mathbf{J}_t = \mathbb{1}[\mathbf{W}_t^\top \mathbf{X} \geq 0]$ be the activation pattern at the t-th iteration. Fix a mask $\mathbf{M}$ whose entries equal to one retain their initial activation pattern on the time interval under consideration, and put $\mathbf{M}^\perp = \mathbf{1}_{2m}\mathbf{1}_N^\top - \mathbf{M}$. For the perturbation bounds in Lemma H.3, $\mathbf{M}$ is chosen from initialization: its entries equal to one are those whose activation pattern is constant throughout the columnwise radius-R_2 ball around $\mathbf{W}_0$. Thus the mask does not depend on future iterates. At $t = 0$, use $\mathbf{P}_{-1} = \mathbf{Q}_{-1} = 0$ and $\mathbf{r}_{-1} = \mathbf{r}_0$. For $j \in [N]$, $S_j = \{i \in [2m] \mid \mathbf{M}_{i,j} = 1\}$, and let $S_j^\perp = [2m] \setminus S_j$. In Proposition E.1, we have

$$\boldsymbol{\xi}_t = \boldsymbol{\iota}_t + \boldsymbol{\zeta}_t, \tag{10}$$

$$\boldsymbol{\iota}_t = (1+\beta)\eta(\mathbf{K}_0 - \mathbf{H}_t)\mathbf{r}_t - \beta\eta(\mathbf{K}_0 - \mathbf{H}_{t-1})\mathbf{r}_{t-1}, \tag{11}$$

$$\begin{aligned}
\boldsymbol{\zeta}_t = \mathrm{vec}\Big(&\beta\eta\mathbf{R}_{t-1}\sigma(\mathbf{X}^\top\mathbf{W}_{t-1})(\sigma(\mathbf{W}_t^\top\mathbf{X}) - \sigma(\mathbf{W}_{t-1}^\top\mathbf{X})) \\
&+ \beta\eta\mathbf{P}_{t-1}\left(\left(((\mathbf{V}_{t-1}^\top\mathbf{R}_{t-1}) \odot \mathbf{J}_{t-1})\mathbf{X}^\top\mathbf{X}\right) \odot \mathbf{J}_{t-1} \odot \mathbf{M}\right) \\
&+ \beta\mathbf{P}_{t-1}\left(\left(\mathbf{Q}_{t-1}^\top\mathbf{X}\right) \odot \mathbf{J}_0 \odot \mathbf{M}\right) \\
&+ \mathbf{P}_t\left(\sigma(\mathbf{W}_{t+1}^\top\mathbf{X}) - \sigma(\mathbf{W}_t^\top\mathbf{X})\right) \\
&+ \mathbf{V}_t\left(\left((\mathbf{W}_{t+1}^\top\mathbf{X}) \odot \mathbf{J}_{t+1} - (\mathbf{W}_t^\top\mathbf{X}) \odot \mathbf{J}_t\right) \odot \mathbf{M}^\perp\right) \\
&- \beta\mathbf{V}_{t-1}\left(\left((\mathbf{W}_t^\top\mathbf{X}) \odot \mathbf{J}_t - (\mathbf{W}_{t-1}^\top\mathbf{X}) \odot \mathbf{J}_{t-1}\right) \odot \mathbf{M}^\perp\right) \\
&+ (1+\beta)\eta\boldsymbol{\varphi}_t - \beta\eta\boldsymbol{\varphi}_{t-1}\Big),
\end{aligned} \tag{12}$$

$$[\boldsymbol{\varphi}_t]_{:,j} = \sum_{i \in S_j^\perp} [\mathbf{A}_{i,t}\mathbf{R}_t\mathbf{B}_{i,t}]_{:,j},\ \forall j \in [N]. \tag{13}$$

E.1 General-Response Residual Dynamics and Contraction

The nonlinear residual recursion decomposes into a fixed linear reference system and an error term as follows.

Proposition E.1 (Dynamics)**.** *NAG admits the following residual dynamics:*

$$\begin{pmatrix} \mathbf{r}_{t+1} \\ \mathbf{r}_t \end{pmatrix} = \mathbf{T}_{\mathrm{NAG}} \begin{pmatrix} \mathbf{r}_t \\ \mathbf{r}_{t-1} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\xi}_t \\ 0 \end{pmatrix}, \tag{14}$$

where $\boldsymbol{\xi}_t$ is specified as (10) *in Appendix E and*

$$\mathbf{T}_{\mathrm{NAG}} \coloneqq \begin{pmatrix} (1+\beta)(\mathbf{I}_{Nn} - \eta\mathbf{K}_0) & -\beta(\mathbf{I}_{Nn} - \eta\mathbf{K}_0) \\ \mathbf{I}_{Nn} & 0 \end{pmatrix}.$$

When $\beta = 0$, (7) *corresponds to the GD dynamics $\mathbf{r}_{t+1} = \mathbf{T}_{\mathrm{GD}}\mathbf{r}_t + \boldsymbol{\xi}_t$, where $\mathbf{T}_{\mathrm{GD}} = \mathbf{I}_{Nn} - \eta\mathbf{K}_0$.*

The proof is provided in Appendix E.2. By Proposition E.1, the evolution of the residual is mainly driven by the transition matrices $\mathbf{T}_{\mathrm{GD}}$ and $\mathbf{T}_{\mathrm{NAG}}$ corresponding to the reference systems for GD and NAG. By Lemma D.2, $\mathbf{K}_0$ has full rank and bounded condition number. Therefore, with proper choice of η and β, $\mathbf{T}_{\mathrm{GD}}$ is a Euclidean contraction and $\mathbf{T}_{\mathrm{NAG}}$ is a contraction in a fixed weighted quadratic norm. We provide the proof in Appendix E.3.

Lemma E.2 (Contraction)**.** *On the event in Lemma D.2, denote $L \coloneqq \lambda_1(\mathbf{K}_0)$, $\mu \coloneqq \lambda_{Nn}(\mathbf{K}_0) > 0$. By setting $\beta = 0$ and $\eta = \frac{1}{L}$ it holds $\|\mathbf{T}_{\mathrm{GD}}\| \leq 1 - \frac{1}{\kappa(\mathbf{K}_0)}$. For any fixed $\alpha \in (0,1)$, by setting $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ and $\eta = \frac{1}{L}$, there exists a positive definite matrix $\mathsf{P}_\alpha \succ 0$ such that $\mathsf{P}_\alpha \succeq \mathbf{I}_{2Nn}$ and*

$$\|\mathbf{T}_{\mathrm{NAG}}\mathbf{z}\|_{\mathsf{P}_\alpha} \leq \left(1 - \frac{\alpha}{\sqrt{\kappa(\mathbf{K}_0)}}\right)\|\mathbf{z}\|_{\mathsf{P}_\alpha}, \qquad \forall \mathbf{z} \in \mathbb{R}^{2Nn},$$

where $\|\mathbf{z}\|_{\mathsf{P}_\alpha}^2 \coloneqq \mathbf{z}^\top\mathsf{P}_\alpha\mathbf{z}$. Moreover, there is a constant C_α depending only on α such that

$$\lambda_1(\mathsf{P}_\alpha) \leq C_\alpha\kappa(\mathbf{K}_0)^{3/2}.$$

E.2 Proof of Proposition E.1

Proof of Proposition E.1. To make the decomposition checkable, terms contributing to the linear part or $\boldsymbol{\iota}_t$ are marked in blue. The NAG update is written componentwise as

$$\begin{aligned}
\mathbf{V}_{t+1} &= (1+\beta)(\mathbf{V}_t - \eta \mathbf{R}_t \sigma(\mathbf{X}^\top \mathbf{W}_t)) - \beta(\mathbf{V}_{t-1} - \eta \mathbf{R}_{t-1}\sigma(\mathbf{X}^\top \mathbf{W}_{t-1})),\\
\mathbf{W}_{t+1} &= (1+\beta)\big(\mathbf{W}_t - \eta \mathbf{X}((\mathbf{R}_t^\top \mathbf{V}_t) \odot \mathbb{1}[\mathbf{X}^\top \mathbf{W}_t \geq 0])\big)\\
&\quad - \beta\big(\mathbf{W}_{t-1} - \eta \mathbf{X}((\mathbf{R}_{t-1}^\top \mathbf{V}_{t-1}) \odot \mathbb{1}[\mathbf{X}^\top \mathbf{W}_{t-1} \geq 0])\big).
\end{aligned}$$

This gives the increments

$$\begin{aligned}
\mathbf{P}_t &= \beta \mathbf{P}_{t-1} - (1+\beta)\eta \mathbf{R}_t \sigma(\mathbf{X}^\top \mathbf{W}_t) + \beta\eta \mathbf{R}_{t-1}\sigma(\mathbf{X}^\top \mathbf{W}_{t-1}),\\
\mathbf{Q}_t &= \beta \mathbf{Q}_{t-1} - (1+\beta)\eta \mathbf{X}((\mathbf{R}_t^\top \mathbf{V}_t) \odot \mathbb{1}[\mathbf{X}^\top \mathbf{W}_t \geq 0])\\
&\quad + \beta\eta \mathbf{X}((\mathbf{R}_{t-1}^\top \mathbf{V}_{t-1}) \odot \mathbb{1}[\mathbf{X}^\top \mathbf{W}_{t-1} \geq 0]).
\end{aligned}$$

According to the NAG update rule, we have

$$\begin{aligned}
\mathbf{R}_{t+1} &= \mathbf{V}_{t+1}\sigma(\mathbf{W}_{t+1}^\top \mathbf{X}) - \mathbf{Y}\\
&= (\mathbf{V}_t + \mathbf{P}_t)\sigma(\mathbf{W}_{t+1}^\top \mathbf{X}) - \mathbf{Y}\\
&= \mathbf{R}_t + \underbrace{\mathbf{P}_t \sigma(\mathbf{W}_t^\top \mathbf{X})}_{T_1} + \underbrace{\mathbf{V}_t\left(\sigma(\mathbf{W}_{t+1}^\top \mathbf{X}) - \sigma(\mathbf{W}_t^\top \mathbf{X})\right)}_{T_2} + \underbrace{\mathbf{P}_t\left(\sigma(\mathbf{W}_{t+1}^\top \mathbf{X}) - \sigma(\mathbf{W}_t^\top \mathbf{X})\right)}_{T_3}.
\end{aligned}$$

For the first part, we have

$$\begin{aligned}
T_1 &= \left(\beta \mathbf{P}_{t-1} - (1+\beta)\eta \mathbf{R}_t \sigma(\mathbf{X}^\top \mathbf{W}_t) + \beta\eta \mathbf{R}_{t-1}\sigma(\mathbf{X}^\top \mathbf{W}_{t-1})\right)\sigma(\mathbf{W}_t^\top \mathbf{X})\\
&= \beta \mathbf{P}_{t-1}\sigma(\mathbf{W}_t^\top \mathbf{X}) - (1+\beta)\eta \mathbf{R}_t \sigma(\mathbf{X}^\top \mathbf{W}_t)\sigma(\mathbf{W}_t^\top \mathbf{X}) + \beta\eta \mathbf{R}_{t-1}\sigma(\mathbf{X}^\top \mathbf{W}_{t-1})\sigma(\mathbf{W}_t^\top \mathbf{X})\\
&= \beta \mathbf{P}_{t-1}\sigma(\mathbf{W}_t^\top \mathbf{X}) - (1+\beta)\eta \mathbf{R}_t \sigma(\mathbf{X}^\top \mathbf{W}_t)\sigma(\mathbf{W}_t^\top \mathbf{X}) + \beta\eta \mathbf{R}_{t-1}\sigma(\mathbf{X}^\top \mathbf{W}_{t-1})(\sigma(\mathbf{W}_{t-1}^\top \mathbf{X}))\\
&\quad + \beta\eta \mathbf{R}_{t-1}\sigma(\mathbf{X}^\top \mathbf{W}_{t-1})(\sigma(\mathbf{W}_t^\top \mathbf{X}) - \sigma(\mathbf{W}_{t-1}^\top \mathbf{X}))\\
&= -(1+\beta)\eta \mathbf{R}_t \sigma(\mathbf{X}^\top \mathbf{W}_t)\sigma(\mathbf{W}_t^\top \mathbf{X}) + \beta\eta \mathbf{R}_{t-1}\sigma(\mathbf{X}^\top \mathbf{W}_{t-1})(\sigma(\mathbf{W}_{t-1}^\top \mathbf{X}))\\
&\quad + \beta(\mathbf{V}_t - \mathbf{V}_{t-1})\sigma(\mathbf{W}_t^\top \mathbf{X}) + \beta\eta \mathbf{R}_{t-1}\sigma(\mathbf{X}^\top \mathbf{W}_{t-1})(\sigma(\mathbf{W}_t^\top \mathbf{X}) - \sigma(\mathbf{W}_{t-1}^\top \mathbf{X}))\\
&= \beta \mathbf{R}_t - \beta \mathbf{R}_{t-1} - (1+\beta)\eta \mathbf{R}_t \sigma(\mathbf{X}^\top \mathbf{W}_t)\sigma(\mathbf{W}_t^\top \mathbf{X}) + \beta\eta \mathbf{R}_{t-1}\sigma(\mathbf{X}^\top \mathbf{W}_{t-1})(\sigma(\mathbf{W}_{t-1}^\top \mathbf{X}))\\
&\quad - \beta \mathbf{V}_{t-1}(\sigma(\mathbf{W}_t^\top \mathbf{X}) - \sigma(\mathbf{W}_{t-1}^\top \mathbf{X})) + \beta\eta \mathbf{R}_{t-1}\sigma(\mathbf{X}^\top \mathbf{W}_{t-1})(\sigma(\mathbf{W}_t^\top \mathbf{X}) - \sigma(\mathbf{W}_{t-1}^\top \mathbf{X})).
\end{aligned}$$

For the second part, we have

$$\begin{aligned}
T_2 &= \mathbf{V}_t\left(\sigma(\mathbf{W}_{t+1}^\top \mathbf{X}) - \sigma(\mathbf{W}_t^\top \mathbf{X})\right)\\
&= \mathbf{V}_t\left((\mathbf{W}_{t+1}^\top \mathbf{X}) \odot \mathbb{1}[\mathbf{W}_{t+1}^\top \mathbf{X} \geq 0] - (\mathbf{W}_t^\top \mathbf{X}) \odot \mathbb{1}[\mathbf{W}_t^\top \mathbf{X} \geq 0]\right)\\
&= \mathbf{V}_t\left((\mathbf{W}_{t+1}^\top \mathbf{X}) \odot \mathbb{1}[\mathbf{W}_0^\top \mathbf{X} \geq 0] \odot \mathbf{M} - (\mathbf{W}_t^\top \mathbf{X}) \odot \mathbb{1}[\mathbf{W}_0^\top \mathbf{X} \geq 0] \odot \mathbf{M}\right)\\
&\quad + \underbrace{\mathbf{V}_t\left((\mathbf{W}_{t+1}^\top \mathbf{X}) \odot \mathbb{1}[\mathbf{W}_{t+1}^\top \mathbf{X} \geq 0] \odot \mathbf{M}^\perp - (\mathbf{W}_t^\top \mathbf{X}) \odot \mathbb{1}[\mathbf{W}_t^\top \mathbf{X} \geq 0] \odot \mathbf{M}^\perp\right)}_{T_5}\\
&= \beta \mathbf{V}_t\left((\mathbf{Q}_{t-1}^\top \mathbf{X}) \odot \mathbb{1}[\mathbf{W}_0^\top \mathbf{X} \geq 0] \odot \mathbf{M}\right) + T_5\\
&\quad - (1+\beta)\eta \underbrace{\mathbf{V}_t\left(\left(((\mathbf{R}_t^\top \mathbf{V}_t) \odot \mathbb{1}[\mathbf{X}^\top \mathbf{W}_t \geq 0])^\top \mathbf{X}^\top \mathbf{X}\right) \odot \mathbb{1}[\mathbf{W}_t^\top \mathbf{X} \geq 0] \odot \mathbf{M}\right)}_{T_7}\\
&\quad + \beta\eta \underbrace{\mathbf{V}_t\left(\left(((\mathbf{R}_{t-1}^\top \mathbf{V}_{t-1}) \odot \mathbb{1}[\mathbf{X}^\top \mathbf{W}_{t-1} \geq 0])^\top \mathbf{X}^\top \mathbf{X}\right) \odot \mathbb{1}[\mathbf{W}_{t-1}^\top \mathbf{X} \geq 0] \odot \mathbf{M}\right)}_{T_8}.
\end{aligned}$$

For any $l \in [n]$, $j \in [N]$,

$$
\begin{aligned}
[T_7]_{l,j} &= \sum_{i=1}^{2m} [\mathbf{V}_t]_{l,i} [\big(((\mathbf{R}_t^\top \mathbf{V}_t) \odot \mathbb{1}[\mathbf{X}^\top \mathbf{W}_t \geq 0])^\top \mathbf{X}^\top \mathbf{X}\big) \odot \mathbb{1}[\mathbf{W}_t^\top \mathbf{X} \geq 0] \odot \mathbf{M}]_{i,j} \\
&= \sum_{i=1}^{2m} [\mathbf{V}_t]_{l,i} [\big(((\mathbf{R}_t^\top \mathbf{V}_t) \odot \mathbb{1}[\mathbf{X}^\top \mathbf{W}_t \geq 0])^\top \mathbf{X}^\top \mathbf{X}\big)]_{i,j} \, \mathbb{1}[\mathbf{W}_t^\top \mathbf{X} \geq 0]_{i,j} [\mathbf{M}]_{i,j} \\
&= \left(\sum_{i=1}^{2m} - \sum_{i \in S_j^\perp} \right) [\mathbf{V}_t]_{l,i} [\big(((\mathbf{V}_t^\top \mathbf{R}_t) \odot \mathbb{1}[\mathbf{W}_t^\top \mathbf{X} \geq 0]) \mathbf{X}^\top \mathbf{X}\big)]_{i,j} \, \mathbb{1}[\mathbf{W}_t^\top \mathbf{X} \geq 0]_{i,j} \\
&= \left(\sum_{i=1}^{2m} - \sum_{i \in S_j^\perp} \right) \sum_{k=1}^{N} [\mathbf{V}_t]_{l,i} [\mathbf{V}_t^\top \mathbf{R}_t]_{i,k} \, \mathbb{1}[\mathbf{W}_t^\top \mathbf{X} \geq 0]_{i,k} [\mathbf{X}^\top \mathbf{X}]_{k,j} \, \mathbb{1}[\mathbf{W}_t^\top \mathbf{X} \geq 0]_{i,j} \\
&= \left(\sum_{i=1}^{2m} - \sum_{i \in S_j^\perp} \right) \sum_{k=1}^{N} \sum_{r=1}^{n} [\mathbf{V}_t]_{l,i} [\mathbf{V}_t]_{r,i} [\mathbf{R}_t]_{r,k} \, \mathbb{1}[\mathbf{W}_t^\top \mathbf{X} \geq 0]_{i,k} [\mathbf{X}^\top \mathbf{X}]_{k,j} \, \mathbb{1}[\mathbf{W}_t^\top \mathbf{X} \geq 0]_{i,j}.
\end{aligned}
$$

For T_7, we have

$$
\begin{aligned}
[T_7]_{l,j} &= \left(\sum_{i=1}^{2m} - \sum_{i \in S_j^\perp} \right) \sum_{k=1}^{N} \sum_{r=1}^{n} [\mathbf{A}_{i,t}]_{l,r} [\mathbf{R}_t]_{r,k} [\mathbf{B}_{i,t}]_{k,j} \\
&= \left(\sum_{i=1}^{2m} - \sum_{i \in S_j^\perp} \right) [\mathbf{A}_{i,t} \mathbf{R}_t \mathbf{B}_{i,t}]_{l,j},
\end{aligned}
$$

thus

$$
T_7 = \sum_{i=1}^{2m} \mathbf{A}_{i,t} \mathbf{R}_t \mathbf{B}_{i,t} - T_9,
$$

where $[T_9]_{l,j} = \sum_{i \in S_j^\perp} [\mathbf{A}_{i,t} \mathbf{R}_t \mathbf{B}_{i,t}]_{l,j}$. Meanwhile,

$$
\begin{aligned}
T_8 &= \mathbf{V}_t \left(\left(((\mathbf{R}_{t-1}^\top \mathbf{V}_{t-1}) \odot \mathbb{1}[\mathbf{X}^\top \mathbf{W}_{t-1} \geq 0])^\top \mathbf{X}^\top \mathbf{X} \right) \odot \mathbb{1}[\mathbf{W}_{t-1}^\top \mathbf{X} \geq 0] \odot \mathbf{M} \right) \\
&= (\mathbf{V}_{t-1} + \mathbf{P}_{t-1}) \left(\left(((\mathbf{R}_{t-1}^\top \mathbf{V}_{t-1}) \odot \mathbb{1}[\mathbf{X}^\top \mathbf{W}_{t-1} \geq 0])^\top \mathbf{X}^\top \mathbf{X} \right) \odot \mathbb{1}[\mathbf{W}_{t-1}^\top \mathbf{X} \geq 0] \odot \mathbf{M} \right) \\
&= \sum_{i=1}^{2m} \mathbf{A}_{i,t-1} \mathbf{R}_{t-1} \mathbf{B}_{i,t-1} - T_{10} \\
&\quad + \mathbf{P}_{t-1} \left(\left(((\mathbf{R}_{t-1}^\top \mathbf{V}_{t-1}) \odot \mathbb{1}[\mathbf{X}^\top \mathbf{W}_{t-1} \geq 0])^\top \mathbf{X}^\top \mathbf{X} \right) \odot \mathbb{1}[\mathbf{W}_{t-1}^\top \mathbf{X} \geq 0] \odot \mathbf{M} \right),
\end{aligned}
$$

where T_{10} is T_9 with t replaced by $t-1$. The remaining part of T_1 and T_2 can be combined:

$$
\begin{aligned}
&\beta \mathbf{V}_t \left((\mathbf{Q}_{t-1}^\top \mathbf{X}) \odot \mathbb{1}[\mathbf{W}_0^\top \mathbf{X} \geq 0] \odot \mathbf{M} \right) - \beta \mathbf{V}_{t-1} (\sigma(\mathbf{W}_t^\top \mathbf{X}) - \sigma(\mathbf{W}_{t-1}^\top \mathbf{X})) \\
=&\beta \mathbf{P}_{t-1} \left((\mathbf{Q}_{t-1}^\top \mathbf{X}) \odot \mathbb{1}[\mathbf{W}_0^\top \mathbf{X} \geq 0] \odot \mathbf{M} \right) - \beta T_6,
\end{aligned}
$$

where T_6 is T_5 with t replaced by $t-1$. Collecting the blue terms gives the linear residual recursion and $\boldsymbol{\iota}_t$, while the red terms are exactly the components of $\boldsymbol{\zeta}_t$ listed in (10). This proves the proposition. □

E.3 Proof of Lemma E.2

Proof of Lemma E.2. Since $\mathbf{K}_0$ and $\mathbf{I}_{Nn}$ can be simultaneously diagonalized, eigenvalues of $\mathbf{T}_{\mathrm{GD}}$ correspond to $1 - \eta \lambda_i(\mathbf{K}_0)$, $i \in [Nn]$. Plugging in $\eta = \frac{1}{L}$ yields $\|\mathbf{T}_{\mathrm{GD}}\| = 1 - \frac{\mu}{L} = 1 - \frac{1}{\kappa(\mathbf{K}_0)}$.

We next prove the NAG contraction. Since $\mathbf{K}_0 = \mathbf{U} \operatorname{diag}(\lambda_1, \ldots, \lambda_{Nn}) \mathbf{U}^\top$ for an orthogonal matrix $\mathbf{U}$, let Π be the permutation that interleaves the two eigen-coordinate vectors, and set $\widetilde{\mathbf{U}} = \operatorname{diag}(\mathbf{U}, \mathbf{U}) \Pi^\top$. This orthogonal change of coordinates block diagonalizes $\mathbf{T}_{\mathrm{NAG}}$ into 2×2 blocks

$$
\mathbf{A}_i = \begin{pmatrix} (1+\beta)(1-\eta\lambda_i) & -\beta(1-\eta\lambda_i) \\ 1 & 0 \end{pmatrix}, \qquad i \in [Nn].
$$

The eigenvalues of each block solve

$$z^2 - (1+\beta)z + \beta + \eta(1+\beta)\lambda_i z - \eta\beta\lambda_i = 0,$$

and hence

$$z = \frac{1}{2}\left((1+\beta)(1-\eta\lambda_i) \pm \sqrt{(1-\eta\lambda_i)\left(-4\beta + (1+\beta)^2(1-\eta\lambda_i)\right)}\right).$$

We spell out the standard scalar NAG bound. Let $s = \sqrt{\mu/L}$, $\varrho = 1 - s$, $x_i = \lambda_i/L \in [s^2, 1]$, and $y_i = 1 - x_i$. Since $\beta = (1-s)/(1+s)$, the characteristic polynomial is

$$p_i(z) = z^2 - (1+\beta)y_i z + \beta y_i.$$

If the two roots are complex, their modulus is $\sqrt{\beta y_i} \le \sqrt{\beta(1-s^2)} = \varrho$. If the roots are real, then $\varrho \ge (1+\beta)y_i/2 = y_i/(1+s)$ and

$$p_i(\varrho) = \varrho\left(\varrho - \frac{y_i}{1+s}\right) \ge 0,$$

which implies that the larger root is at most ϱ: since p_i opens upward and ϱ lies to the right of the vertex, $p_i(\varrho) \ge 0$ places ϱ outside the interval between the two real roots. Thus $\rho(\mathbf{A}_i) \le \varrho = 1 - \sqrt{\mu/L}$ for every i. This spectral-radius bound alone does not imply Euclidean operator-norm contraction because the lifted NAG matrix is non-normal. We therefore build a Lyapunov norm.

Fix $\alpha \in (0,1)$ and set $q \coloneqq 1 - \frac{\alpha}{\sqrt{\kappa(\mathbf{K}_0)}}$. Since $q > \rho(\mathbf{A}_i)$ for every i, the series

$$\mathsf{P}_i \coloneqq \sum_{s=0}^{\infty} q^{-2s}(\mathbf{A}_i^s)^\top \mathbf{A}_i^s$$

is well-defined and positive definite. Moreover, $\mathsf{P}_i \succeq \mathbf{I}_2$ and

$$\mathbf{A}_i^\top \mathsf{P}_i \mathbf{A}_i = q^2(\mathsf{P}_i - \mathbf{I}_2) \preceq q^2 \mathsf{P}_i.$$

We also record an explicit norm-equivalence bound. For $\kappa(\mathbf{K}_0) \ge 4$, the Cayley-Hamilton formula for the 2×2 block $\mathbf{A}_i$ writes $\mathbf{A}_i^s = a_{i,s}\mathbf{A}_i + b_{i,s}\mathbf{I}_2$, where $a_{i,s}$ and $b_{i,s}$ are divided differences of x^s at the two eigenvalues of $\mathbf{A}_i$. Together with $\rho(\mathbf{A}_i) \le \varrho$ and uniformly bounded entries of $\mathbf{A}_i$, this gives

$$\|\mathbf{A}_i^s\| \le C(s+1)\varrho^{s-1}, \qquad s \ge 1,$$

for a numerical constant C. Therefore,

$$\lambda_1(\mathsf{P}_i) \le 1 + C\sum_{s=1}^{\infty}(s+1)^2 q^{-2s}\varrho^{2s-2} \le C_\alpha\left(1 - \frac{\varrho^2}{q^2}\right)^{-3} \le C_\alpha \kappa(\mathbf{K}_0)^{3/2},$$

where the middle inequality uses $\sum_{s=0}^{\infty}(s+1)^2 t^s = (1+t)/(1-t)^3$ for $t \in (0,1)$, and the last inequality uses $q - \varrho = \frac{1-\alpha}{\sqrt{\kappa(\mathbf{K}_0)}}$. When $\kappa(\mathbf{K}_0) \le 4$, the same bound follows by increasing C_α, since the block family has a fixed spectral gap $q - \rho(\mathbf{A}_i)$ depending only on α. Define

$$\mathsf{P}_\alpha \coloneqq \widetilde{\mathbf{U}}\,\mathrm{diag}(\mathsf{P}_1, \dots, \mathsf{P}_{Nn})\widetilde{\mathbf{U}}^\top.$$

Then $\mathsf{P}_\alpha \succeq \mathbf{I}_{2Nn}$, $\lambda_1(\mathsf{P}_\alpha) \le C_\alpha \kappa(\mathbf{K}_0)^{3/2}$, and $\mathbf{T}_{\mathrm{NAG}}^\top \mathsf{P}_\alpha \mathbf{T}_{\mathrm{NAG}} \preceq q^2 \mathsf{P}_\alpha$, which gives the desired weighted-norm contraction. □

E.4 Proof of Lemma 4.8

Proof of Lemma 4.8. Let $\mathcal{E} \coloneqq \mathrm{col}(\mathbf{K}_0)$. By Lemma 4.6, $\mathcal{E} = \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$, and the restriction $\mathbf{K}_0|_{\mathcal{E}}$ has positive eigenvalues in $[\mu, L]$. When $r < n$, the lifted NAG map on $\ker(\mathbf{K}_0)$ has eigenvalue 1 on states $(\mathbf{u}^\top, \mathbf{u}^\top)^\top$ with $\mathbf{u} \in \ker(\mathbf{K}_0)$, so a strict contraction on the full space is impossible.

It remains to restrict the reference dynamics to $\mathcal{E}\times\mathcal{E}$. Diagonalizing $\mathbf{K}_0|_{\mathcal{E}}$ and repeating the blockwise Lyapunov construction in Appendix E.3 over the Nr positive eigenvalues yields a positive definite matrix $\mathsf{P}_{\alpha,\mathcal{E}} \succeq \mathbf{I}_{\mathcal{E}\times\mathcal{E}}$ such that

$$\mathbf{T}_{\mathrm{NAG}}^\top \mathsf{P}_{\alpha,\mathcal{E}} \mathbf{T}_{\mathrm{NAG}} \preceq \left(1 - \frac{\alpha}{\sqrt{\kappa(\mathbf{K}_0)}}\right)^2 \mathsf{P}_{\alpha,\mathcal{E}} \quad \text{on } \mathcal{E}\times\mathcal{E}.$$

The bound $\lambda_1(\mathsf{P}_{\alpha,\mathcal{E}}) \le C_\alpha \kappa(\mathbf{K}_0)^{3/2}$ follows from the same blockwise estimate in Appendix E.3: in the eigenbasis of $\mathbf{K}_0|_{\mathcal{E}}$, the Lyapunov matrix is block diagonal with 2×2 blocks, so $\lambda_1(\mathsf{P}_{\alpha,\mathcal{E}}) = \max_{i \le Nr} \lambda_1(\mathsf{P}_i)$. By Lemma 4.7, this gives reference-system stability on $\mathcal{E} \times \mathcal{E}$. The nonlinear convergence proof in Appendix H uses the residual together with the parameter momentum. □

F Kernel-Shift Proofs

We first state an auxiliary result on the mean absolute value of the product of correlated Gaussian random variables.

Proposition F.1 (Corollary 3.1 in Li and Wei [2009]). *Let $\mathbf{u}, \mathbf{v} \in \mathbb{R}^N$ be nonzero, $\boldsymbol{\phi} \sim \mathcal{N}(0, \mathbf{I}_N)$, $\rho \coloneqq \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$, then*

$$\mathbb{E}\left|(\mathbf{u}^\top \boldsymbol{\phi})(\mathbf{v}^\top \boldsymbol{\phi})\right| = \frac{2}{\pi}\left(\sqrt{1-\rho^2} + \rho \arcsin \rho\right) \|\mathbf{u}\| \, \|\mathbf{v}\| \ge \frac{2}{\pi} \|\mathbf{u}\| \, \|\mathbf{v}\| .$$

With this result, we show the kernel shift can be properly controlled by the change in weights with high probability.

F.1 General-Response Kernel-Shift

We bound the deviation from the reference system, including the kernel shift in NTK, $\mathbf{K}_t - \mathbf{K}_0$, and the whole CK, $\mathbf{G}_t$, under the assumption that the weights remain close to their unbalanced initialization.

Lemma F.2 (Kernel shift). *For any $\delta \in (0,1)$, $R_1 > 0$ and $R_2 \in (0, c_2)$, suppose the width satisfies*

$$m \ge \Omega\left(\frac{c_2}{R_2} \cdot \log^2\left(\frac{2N^2 n^2 m}{\delta}\right)\right).$$

Then with probability at least $1-\delta$, the following inequalities hold for any $\mathbf{V}_t$ and $\mathbf{W}_t$ such that $\|[\mathbf{V}_t]_{:,i} - [\mathbf{V}_0]_{:,i}\| \le R_1$, $\|[\mathbf{W}_t]_{:,i} - [\mathbf{W}_0]_{:,i}\| \le R_2$, $\forall i \in [2m]$:

$$\|\mathbf{K}_0 - \mathbf{K}_t\| \le 6mNR_V^2(R_2/c_2) + 4mNR_1(2R_V + R_1), \ \|\mathbf{G}_t\| \le 2m \|\mathbf{X}\|^2 (R_W + R_2)^2.$$

Moreover, if $R_V^2 \cdot \frac{R_2}{c_2} \ge R_1(2R_V + R_1)$, then $\|\mathbf{K}_0 - \mathbf{K}_t\| \le 10mNR_V^2(R_2/c_2)$. The probability event is simultaneous over all radius-bounded $\mathbf{V}_t, \mathbf{W}_t$ satisfying the displayed inequalities and depends only on the initialization, so it can be applied to adaptive training trajectories without an additional union bound over time.

The proof is provided in Appendix F.2. Lemma F.2 gives a time-uniform radius-ball event. For the radii used in the main theorems, Lemma H.7 controls the Jacobian directly.

F.2 Proof of Lemma F.2

Proof of Lemma F.2. We will bound each of the two coupled halves separately and then add the bounds. The estimates below are written for the first m neurons; the same argument applies to the second half after relabeling, since the initialization and the radius bounds are uniform over all $i \in [2m]$. *Step 1: kernel-shift decomposition.* Let $\mathbf{D}_i \coloneqq \mathbf{B}_{i,t}^\top \otimes \mathbf{A}_{i,t} - \mathbf{B}_{i,0}^\top \otimes \mathbf{A}_{i,0}$. We first split the full $2m$-neuron kernel shift into the two halves:

$$\|\mathbf{K}_0 - \mathbf{K}_t\| \le \left\|\sum_{i=1}^{m} \mathbf{D}_i\right\|_{\mathrm{F}} + \left\|\sum_{i=m+1}^{2m} \mathbf{D}_i\right\|_{\mathrm{F}}.$$

For the first half,

$$\left\|\sum_{i=1}^{m} \mathbf{D}_i\right\|_{\mathrm{F}} \le \left\|\sum_{i=1}^{m} (\mathbf{B}_{i,t} - \mathbf{B}_{i,0})^\top \otimes \mathbf{A}_{i,0}\right\|_{\mathrm{F}} + \left\|\sum_{i=1}^{m} \mathbf{B}_{i,t}^\top \otimes (\mathbf{A}_{i,t} - \mathbf{A}_{i,0})\right\|_{\mathrm{F}}.$$

For the first part, we have

$$\left\|\sum_{i=1}^{m}\left(\mathbf{B}_{i,t}-\mathbf{B}_{i,0}\right)^{\top}\otimes\mathbf{A}_{i,0}\right\|_{\mathrm{F}}^{2}=\sum_{p=1}^{n}\sum_{q=1}^{n}\sum_{j=1}^{N}\sum_{k=1}^{N}\left(\sum_{i=1}^{m}[\mathbf{B}_{i,t}-\mathbf{B}_{i,0}]_{j,k}[\mathbf{A}_{i,0}]_{p,q}\right)^{2}$$
$$\leq\sum_{p=1}^{n}\sum_{q=1}^{n}\sum_{j=1}^{N}\sum_{k=1}^{N}\left(\sum_{i=1}^{m}s_{i,j,k}a_{i,p,q}\right)^{2},$$

where we denote

$$a_{i,p,q}\coloneqq\left|[\mathbf{A}_{i,0}]_{p,q}\right|,\ s_{i,j,k}\coloneqq\left|[\mathbf{J}_t]_{i,j}[\mathbf{J}_t]_{i,k}-[\mathbf{J}_0]_{i,j}[\mathbf{J}_0]_{i,k}\right|.$$

Step 2: time-uniform sign-change domination. For notation simplicity, we fix p,q,j,k for now and denote $s_i=s_{i,j,k}$ and $a_i=a_{i,p,q}$. To make the event uniform over any future iterate inside the radius ball, define the worst-case indicator

$$\widehat{s}_{i,j,k}\coloneqq\sup_{\|\mathbf{w}-[\mathbf{W}_0]_{:,i}\|\leq R_2}\left|\mathbb{1}[\mathbf{w}^{\top}\mathbf{x}_j\geq 0]\,\mathbb{1}[\mathbf{w}^{\top}\mathbf{x}_k\geq 0]-\mathbb{1}[[\mathbf{W}_0]_{:,i}^{\top}\mathbf{x}_j\geq 0]\,\mathbb{1}[[\mathbf{W}_0]_{:,i}^{\top}\mathbf{x}_k\geq 0]\right|.$$

Whenever $\|[\mathbf{W}_t]_{:,i}-[\mathbf{W}_0]_{:,i}\|\leq R_2$, we have $s_{i,j,k}\leq\widehat{s}_{i,j,k}$. The variables $\widehat{s}_{i,j,k}$ depend only on the independent initialization vectors $[\mathbf{W}_0]_{:,i}$, hence are independent over i. The event $\widehat{s}_{i,j,k}=1$ implies that at least one of the two margins can be crossed inside the radius-R_2 ball, and therefore

$$\widehat{s}_{i,j,k}\leq\mathbb{1}\left[\left|[\mathbf{W}_0]_{:,i}^{\top}\mathbf{x}_j\right|\leq R_2\left\|\mathbf{x}_j\right\|\right]+\mathbb{1}\left[\left|[\mathbf{W}_0]_{:,i}^{\top}\mathbf{x}_k\right|\leq R_2\left\|\mathbf{x}_k\right\|\right].$$

Since the data are normalized and $[\mathbf{W}_0]_{:,i}\sim\mathcal{N}(0,c_2^2\mathbf{I}_d)$, Gaussian anti-concentration (the same estimate used in Lemma G.2 of Munteanu et al. [2022]) gives, after adjusting absolute constants,

$$\mathbb{E}[\widehat{s}_{i,j,k}]\leq\frac{2R_2}{c_2}.$$

If $R_2/c_2\geq 1/3$, then $\sum_i a_i s_i\leq\sum_i a_i\leq(3R_2/c_2)\sum_i a_i$ deterministically; thus it suffices to consider $R_2/c_2<1/3$. In this case set $p=2R_2/c_2$ and $q_i=\mathbb{E}[\widehat{s}_{i,j,k}]\leq p$. Enlarging the probability space if necessary, let U_i be independent uniform random variables on $[0,1]$, independent of the initialization, and define

$$X_i=\widehat{s}_{i,j,k}+(1-\widehat{s}_{i,j,k})\,\mathbb{1}\left[U_i\leq\frac{p-q_i}{1-q_i}\right].$$

Then $X_1,\ldots,X_m$ are independent Bernoulli random variables with $\mathbb{E}[X_i]=p=\frac{2R_2}{c_2}$ and $X_i\geq\widehat{s}_{i,j,k}\geq s_i$ for all $i\in[m]$. Condition on $\mathbf{\Phi}$, so the weights a_i are fixed, and write $\widehat{s}_i=\widehat{s}_{i,j,k}$. By the weighted Chernoff inequality (Lemma C.3), we have the following bound:

$$\mathbb{P}\left(\sum_{i=1}^{m}a_i\widehat{s}_i\geq\frac{3R_2}{c_2}\sum_{i=1}^{m}a_i\,\middle|\,\mathbf{\Phi}\right)\leq\mathbb{P}\left(\sum_{i=1}^{m}a_iX_i\geq\frac{3}{2}\cdot\sum_{i=1}^{m}a_i\mathbb{E}[X_i]\,\middle|\,\mathbf{\Phi}\right)$$
$$\leq\exp\left(-\frac{1}{10}\frac{\sum_{i=1}^{m}a_i\cdot\frac{2R_2}{c_2}}{\max_{i\in[m]}a_i}\right).\tag{15}$$

Step 3: concentration of the output weights. Now we study the concentration properties for a_i's to ensure the above bound is valid. By Proposition F.1 we have

$$\mathbb{E}[a_i]=c_1^2\mathbb{E}\left|[\mathbf{\Phi}]_{p,i}[\mathbf{\Phi}]_{q,i}\right|=c_1^2\mathbb{E}\left|(\mathbf{e}_p^{\top}\boldsymbol{\phi}_i)(\mathbf{e}_q^{\top}\boldsymbol{\phi}_i)\right|\geq\frac{2c_1^2}{\pi}.$$

On the other hand, since $[\mathbf{\Phi}]_{p,i}$ and $[\mathbf{\Phi}]_{q,i}$ are Gaussian, a_i and its centered version are sub-exponential with bounded ψ_1-norm

$$\left\|a_i-\mathbb{E}[a_i]\right\|_{\psi_1}\lesssim\left\|a_i\right\|_{\psi_1}\lesssim c_1^2.$$

By plugging in $t=\frac{c_1^2m}{\pi}$ in Bernstein's inequality (Lemma C.2), we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{m}(a_i-\mathbb{E}[a_i])\right|\geq\frac{c_1^2m}{\pi}\right)\leq 2\exp\left(-c'm\right),$$

where c' is an absolute constant. Therefore, for any $\delta_2 \in (0,1)$, by setting $m \geq \Omega(\log \frac{2n^2}{\delta_2})$ and applying a union bound over $(p,q) \in [n]^2$, with probability at least $1-\delta_2$ we have

$$\sum_{i=1}^{m} a_{i,p,q} \geq \frac{c_1^2 m}{\pi},\ \forall p,q \in [n].$$

For the denominator, by union bound on sub-Gaussian tails, it holds

$$\mathbb{P}\left(\exists i \in [m], \exists p \in [n],\ |[\mathbf{\Phi}]_{p,i}| > \sqrt{c \log \frac{2mn}{\delta_3}}\right) \leq \delta_3$$

for any $\delta_3 \in (0,1)$, where c is an absolute constant. Thus with probability at least $1-\delta_3$, we have

$$a_{i,p,q} \lesssim c_1^2 \log\left(\frac{2mn}{\delta_3}\right),\ \forall i \in [m], \forall p,q \in [n].$$

Plug the lower bound on the numerator and upper bound on the denominator back to (15), with probability at least $1-\delta_2-\delta_3$ we have

$$\mathbb{P}\left(\sum_{i=1}^{m} a_i \widehat{s}_i \geq \frac{3R_2}{c_2} \sum_{i=1}^{m} a_i \,\middle|\, \mathbf{\Phi}\right) \leq \exp\left(-\frac{c'' m R_2}{c_2 \log \frac{2mn}{\delta_3}}\right)$$

for some absolute constant $c'' > 0$. Applying a union bound over $(p,q) \in [n]^2$ and $(j,k) \in [N]^2$, for any $\delta_4 \in (0,1)$, by setting

$$m = \max\left\{\Omega\left(\frac{c_2}{R_2} \cdot \log\left(\frac{2mn}{\delta_3}\right) \log\left(\frac{2N^2n^2}{\delta_4}\right)\right), \Omega\left(\log\left(\frac{2n^2}{\delta_2}\right)\right)\right\},$$

we have with probability at least $1-\delta_2-\delta_3-\delta_4$ that

$$\begin{aligned}
\left\|\sum_{i=1}^{m} \left(\mathbf{B}_{i,t} - \mathbf{B}_{i,0}\right)^\top \otimes \mathbf{A}_{i,0}\right\|_{\mathrm{F}}^2 &\leq \left(\frac{3NR_2}{c_2}\right)^2 \sum_{p=1}^{n}\sum_{q=1}^{n}\left(\sum_{i=1}^{m} |[\mathbf{A}_{i,0}]_{p,q}|\right)^2 \\
&\leq \left(\frac{3NR_2}{c_2}\right)^2 \sum_{p=1}^{n}\sum_{q=1}^{n}\left(m\sum_{i=1}^{m} |[\mathbf{A}_{i,0}]_{p,q}|^2\right) \\
&\leq \left(\frac{3NR_2}{c_2}\right)^2 m \sum_{i=1}^{m} \|\mathbf{A}_{i,0}\|_{\mathrm{F}}^2 \\
&\leq \left(\frac{3mNR_2}{c_2}\right)^2 R_V^4,
\end{aligned}$$

where the second line uses the Cauchy–Schwarz inequality.

Step 4: deterministic control of the remaining terms. For the second part, we have

$$\begin{aligned}
\left\|\sum_{i=1}^{m} \mathbf{B}_{i,t}^\top \otimes (\mathbf{A}_{i,t} - \mathbf{A}_{i,0})\right\|_{\mathrm{F}}^2 &\leq \sum_{p=1}^{n}\sum_{q=1}^{n}\sum_{j=1}^{N}\sum_{k=1}^{N}\left(\sum_{i=1}^{m}[\mathbf{B}_{i,t}]_{j,k}^2\right)\left(\sum_{i=1}^{m}[\mathbf{A}_{i,t} - \mathbf{A}_{i,0}]_{p,q}^2\right) \\
&\leq mN^2 \sum_{i=1}^{m} \|\mathbf{A}_{i,t} - \mathbf{A}_{i,0}\|_{\mathrm{F}}^2 \\
&\leq 4m^2N^2R_1^2(2R_V + R_1)^2,
\end{aligned}$$

where we use Cauchy–Schwarz for the first line and $[\mathbf{B}_{i,t}]_{j,k} \leq 1$ for the second line. Combining the two parts gives the first-half bound

$$\left\|\sum_{i=1}^{m} \mathbf{D}_i\right\|_{\mathrm{F}} \leq 3mN\frac{R_2}{c_2}R_V^2 + 2mNR_1(2R_V + R_1).$$

The second half obeys the same bound by the same proof. We allocate the auxiliary failure budgets across the two halves; this only changes absolute constants inside the logarithms and is absorbed by the displayed width condition. Therefore the preceding two-half decomposition yields

$$\|\mathbf{K}_0 - \mathbf{K}_t\| \le 6mN\frac{R_2}{c_2}R_V^2 + 4mNR_1(2R_V + R_1).$$

Moreover, if $\frac{R_2}{c_2}R_V^2 \ge R_1(2R_V + R_1)$, then

$$\|\mathbf{K}_0 - \mathbf{K}_t\| \le 10mN\frac{R_2}{c_2}R_V^2.$$

Meanwhile, by 1-Lipschitzness of ReLU, we have

$$\|\mathbf{G}_t\| \le \left\|\sigma(\mathbf{X}^\top\mathbf{W}_t)\sigma(\mathbf{W}_t^\top\mathbf{X})\right\| \le \left\|\mathbf{W}_t^\top\mathbf{X}\right\|_{\mathrm{F}}^2 \le 2m(R_W + R_2)^2\|\mathbf{X}\|^2 \le 2mN(R_W + R_2)^2.$$

Picking $\delta_i = \frac{\delta}{3}$ for $i = 2, 3, 4$ gives the final result. □

F.3 Low-Rank Kernel Shift

Lemma F.3 (Low-rank kernel shift)**.** *For sketching initialization* (8)*, the conclusions of Lemma F.2 hold with the same width condition and the same time-uniform interpretation over radius-bounded iterates.*

Proof of Lemma F.3. We use the same kernel-shift decomposition as in Appendix F.2. The uniform sign-change indicators $\widehat{s}_{i,j,k}$ from Appendix F.2 only depend on the hidden initialization $\mathbf{W}_0$, which is the same in (8) and (4); hence their independence, monotone domination, and anti-concentration bounds are unchanged. The only separate check is that, under the sketching initialization (8), the output-layer weights in $\sum_{i=1}^m a_i s_i$ have the numerator/denominator concentration needed for the weighted Chernoff step, where a_i and s_i are defined as in Appendix F.2. For nonzero row pairs, conditioning on $\mathbf{\Phi}$ as above gives

$$\mathbb{P}\left(\sum_{i=1}^m a_i\widehat{s}_i \ge \frac{3R_2}{c_2}\sum_{i=1}^m a_i \,\middle|\, \mathbf{\Phi}\right) \le \exp\left(-\frac{1}{10}\frac{\sum_{i=1}^m a_i \cdot \frac{2R_2}{c_2}}{\max_{i\in[m]} a_i}\right). \tag{16}$$

Given sketching initialization (8) and Proposition F.1, we have

$$\mathbb{E}[a_i] = c_1^2\mathbb{E}\left|[\mathbf{Y}\mathbf{\Phi}]_{p,i}[\mathbf{Y}\mathbf{\Phi}]_{q,i}\right| \ge \frac{2c_1^2}{\pi}\left\|[\mathbf{Y}]_{p,:}\right\|\left\|[\mathbf{Y}]_{q,:}\right\|.$$

If $\left\|[\mathbf{Y}]_{p,:}\right\|\left\|[\mathbf{Y}]_{q,:}\right\| = 0$, then $a_{i,p,q} = 0$ for every i and this (p, q) contribution is identically zero. We therefore only need the following concentration argument for pairs with $\left\|[\mathbf{Y}]_{p,:}\right\|\left\|[\mathbf{Y}]_{q,:}\right\| > 0$. On the other hand, we have

$$\left\|[\mathbf{Y}\mathbf{\Phi}]_{p,i}\right\|_{\psi_2} \lesssim \left\|[\mathbf{Y}]_{p,:}\right\|,$$

thus a_i and its centered version are sub-exponential with ψ_1-norm bounded by

$$\left\|a_i - \mathbb{E}[a_i]\right\|_{\psi_1} \lesssim \left\|a_i\right\|_{\psi_1} \lesssim c_1^2\left\|[\mathbf{Y}]_{p,:}\right\|\left\|[\mathbf{Y}]_{q,:}\right\|.$$

By plugging in $t = \frac{c_1^2 m\|[\mathbf{Y}]_{p,:}\|\|[\mathbf{Y}]_{q,:}\|}{\pi}$ in Bernstein's inequality (Lemma C.2), we have

$$\mathbb{P}\left(\left|\sum_{i=1}^m (a_i - \mathbb{E}[a_i])\right| \ge \frac{c_1^2 m}{\pi}\left\|[\mathbf{Y}]_{p,:}\right\|\left\|[\mathbf{Y}]_{q,:}\right\|\right) \le 2\exp\left(-c'm\right)$$

for some absolute constant c'. Therefore, for any $\delta_2 \in (0, 1)$, by setting $m \ge \Omega(\log\frac{2n^2}{\delta_2})$ and applying a union bound over $(p, q) \in [n]^2$, with probability at least $1 - \delta_2$ we have

$$\sum_{i=1}^m a_{i,p,q} \ge \frac{c_1^2 m}{\pi}\left\|[\mathbf{Y}]_{p,:}\right\|\left\|[\mathbf{Y}]_{q,:}\right\|, \ \forall p, q \in [n].$$

For the denominator, by union bound on sub-Gaussian tails, it holds

$$\mathbb{P}\left(\exists i \in [m], \exists p \in [n],\ |[\mathbf{Y}\mathbf{\Phi}]_{p,i}| > \sqrt{c \log \frac{2mn}{\delta_3}}\, \|[\mathbf{Y}]_{p,:}\|\right) \leq \delta_3$$

for any $\delta_3 \in (0,1)$, where c is an absolute constant. Thus with probability at least $1-\delta_3$, we have

$$a_{i,p,q} \lesssim c_1^2\, \|[\mathbf{Y}]_{p,:}\|\, \|[\mathbf{Y}]_{q,:}\| \log\left(\frac{2mn}{\delta_3}\right),\ \forall i \in [m], \forall p, q \in [n].$$

Plug the lower bound on the numerator and upper bound on the denominator back to (16), with probability at least $1-\delta_2-\delta_3$ we have

$$\mathbb{P}\left(\sum_{i=1}^m a_i \widehat{s}_i \geq \frac{3R_2}{c_2} \sum_{i=1}^m a_i \,\middle|\, \mathbf{\Phi}\right) \leq \exp\left(-\frac{c'' m R_2}{c_2 \log \frac{2mn}{\delta_3}}\right)$$

for some absolute constant $c'' > 0$. Combining this numerator/denominator ratio with (16) and applying the union bound over $(j,k) \in [N]^2$ and $(p,q) \in [n]^2$ gives the same first-part kernel-shift bound as in Appendix F.2. The resulting $\log(2N^2n^2m/\delta)$ factors are exactly those recorded in Lemma F.2 and are absorbed into the theorem-level polylogarithmic width condition. The deterministic bounds for the second kernel-shift part and for $\|\mathbf{G}_t\|$ use only the radius assumptions and normalized data, so they are unchanged under the sketching initialization. This proves the two inequalities in Lemma F.3, and the sharpened bound follows from the same condition $R_V^2(R_2/c_2) \geq R_1(2R_V + R_1)$. □

G Subspace Alignment Proofs

This section identifies the positive-eigenvalue subspace in the low-rank setting and proves that the residual and error terms stay in that subspace.

G.1 Proof of Lemma 4.6

Proof of Lemma 4.6. By sketching initialization (8), $\mathrm{col}(\mathbf{V}_0) = \mathrm{col}(\mathbf{Y}\mathbf{\Phi}) \subseteq \mathrm{col}(\mathbf{Y})$. Since $\mathbf{K}_0 = \sum_{i=1}^{2m} \mathbf{B}_i^\top \otimes \mathbf{A}_i$ is symmetric, its row space is equivalent to column space. For any $\mathbf{w} = (\mathbf{w}_1^\top, \ldots, \mathbf{w}_N^\top)^\top \in \mathbb{R}^{Nn}$, $\mathbf{w}_j \in \mathbb{R}^n$, $\forall j \in [N]$, it holds for every $i \in [2m]$ that

$$\begin{aligned}(\mathbf{B}_i^\top \otimes \mathbf{A}_i)\mathbf{w} &= \begin{pmatrix} [\mathbf{B}_i]_{1,1}\mathbf{A}_i & \ldots & [\mathbf{B}_i]_{1,N}\mathbf{A}_i \\ \vdots & & \vdots \\ [\mathbf{B}_i]_{N,1}\mathbf{A}_i & \ldots & [\mathbf{B}_i]_{N,N}\mathbf{A}_i \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_N \end{pmatrix} \\ &= \begin{pmatrix} \sum_{j=1}^N [\mathbf{B}_i]_{1,j}\mathbf{A}_i\mathbf{w}_j \\ \vdots \\ \sum_{j=1}^N [\mathbf{B}_i]_{N,j}\mathbf{A}_i\mathbf{w}_j \end{pmatrix} \in \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y}),\end{aligned}$$

where the second line is from that $\mathrm{col}(\mathbf{A}_i) = \mathrm{col}([\mathbf{V}_0]_{:,i}[\mathbf{V}_0]_{:,i}^\top) \subseteq \mathrm{col}(\mathbf{V}_0) \subseteq \mathrm{col}(\mathbf{Y})$. Therefore, we have $\mathrm{col}(\mathbf{K}_0) \subseteq \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$. By assumption, $\mathrm{rank}(\mathbf{Y}) = r$, hence $\dim(\mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})) = Nr$. By Lemma 4.5, $\lambda_{Nr}(\mathbf{K}_0) \geq \frac{1}{2}c_1^2 m\lambda\sigma_r^2(\mathbf{Y})$, thus we have $Nr \leq \mathrm{rank}(\mathbf{K}_0) \leq \dim(\mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})) = Nr$, which implies that $\mathrm{col}(\mathbf{K}_0) = \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$. □

G.2 Proof of Lemma 4.7

Proof of Lemma 4.7. We prove the lemma by induction. For any subspace $U \subseteq \mathbb{R}^n$, $\mathrm{vec}(\mathbf{M}) \in \mathbb{R}^N \otimes U$ is equivalent to every column of $\mathbf{M} \in \mathbb{R}^{n \times N}$ lying in U under our column-first vectorization convention. We use the following closure facts repeatedly. First, if the columns of a matrix $\mathbf{M}$ lie in $\mathrm{col}(\mathbf{Y})$, then so do the columns of $\mathbf{MN}$ for any compatible matrix $\mathbf{N}$. Second, if $\mathrm{col}(\mathbf{V}_t) \subseteq \mathrm{col}(\mathbf{Y})$, then each $\mathbf{A}_{i,t} = [\mathbf{V}_t]_{:,i}[\mathbf{V}_t]_{:,i}^\top$ maps into $\mathrm{col}(\mathbf{Y})$, while $\mathbf{G}_t = (\text{an } N \times N \text{ matrix}) \otimes \mathbf{I}_n$ preserves $\mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$.

We first show by induction that $\mathrm{col}(\mathbf{V}_t), \mathrm{col}(\mathbf{P}_t), \mathrm{col}(\mathbf{R}_t) \subseteq \mathrm{col}(\mathbf{Y})$ for all $t \geq 0$, with the convention $\mathbf{P}_{-1} = 0$ and $\mathbf{R}_{-1} = \mathbf{R}_0$. By initialization (8), $\mathrm{col}(\mathbf{V}_0) \subseteq \mathrm{col}(\mathbf{Y})$, hence $\mathrm{col}(\mathbf{R}_0) = \mathrm{col}(\mathbf{V}_0\sigma(\mathbf{W}_0^\top \mathbf{X}) - \mathbf{Y}) \subseteq \mathrm{col}(\mathbf{Y})$. If the claim holds up to time t, then the update formula for $\mathbf{P}_t$ in Appendix E.2 gives $\mathrm{col}(\mathbf{P}_t) \subseteq \mathrm{col}(\mathbf{Y})$ because it is a linear combination of $\mathbf{P}_{t-1}$, $\mathbf{R}_t\sigma(\mathbf{X}^\top \mathbf{W}_t)$, and $\mathbf{R}_{t-1}\sigma(\mathbf{X}^\top \mathbf{W}_{t-1})$. Therefore $\mathrm{col}(\mathbf{V}_{t+1}) \subseteq \mathrm{col}(\mathbf{Y})$, and then $\mathrm{col}(\mathbf{R}_{t+1}) = \mathrm{col}(\mathbf{V}_{t+1}\sigma(\mathbf{W}_{t+1}^\top \mathbf{X}) - \mathbf{Y}) \subseteq \mathrm{col}(\mathbf{Y})$. This proves $\mathbf{r}_t = \mathrm{vec}(\mathbf{R}_t) \in \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$ for all t.

It remains to show $\boldsymbol{\xi}_t \in \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$. By definition of $\boldsymbol{\varphi}_t$, for any $j \in [N]$,

$$[\boldsymbol{\varphi}_t]_{:,j} = \sum_{i \in S_j^\perp} [\mathbf{A}_{i,t}\mathbf{R}_t\mathbf{B}_{i,t}]_{:,j} \in \mathrm{col}(\mathbf{V}_t) \subseteq \mathrm{col}(\mathbf{Y}),$$

where the inclusion uses $\mathbf{A}_{i,t} = [\mathbf{V}_t]_{:,i}[\mathbf{V}_t]_{:,i}^\top$. Thus $\mathrm{vec}(\boldsymbol{\varphi}_t) \in \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$. For the kernel terms in $\boldsymbol{\iota}_t$, each NTK summand satisfies the standard identity $(\mathbf{B}_{i,t}^\top \otimes \mathbf{A}_{i,t})\,\mathrm{vec}(\mathbf{R}_t) = \mathrm{vec}(\mathbf{A}_{i,t}\mathbf{R}_t\mathbf{B}_{i,t})$ and hence lies in $\mathbb{R}^N \otimes \mathrm{col}(\mathbf{A}_{i,t})$. Equivalently, in block form,

$$\begin{aligned}(\mathbf{B}_{i,t}^\top \otimes \mathbf{A}_{i,t})\mathbf{r}_t &= \begin{pmatrix} [\mathbf{B}_{i,t}]_{1,1}\mathbf{A}_{i,t} & \dots & [\mathbf{B}_{i,t}]_{1,N}\mathbf{A}_{i,t} \\ \vdots & & \vdots \\ [\mathbf{B}_{i,t}]_{N,1}\mathbf{A}_{i,t} & \dots & [\mathbf{B}_{i,t}]_{N,N}\mathbf{A}_{i,t} \end{pmatrix} \begin{pmatrix} [\mathbf{R}_t]_{:,1} \\ \vdots \\ [\mathbf{R}_t]_{:,N} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{j=1}^N [\mathbf{B}_{i,t}]_{1,j}[\mathbf{A}_{i,t}\mathbf{R}_t]_{:,j} \\ \vdots \\ \sum_{j=1}^N [\mathbf{B}_{i,t}]_{N,j}[\mathbf{A}_{i,t}\mathbf{R}_t]_{:,j} \end{pmatrix} \in \mathbb{R}^N \otimes \mathrm{col}(\mathbf{V}_t).\end{aligned}$$

Summing over i gives $\mathbf{K}_t\mathbf{r}_t \in \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$. Similarly, we have

$$\begin{aligned}\mathbf{G}_t\mathbf{r}_t &= \left((\sigma(\mathbf{x}^\top \mathbf{W}_t)\sigma(\mathbf{W}_t^\top \mathbf{x})) \otimes \mathbf{I}_n\right)\mathbf{r}_t \\ &= \begin{pmatrix} (\sigma(\mathbf{x}_1^\top \mathbf{W}_t)\sigma(\mathbf{W}_t^\top \mathbf{x}_1))\mathbf{I}_n & \dots & (\sigma(\mathbf{x}_1^\top \mathbf{W}_t)\sigma(\mathbf{W}_t^\top \mathbf{x}_N))\mathbf{I}_n \\ \vdots & & \vdots \\ (\sigma(\mathbf{x}_N^\top \mathbf{W}_t)\sigma(\mathbf{W}_t^\top \mathbf{x}_1))\mathbf{I}_n & \dots & (\sigma(\mathbf{x}_N^\top \mathbf{W}_t)\sigma(\mathbf{W}_t^\top \mathbf{x}_N))\mathbf{I}_n \end{pmatrix} \begin{pmatrix} [\mathbf{R}_t]_{:,1} \\ \vdots \\ [\mathbf{R}_t]_{:,N} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{j=1}^N (\sigma(\mathbf{x}_1^\top \mathbf{W}_t)\sigma(\mathbf{W}_t^\top \mathbf{x}_j))[\mathbf{R}_t]_{:,j} \\ \vdots \\ \sum_{j=1}^N (\sigma(\mathbf{x}_N^\top \mathbf{W}_t)\sigma(\mathbf{W}_t^\top \mathbf{x}_j))[\mathbf{R}_t]_{:,j} \end{pmatrix} \in \mathbb{R}^N \otimes \mathrm{col}(\mathbf{R}_t).\end{aligned}$$

Therefore $\mathbf{H}_t\mathbf{r}_t = (\mathbf{K}_t + \mathbf{G}_t)\mathbf{r}_t \in \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$, and Lemma 4.6 gives $\mathbf{K}_0\mathbf{r}_t \in \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$. The same argument applies to the $t-1$ terms in (10). Hence $\boldsymbol{\iota}_t \in \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$. Finally, every matrix term inside $\boldsymbol{\zeta}_t$ in (10) has columns in $\mathrm{col}(\mathbf{Y})$: the terms with $\mathbf{R}_s$ or $\mathbf{P}_s$ preserve this column space by right multiplication, the terms with $\mathbf{V}_s$ are in $\mathrm{col}(\mathbf{V}_s) \subseteq \mathrm{col}(\mathbf{Y})$, and the $\boldsymbol{\varphi}_s$ terms were handled above. Thus $\boldsymbol{\zeta}_t \in \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$, and $\boldsymbol{\xi}_t = \boldsymbol{\iota}_t + \boldsymbol{\zeta}_t \in \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$ for all $t \geq 0$. □

H Convergence Proofs

We first give auxiliary bounds for the residual decomposition in Appendix E. The proofs of the three main theorems are given in Appendices H.6 to H.8, respectively.

H.1 Auxiliary Lemma

Lemma H.1. *Suppose $\{a_t\}_{t\geq 0}$ and $\{b_t\}_{t\geq 0}$ are two non-negative sequences satisfying*

$$a_{t+1} \leq \rho \cdot a_t + b_t, \quad b_t \leq \theta^t \cdot c_0,$$

where $0 \leq \rho < \theta < 1$, $c_0 \geq 0$, then the following holds for all $t \geq 0$:

$$a_t \leq \theta^t \cdot \left(a_0 + \frac{c_0}{\theta - \rho}\right).$$

Proof of Lemma H.1. The inequality holds trivially for $t = 0$. Iterating the recursion and using $b_s \le \theta^s c_0$, for $t \ge 0$ we have

$$\begin{aligned} a_{t+1} &\le \rho^{t+1} \cdot a_0 + \sum_{s=0}^{t} \rho^{t-s}\theta^s \cdot c_0 \\ &= \rho^{t+1} \cdot a_0 + \frac{\theta^{t+1} - \rho^{t+1}}{\theta - \rho} \cdot c_0 \\ &\le \theta^{t+1} \cdot \left(a_0 + \frac{1}{\theta - \rho} \cdot c_0 \right), \end{aligned}$$

where the last inequality uses $\rho < \theta$. □

H.2 General-Response Error Bounds

We then bound the error induced by the change in the activation pattern and higher-order error terms. Here, higher-order terms refer to those whose norm can be bounded by degree-2 monomials of $\|\mathbf{r}_t\|$, $\|\mathbf{P}_t\|_{\mathrm{F}}$ and $\|\mathbf{Q}_t\|_{\mathrm{F}}$. For GD, we have $\|\mathbf{P}_t\|_{\mathrm{F}} = \eta \|\nabla_V \mathcal{L}(\mathbf{V}_t, \mathbf{W}_t)\|_{\mathrm{F}} = O(\|\mathbf{r}_t\|)$, and similarly $\|\mathbf{Q}_t\|_{\mathrm{F}} = O(\|\mathbf{r}_t\|)$, so these terms obey the induction scale. For NAG, by recursion,

$$\mathbf{P}_t = -\eta \nabla_V \mathcal{L}(\mathbf{V}_t, \mathbf{W}_t) + \eta\beta^{t+1} \nabla_V \mathcal{L}(\mathbf{V}_0, \mathbf{W}_0) - \eta \sum_{s=0}^{t} \beta^{t+1-s} \nabla_V \mathcal{L}(\mathbf{V}_s, \mathbf{W}_s).$$

If the residual is $O(\theta^t)$, the weights remain bounded and $0 \le \beta \le \theta^2 < 1$, the triangle inequality gives $\|\mathbf{P}_t\|_{\mathrm{F}} = O(\theta^t)$. Similarly we have $\|\mathbf{Q}_t\|_{\mathrm{F}} = O(\theta^t)$, and thus higher-order terms can be controlled in $O(\theta^{2t})$.

For the change in activation pattern, we first call the following result from Wang et al. [2021].

Lemma H.2 (Lemma 11 in Wang et al. [2021])**.** *Suppose $\mathbf{W}_0$ is as* (4) *and $R_2 \in (0, c_2)$. Then with probability at least $1 - N \cdot \exp(-mR_2/c_2)$, the following statement holds uniformly over time intervals: for any collection of iterates $\{\mathbf{W}_t\}_{t \in \mathcal{T}}$ satisfying $\|[\mathbf{W}_t]_{:,i} - [\mathbf{W}_0]_{:,i}\| \le R_2$ for all $t \in \mathcal{T}$ and $i \in [2m]$, if $S_j^{\perp}$ denotes the neurons whose activation on $\mathbf{x}_j$ differs from initialization for some $t \in \mathcal{T}$, then*

$$|S_j^{\perp}| \le 8mR_2/c_2, \ \forall j \in [N].$$

For completeness, we provide a proof in Appendix H.3. Lemma H.2 relates the pattern change to the hidden layer weight change. It is an initialization event over the radius balls around $\mathbf{W}_0$: in induction proofs it may be applied on any interval on which the hidden weights remain within radius R_2, without an additional union bound over time. The residual-dynamics error terms caused by activation changes are bounded in terms of a uniform sign-flip count $\bar{S}$. In the following lemma, $\bar{S}$ bounds the number of possibly unstable neurons per sample throughout the radius ball; by the proof of Lemma H.2, one may take $\bar{S} = 8mR_2/c_2$. Wrapping up these terms, we have the following bound on the error $\|\boldsymbol{\xi}_t\|$.

Lemma H.3. *Suppose $0 < \theta < 1$, $0 \le \beta \le \theta^2$, and $R_V^2(R_2/c_2) \ge R_1(2R_V + R_1)$. If for all $0 \le s \le t$ and $i \in [2m]$, $\|\mathbf{r}_s\| \le C_1\theta^s \|\mathbf{Y}\|_{\mathrm{F}}$, $\|[\mathbf{V}_s]_{:,i} - [\mathbf{V}_0]_{:,i}\| \le R_1$, $\|[\mathbf{W}_s]_{:,i} - [\mathbf{W}_0]_{:,i}\| \le R_2$ and the kernel bounds in Lemma F.2 hold, with the parameter-radius bounds also valid at step $t+1$, then*

$$\begin{aligned} \|\boldsymbol{\xi}_t\| \le{}& \frac{2\theta^t}{1-\theta} \eta C_1 \bar{S} C_V^2 \|\mathbf{Y}\|_{\mathrm{F}} + 3\eta\bar{S}N \left(R_V + R_1\right)^2 C_1 \theta^t \|\mathbf{Y}\|_{\mathrm{F}} \\ &+ \frac{8\theta^{2t}}{(1-\theta)^2} \eta^2 C_1^2 m C_V C_W \|\mathbf{X}\|_{\mathrm{F}} \|\mathbf{Y}\|_{\mathrm{F}}^2 + 6\left(5N\frac{R_2}{c_2}R_V^2 + C_W^2\right) m\eta C_1 \theta^t \|\mathbf{Y}\|_{\mathrm{F}}, \end{aligned}$$

where $C_V \coloneqq \|\mathbf{X}\|_{\mathrm{F}} (R_V + R_1)$, $C_W \coloneqq \|\mathbf{X}\|_{\mathrm{F}} (R_W + R_2)$. The four summands correspond respectively to the $\bar{S}$ activation-change term, the linear $\bar{S}$ sign-flip term, the quadratic NAG-error term, and the sharpened kernel-shift/CK term.

Moreover, the weights at the next step remain close to their initialization under compatible radius conditions. The proofs of Lemmas H.3 and H.4 are given below in Appendices H.4 and H.5.

Lemma H.4. *Assume $0 < \theta < 1$, $0 \le \beta \le \theta^2$, and the residual and parameter-radius bounds through step t in Lemma H.3, without assuming the radii at step $t+1$. If*

$$\eta C_1 C_W \|\mathbf{Y}\|_{\mathrm{F}} \le (1-\theta)^2 R_1,\ \eta C_1 C_V \|\mathbf{Y}\|_{\mathrm{F}} \le (1-\theta)^2 R_2,$$

then for all $i \in [2m]$, $\|[\mathbf{V}_{t+1}]_{:,i} - [\mathbf{V}_0]_{:,i}\| \le R_1$, and $\|[\mathbf{W}_{t+1}]_{:,i} - [\mathbf{W}_0]_{:,i}\| \le R_2$. The cross-layer pairing is intentional: the $\mathbf{V}$-update depends on the hidden-layer features and therefore uses C_W, while the $\mathbf{W}$-update depends on the output weights and therefore uses C_V.

Remark H.5. While initialization only requires $\widetilde{\Omega}(\|\mathbf{X}\|^2 n/\lambda)$ width, our full convergence proof requires $\widetilde{\Omega}(Nn/\lambda)$ to control gate changes and weight movement along the trajectory (Appendix H).

H.3 Proof of Lemma H.2

Proof of Lemma H.2. The proof follows the proof of Claim 3.12 in [Song and Yang, 2019]. Recall

$$S_j^{\perp} = \left\{ i \in [2m] \mid \exists t \in \mathcal{T},\ \mathbb{1}[\mathbf{W}_t^\top \mathbf{X} \ge 0]_{i,j} \ne \mathbb{1}[\mathbf{W}_0^\top \mathbf{X} \ge 0]_{i,j} \right\}.$$

Given the condition that $\|[\mathbf{W}_t]_{:,i} - [\mathbf{W}_0]_{:,i}\| \le R_2$ for all $t \in \mathcal{T}$, we can relax the constraint in $S_j^{\perp}$ and consider the following random variable s_j for each $j \in [N]$:

$$s_j \coloneqq \sum_{i=1}^{2m} \mathbb{1}\left[\exists \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w} - [\mathbf{W}_0]_{:,i}\| \le R_2, (\mathbf{w}^\top \mathbf{x}_j) \cdot ([\mathbf{W}_0]_{:,i}^\top \mathbf{x}_j) \le 0\right].$$

By definition, we always have $\left|S_j^{\perp}\right| \le s_j$, so we only need to bound s_j with high probability. Denote event $A_{i,j} \coloneqq \left\{\exists \mathbf{w} \in \mathbb{R}^d, \|\mathbf{w} - [\mathbf{W}_0]_{:,i}\| \le R_2, (\mathbf{w}^\top \mathbf{x}_j) \cdot ([\mathbf{W}_0]_{:,i}^\top \mathbf{x}_j) \le 0\right\}$. By anti-concentration inequality of Gaussian (Lemma A.4 in Song and Yang 2019),

$$\mathbb{P}(A_{i,j}) = \mathbb{P}_{z \sim \mathcal{N}(0, c_2^2)}(|z| \le R_2) \le \frac{2R_2}{\sqrt{2\pi} c_2} < \frac{R_2}{c_2},$$

hence $\mathbb{E}[\mathbb{1}[A_{i,j}]] \le \frac{R_2}{c_2}$. We can also bound the absolute value $|\mathbb{1}[A_{i,j}]| \le 1$ and variance

$$\mathbb{E}[(\mathbb{1}[A_{i,j}] - \mathbb{E}[\mathbb{1}[A_{i,j}]])^2] = \mathbb{E}[\mathbb{1}[A_{i,j}]^2] - (\mathbb{E}[\mathbb{1}[A_{i,j}]])^2 \le \mathbb{E}[\mathbb{1}[A_{i,j}]] \le \frac{R_2}{c_2}.$$

Since $\{A_{i,j}\}_{i=1}^m$ are independent, we can apply Bernstein's inequality and get that for all $\Delta > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{m} \mathbb{1}[A_{i,j}] > m \cdot \frac{R_2}{c_2} + \Delta\right) \le \exp\left(-\frac{\Delta^2/2}{m\frac{R_2}{c_2} + \Delta/3}\right).$$

By setting $\Delta = \frac{3mR_2}{c_2}$ and using the fact that $A_{i,j} = A_{m+i,j}$ for $i \in [m]$ due to the coupled initialization, we have $s_j = 2\sum_{i=1}^m \mathbb{1}[A_{i,j}]$ and thus

$$\mathbb{P}\left(s_j > \frac{8mR_2}{c_2}\right) \le \exp\left(-\frac{9mR_2}{4c_2}\right) < \exp\left(-mR_2/c_2\right).$$

Taking a union bound over $j \in [N]$ proves Lemma H.2. □

H.4 Proof of Lemma H.3

Proof of Lemma H.3. The proof has three parts: first we bound the update increments $\mathbf{P}_t$ and $\mathbf{Q}_t$ from the residual induction hypothesis; next we bound the ReLU perturbation $\boldsymbol{\zeta}_t$; finally we bound the kernel-mismatch term $\boldsymbol{\iota}_t$ using the kernel-shift inequalities. *Step 1: update increments.* By assumptions and the 1-Lipschitzness of $\sigma(\cdot)$, we have

$$\left\|[\sigma(\mathbf{X}^\top \mathbf{W}_s)]_{:,i}\right\| \le \|\mathbf{X}\|_{\mathrm{F}} \left(\|[\mathbf{W}_0]_{:,i}\| + \|[\mathbf{W}_s - \mathbf{W}_0]_{:,i}\|\right) \le \|\mathbf{X}\|_{\mathrm{F}} \left(\|[\mathbf{W}_0]_{:,i}\| + R_2\right) \le C_W.$$

By recursion, we have

$$\begin{aligned}
\mathbf{P}_t &= \beta \mathbf{P}_{t-1} - (1+\beta)\eta \mathbf{R}_t \sigma(\mathbf{X}^\top \mathbf{W}_t) + \beta\eta \mathbf{R}_{t-1} \sigma(\mathbf{X}^\top \mathbf{W}_{t-1}) \\
&= -\eta \mathbf{R}_t \sigma(\mathbf{X}^\top \mathbf{W}_t) - \beta\eta \sum_{s=0}^{t} \beta^{t-s} \mathbf{R}_s \sigma(\mathbf{X}^\top \mathbf{W}_s) + \eta\beta^{t+1} \mathbf{R}_0 \sigma(\mathbf{X}^\top \mathbf{W}_0),
\end{aligned}$$

thus

$$\begin{aligned}\|[\mathbf{P}_t]_{:,i}\| &\le \left(\theta^t + \beta^{t+1} + \beta\sum_{s=0}^{t}\beta^{t-s}\theta^s\right)\eta C_1 \|\mathbf{R}_0\|_{\mathrm{F}} C_W \\ &\le \frac{\theta^t}{1-\theta}\eta C_1 \|\mathbf{R}_0\|_{\mathrm{F}} C_W,\end{aligned}$$

and

$$\|\mathbf{P}_t\|_{\mathrm{F}} \le \frac{\theta^t}{1-\theta}\eta C_1\sqrt{2m}\,\|\mathbf{R}_0\|_{\mathrm{F}} C_W,$$

where we use $C_1 \ge 1$ and $\beta \le \theta^2 < \theta < 1$. Similarly, we have

$$\begin{aligned}\mathbf{Q}_t &= -\eta\nabla_W\mathcal{L}(\mathbf{V}_t,\mathbf{W}_t) - \beta\eta\sum_{s=0}^{t}\beta^{t-s}\nabla_W\mathcal{L}(\mathbf{V}_s,\mathbf{W}_s) + \eta\beta^{t+1}\nabla_W\mathcal{L}(\mathbf{V}_0,\mathbf{W}_0) \\ &= -\eta\mathbf{X}((\mathbf{R}_t^\top\mathbf{V}_t)\odot\mathbb{1}[\mathbf{X}^\top\mathbf{W}_t\ge 0]) + \eta\beta^{t+1}\mathbf{X}((\mathbf{R}_0^\top\mathbf{V}_0)\odot\mathbb{1}[\mathbf{X}^\top\mathbf{W}_0\ge 0]) \\ &\quad - \beta\eta\sum_{s=0}^{t}\beta^{t-s}\mathbf{X}((\mathbf{R}_s^\top\mathbf{V}_s)\odot\mathbb{1}[\mathbf{X}^\top\mathbf{W}_s\ge 0]).\end{aligned}$$

We can bound the norm of each column of $\mathbf{Q}_t$ by

$$\begin{aligned}\|[\mathbf{Q}_t]_{:,i}\| &\le \eta\|\mathbf{X}\|_{\mathrm{F}}\left(\left\|[\mathbf{R}_t^\top\mathbf{V}_t]_{:,i}\right\| + \beta\sum_{s=0}^{t}\beta^{t-s}\left\|[\mathbf{R}_s^\top\mathbf{V}_s]_{:,i}\right\| + \beta^{t+1}\left\|[\mathbf{R}_0^\top\mathbf{V}_0]_{:,i}\right\|\right) \\ &\le \frac{\theta^t}{1-\theta}\eta C_1\|\mathbf{R}_0\|_{\mathrm{F}} C_V,\end{aligned}$$

and bound the overall Frobenius norm by

$$\|\mathbf{Q}_t\|_{\mathrm{F}} \le \frac{\theta^t}{1-\theta}\eta C_1\sqrt{2m}\,\|\mathbf{R}_0\|_{\mathrm{F}} C_V.$$

Step 2: ReLU perturbation $\boldsymbol{\zeta}_t$. We begin with $\|\boldsymbol{\zeta}_t\|$.

$$\begin{aligned}\|\boldsymbol{\zeta}_t\| \le\ & \underbrace{\left\|\beta\eta\mathbf{R}_{t-1}\sigma(\mathbf{X}^\top\mathbf{W}_{t-1})(\sigma(\mathbf{W}_t^\top\mathbf{X}) - \sigma(\mathbf{W}_{t-1}^\top\mathbf{X}))\right\|_{\mathrm{F}}}_{\ell_1} \\ &+ \underbrace{\left\|\beta\eta\mathbf{P}_{t-1}\left(\left(((\mathbf{R}_{t-1}^\top\mathbf{V}_{t-1})\odot\mathbb{1}[\mathbf{X}^\top\mathbf{W}_{t-1}\ge 0])^\top\mathbf{X}^\top\mathbf{X}\right)\odot\mathbb{1}[\mathbf{W}_{t-1}^\top\mathbf{X}\ge 0]\odot\mathbf{M}\right)\right\|_{\mathrm{F}}}_{\ell_2} \\ &+ \underbrace{\|T_5 - \beta T_6 + (1+\beta)\eta T_9 - \beta\eta T_{10}\|_{\mathrm{F}}}_{\ell_3} \\ &+ \underbrace{\left\|\beta\mathbf{P}_{t-1}\left(\left(\mathbf{Q}_{t-1}^\top\mathbf{X}\right)\odot\mathbb{1}[\mathbf{W}_0^\top\mathbf{X}\ge 0]\odot\mathbf{M}\right) + T_3\right\|_{\mathrm{F}}}_{\ell_4},\end{aligned}$$

where T_3, T_5, T_6, T_9 and T_{10} are defined in the proof of Proposition E.1 in Appendix E.2. We provide upper bounds for $\|T_3\|_{\mathrm{F}}$, $\|T_5\|_{\mathrm{F}}$, and $\|T_9\|_{\mathrm{F}}$; the bounds for $\|T_6\|_{\mathrm{F}}$ and $\|T_{10}\|_{\mathrm{F}}$ follow by replacing t with $t-1$. By 1-Lipschitzness of ReLU activation $\sigma(\cdot)$,

$$\|T_3\|_{\mathrm{F}} \le \|\mathbf{P}_t\|_{\mathrm{F}}\left\|\sigma(\mathbf{W}_{t+1}^\top\mathbf{X}) - \sigma(\mathbf{W}_t^\top\mathbf{X})\right\|_{\mathrm{F}} \le \|\mathbf{P}_t\|_{\mathrm{F}}\left\|\mathbf{Q}_t^\top\mathbf{X}\right\|_{\mathrm{F}}.$$

For T_5, use the maximum column norm directly. For each $j\in[N]$,

$$\begin{aligned}\|[T_5]_{:,j}\| &\le \sum_{i\in S_j^\perp}\|[\mathbf{V}_t]_{:,i}\|\left|\sigma([\mathbf{W}_{t+1}]_{:,i}^\top\mathbf{x}_j) - \sigma([\mathbf{W}_t]_{:,i}^\top\mathbf{x}_j)\right| \\ &\le \overline{S}(R_V + R_1)\max_i\|[\mathbf{Q}_t]_{:,i}\|\,\|\mathbf{x}_j\|.\end{aligned}$$

Squaring and summing over j gives $\|T_5\|_{\mathrm{F}} \le \overline{S}(R_V + R_1)\|\mathbf{X}\|_{\mathrm{F}} \max_i \|[\mathbf{Q}_t]_{:,i}\|$. For T_9, we have

$$\begin{aligned}
\|T_9\|_{\mathrm{F}}^2 &= \sum_{l=1}^{n}\sum_{j=1}^{N}\left(\sum_{i\in S_j^\perp}[\mathbf{A}_{i,t}\mathbf{R}_t\mathbf{B}_{i,t}]_{l,j}\right)^2 \\
&\le \sum_{l=1}^{n}\sum_{j=1}^{N}\left|S_j^\perp\right|\sum_{i\in S_j^\perp}[\mathbf{A}_{i,t}\mathbf{R}_t\mathbf{B}_{i,t}]_{l,j}^2 \\
&= \sum_{j=1}^{N}\left|S_j^\perp\right|\sum_{i\in S_j^\perp}\|[\mathbf{A}_{i,t}\mathbf{R}_t\mathbf{B}_{i,t}]_{:,j}\|^2 \\
&\le \sum_{j=1}^{N}\left|S_j^\perp\right|\sum_{i\in S_j^\perp}\|\mathbf{A}_{i,t}\|^2\|\mathbf{R}_t\|^2\|[\mathbf{B}_{i,t}]_{:,j}\|^2 \\
&\le \|\mathbf{R}_t\|^2 \max_{1\le i\le 2m}\|\mathbf{A}_{i,t}\|^2 \sum_{j=1}^{N}\sum_{k=1}^{N}\left|S_j^\perp\right|\sum_{i\in S_j^\perp}[\mathbf{B}_{i,t}]_{k,j}^2 \\
&\le \|\mathbf{R}_t\|^2 \max_{1\le i\le 2m}\|\mathbf{A}_{i,t}\|^2 \overline{S}^2N^2,
\end{aligned}$$

where the second line is by convexity of quadratic function, and the fifth line is by $|[\mathbf{B}_{i,t}]_{k,j}| \le 1$. We now bound terms ℓ_1, ℓ_2, ℓ_3 and ℓ_4. For ℓ_1, we have

$$\begin{aligned}
\ell_1 &= \beta\eta\left\|\mathbf{R}_{t-1}\sigma(\mathbf{X}^\top\mathbf{W}_{t-1})(\sigma(\mathbf{W}_t^\top\mathbf{X}) - \sigma(\mathbf{W}_{t-1}^\top\mathbf{X}))\right\|_{\mathrm{F}} \\
&\le \beta\eta\|\mathbf{R}_{t-1}\|_{\mathrm{F}}\left\|\sigma(\mathbf{X}^\top\mathbf{W}_{t-1})\right\|_{\mathrm{F}}\left\|\sigma(\mathbf{W}_t^\top\mathbf{X}) - \sigma(\mathbf{W}_{t-1}^\top\mathbf{X})\right\|_{\mathrm{F}} \\
&\le \beta\eta\|\mathbf{R}_{t-1}\|_{\mathrm{F}}\left\|\mathbf{X}^\top\mathbf{W}_{t-1}\right\|_{\mathrm{F}}\left\|\mathbf{Q}_{t-1}^\top\mathbf{X}\right\|_{\mathrm{F}} \\
&\le \frac{2\theta^{2t}}{1-\theta}\eta^2C_1^2mC_VC_W\|\mathbf{X}\|_{\mathrm{F}}\|\mathbf{R}_0\|_{\mathrm{F}}^2,
\end{aligned}$$

where the second inequality uses 1-Lipschitzness of ReLU. For ℓ_2, we have

$$\begin{aligned}
\ell_2 &= \beta\eta\left\|\mathbf{P}_{t-1}\left(\left(((\mathbf{R}_{t-1}^\top\mathbf{V}_{t-1})\odot\mathbb{1}[\mathbf{X}^\top\mathbf{W}_{t-1}\ge 0]\right)^\top\mathbf{X}^\top\mathbf{X}\right)\odot\mathbb{1}[\mathbf{W}_{t-1}^\top\mathbf{X}\ge 0]\odot\mathbf{M}\right)\right\|_{\mathrm{F}} \\
&\le \beta\eta\|\mathbf{P}_{t-1}\|_{\mathrm{F}}\|\mathbf{R}_{t-1}\|_{\mathrm{F}}\|\mathbf{V}_{t-1}\|_{\mathrm{F}}\left\|\mathbf{X}^\top\mathbf{X}\right\| \\
&\le \frac{2\theta^{2t}}{1-\theta}\eta^2C_1^2mC_VC_W\|\mathbf{X}\|_{\mathrm{F}}\|\mathbf{R}_0\|_{\mathrm{F}}^2.
\end{aligned}$$

For ℓ_3, we have

$$\begin{aligned}
\ell_3 &\le \|T_5\|_{\mathrm{F}} + \beta\|T_6\|_{\mathrm{F}} + (1+\beta)\eta\|T_9\|_{\mathrm{F}} + \beta\eta\|T_{10}\|_{\mathrm{F}} \\
&\le \|T_5\|_{\mathrm{F}} + \theta^2\|T_6\|_{\mathrm{F}} + \eta\overline{S}N(R_V+R_1)^2C_1\left(1+\theta+\theta^2\right)\theta^t\|\mathbf{R}_0\|_{\mathrm{F}} \\
&\le \frac{2\theta^t}{1-\theta}\eta C_1\overline{S}C_V^2\|\mathbf{R}_0\|_{\mathrm{F}} + \eta\overline{S}N(R_V+R_1)^2C_1\left(1+\theta+\theta^2\right)\theta^t\|\mathbf{R}_0\|_{\mathrm{F}}
\end{aligned}$$

where we use the bounds for T_5, T_9, and T_{10} from Proposition E.1. For ℓ_4, we have

$$\begin{aligned}
\ell_4 &= \left\|\beta\mathbf{P}_{t-1}\left(\left(\mathbf{Q}_{t-1}^\top\mathbf{X}\right)\odot\mathbb{1}[\mathbf{W}_0^\top\mathbf{X}\ge 0]\odot\mathbf{M}\right) + \mathbf{P}_t\left(\sigma(\mathbf{W}_{t+1}^\top\mathbf{X}) - \sigma(\mathbf{W}_t^\top\mathbf{X})\right)\right\|_{\mathrm{F}} \\
&\le \beta\|\mathbf{P}_{t-1}\|_{\mathrm{F}}\left\|\mathbf{Q}_{t-1}^\top\mathbf{X}\right\|_{\mathrm{F}} + \|\mathbf{P}_t\|_{\mathrm{F}}\left\|\mathbf{Q}_t^\top\mathbf{X}\right\|_{\mathrm{F}} \\
&\le \frac{4\theta^{2t}}{(1-\theta)^2}\eta^2C_1^2mC_VC_W\|\mathbf{X}\|_{\mathrm{F}}\|\mathbf{R}_0\|_{\mathrm{F}}^2.
\end{aligned}$$

Combining these bounds gives

$$\begin{aligned}
\|\boldsymbol{\zeta}_t\| \le{}& \frac{\theta^{2t}(8-4\theta)}{(1-\theta)^2}\eta^2C_1^2mC_VC_W\|\mathbf{X}\|_{\mathrm{F}}\|\mathbf{R}_0\|_{\mathrm{F}}^2 \\
&+ \frac{2\theta^t}{1-\theta}\eta C_1\overline{S}C_V^2\|\mathbf{R}_0\|_{\mathrm{F}} + \eta\overline{S}N(R_V+R_1)^2C_1\left(1+\theta+\theta^2\right)\theta^t\|\mathbf{R}_0\|_{\mathrm{F}}.
\end{aligned}$$

Step 3: kernel-mismatch term $\boldsymbol{\iota}_t$. Now we turn to $\|\boldsymbol{\iota}_t\|$. By Lemma F.2, we have

$$\|\mathbf{K}_0 - \mathbf{K}_t\| \le 10mN\frac{R_2}{c_2}R_V^2, \quad \|\mathbf{G}_t\| \le 2mC_W^2.$$

Therefore,

$$\begin{aligned}\|\boldsymbol{\iota}_t\| &\le (1+\beta)\eta\,\|(\mathbf{K}_0 - \mathbf{H}_t)\mathbf{r}_t\| + \beta\eta\,\|(\mathbf{K}_0 - \mathbf{H}_{t-1})\mathbf{r}_{t-1}\| \\ &\le (1+\beta)\eta\,(\|\mathbf{K}_0 - \mathbf{K}_t\| + \|\mathbf{G}_t\|)\,\|\mathbf{R}_t\|_{\mathrm{F}} + \beta\eta\,(\|\mathbf{K}_0 - \mathbf{K}_{t-1}\| + \|\mathbf{G}_{t-1}\|)\,\|\mathbf{R}_{t-1}\|_{\mathrm{F}} \\ &\le 6\left(5N\frac{R_2}{c_2}R_V^2 + C_W^2\right)m\eta C_1\theta^t\,\|\mathbf{R}_0\|_{\mathrm{F}}\,.\end{aligned}$$

Combining the bounds for $\|\boldsymbol{\zeta}_t\|$ and $\|\boldsymbol{\iota}_t\|$ and using $\theta < 1$ gives the final result. □

H.5 Proof of Lemma H.4

Proof of Lemma H.4. The update-increment estimates in Step 1 of Appendix H.4 use only iterates through step t. They give $\|[\mathbf{P}_s]_{:,i}\| \le \frac{\theta^s}{1-\theta}\eta C_1 C_W\,\|\mathbf{Y}\|_{\mathrm{F}}$. When $\eta C_1 C_W\,\|\mathbf{Y}\|_{\mathrm{F}} \le (1-\theta)^2 R_1$, for any $i \in [2m]$ we have

$$\begin{aligned}\|[\mathbf{V}_{t+1}]_{:,i} - [\mathbf{V}_0]_{:,i}\| &\le \sum_{s=0}^{t}\|[\mathbf{P}_s]_{:,i}\| \le \frac{\eta C_1\,\|\mathbf{Y}\|_{\mathrm{F}}\,C_W}{1-\theta}\sum_{s=0}^{t}\theta^s \\ &\le \frac{1}{(1-\theta)^2}\eta C_1\,\|\mathbf{Y}\|_{\mathrm{F}}\,C_W \le R_1.\end{aligned}$$

Similarly, when $\eta C_1 C_V\,\|\mathbf{Y}\|_{\mathrm{F}} \le (1-\theta)^2 R_2$, we obtain

$$\|[\mathbf{W}_{t+1}]_{:,i} - [\mathbf{W}_0]_{:,i}\| \le R_2, \qquad\qquad i \in [2m].$$

□

H.6 Proof of Theorem 3.2

We first record two estimates on the training Jacobian that will also be used in the NAG proofs. For general responses, set $(k,a,b) = (n,1,1)$; for sketched responses, set $(k,a,b) = (r, \|\mathbf{Y}\|, \sigma_r(\mathbf{Y}))$ and use orthonormal coordinates on $\mathrm{col}(\mathbf{Y})$. In either case there are Nk residual coordinates; the reduction in the sketched case is justified in Appendix H.8. Write $\mathcal{L}_t = \mathcal{L}(\mathbf{V}_t, \mathbf{W}_t) = \|\mathbf{r}_t\|^2/2$ and $B = \|\mathbf{Y}\|_{\mathrm{F}}$.

Write $\chi = a/b$, $p = c_1c_2$, $q = c_1/c_2$ and $\ell = \log(e + 128mN^2n^2/\delta)$. We shall use the following sufficient conditions, with a sufficiently large absolute constant C:

$$m \ge C\frac{Nk\chi^2}{\lambda}\ell^4, \qquad q \ge C\frac{\sqrt{Nd}}{b\sqrt{\lambda}}, \qquad p \ge C\frac{a\sqrt{Nk}}{\lambda b^2}. \tag{17}$$

The first inequality follows from the width assumption by the log self-bounding convention in Appendix A, with $D = Nk\chi^2/\lambda$. Set

$$h = \frac{\sqrt{N}\,\ell^2}{m}, \qquad R_2 = c_2 h, \qquad R_1 = c_1 b\sqrt{k}\,h. \tag{18}$$

In particular, $h < 1$. We shall work on the columnwise radius ball of radii R_1, R_2 around $(\mathbf{V}_0, \mathbf{W}_0)$.

Let L and μ be the extreme positive eigenvalues of $\mathbf{K}_0$. By Lemmas D.1 and D.2 or Corollary 4.4 and lemmas 4.5 and 4.6, the following bounds hold:

$$\begin{aligned} R_V &\lesssim c_1 a\sqrt{k+\ell}, & R_W &\lesssim c_2\sqrt{d+\ell}, & \|\mathbf{W}_0\|_{\mathrm{F}} &\lesssim c_2\sqrt{md}, \\ \mu &\gtrsim mc_1^2b^2\lambda, & L &\lesssim mc_1^2a^2k\lambda_1(\overline{\mathbf{K}}). \end{aligned}$$

Here the column-radius event costs at most $\delta/8$ and the spectrum event at most $\delta/4$. For the Frobenius bound, $\|\mathbf{W}_0\|_{\mathrm{F}}^2/(2c_2^2) \sim \chi^2(md)$, so Proposition C.1 gives it at cost at most $\delta/4$ under (17). We use one further event to control changes of activation over the entire ball.

Lemma H.6 (Joint activation margins)**.** *For $i \in [m]$, write $\boldsymbol{\psi}_i = [\mathbf{W}_0]_{:,i}/c_2$ and $A_{ij} = \mathbb{1}[|\boldsymbol{\psi}_i^\top \mathbf{x}_j| \le h]$. With probability at least $1 - \delta/4$,*

$$\max_{j\in[N]} \sum_{l=1}^{N} \sum_{i=1}^{m} A_{ij} A_{il} \lesssim m(h + Nh^2/\lambda) + N\ell. \tag{19}$$

Proof. For two distinct samples at angle α, a Rayleigh quotient of $\overline{\mathbf{K}}$ on $(\mathbf{e}_j - \mathbf{e}_l)/\sqrt{2}$ gives $\lambda \le 1/2 - \cos(\alpha)(\pi - \alpha)/(2\pi) \le \alpha$, using the Gaussian halfspace probability and $1 - \cos\alpha \le \alpha$. For each j, at most one other sample can lie within angular distance $\lambda/3$ of $-\mathbf{x}_j$: otherwise two such samples would be less than λ apart. Apart from this exception and j itself, $\sin\alpha \gtrsim \lambda$. The joint density of the two Gaussian projections is bounded by $(2\pi \sin\alpha)^{-1}$, so $\mathbb{E}[A_{ij}A_{il}] \lesssim h^2/\lambda$ for these pairs. For the exceptions use the single-margin bound $\mathbb{E}A_{ij} \lesssim h$. This also covers exact antipodes and $N = 1$. For fixed j, l the products are independent across the first m neurons. The exponential Markov inequality at parameter one gives $\sum_i A_{ij}A_{il} \le (e-1)m\mathbb{E}[A_{1j}A_{1l}] + \log(4N^2/\delta)$ except with probability $\delta/(4N^2)$. Union over the pairs and sum along each row. □

Write $\mathcal{J}(\vartheta)$ for the Jacobian of the vectorized output in the effective coordinates, with ϑ the stacked parameters $(\mathbf{V}, \mathbf{W})$ and the gate $\mathbb{1}[z \ge 0]$ at zero. Let $\mathcal{J}_* = [0, \mathcal{J}_{W,0}]$ retain only the hidden-weight block at initialization, so $\mathcal{J}_*\mathcal{J}_*^\top = \mathbf{K}_0$ on the effective space.

Lemma H.7 (Uniform Jacobian bound)**.** *On the intersection of the above events, throughout the ball* (18)*, $\|\mathcal{J}(\vartheta) - \mathcal{J}_*\| \le \gamma\sqrt{\mu}$ for any fixed sufficiently small absolute $\gamma > 0$, by choosing the constant in* (17) *large enough. The same bound holds for the averaged Jacobian on any segment in the ball.*

Proof. First change only the gates of the hidden-weight block, leaving the output weights at their initial values. The absolute change of gate at (i, j) is at most A_{ij} for either neuron in pair i. For a test vector with blocks $\mathbf{u}_j \in \mathbb{R}^k$, expansion of the squared norm of the transposed difference acting on this vector gives the upper bound

$$2R_V^2 \sum_{j,l} \Big(\sum_{i\le m} A_{ij}A_{il} \Big) \|\mathbf{u}_j\| \, \|\mathbf{u}_l\| \,,$$

since $|\mathbf{x}_j^\top \mathbf{x}_l| \le 1$. The factor 2 counts the two neurons; their later gates need not agree. The coefficient matrix is symmetric and nonnegative entrywise, so its norm is bounded by its maximum row sum. By (19), the squared operator norm of this change divided by μ is at most a constant times

$$\chi^2(k + \ell) \left(\frac{h}{\lambda} + \frac{Nh^2}{\lambda^2} + \frac{N\ell}{m\lambda} \right).$$

Substituting (17) and (18) and $k + \ell \le 2k\ell$, the three terms are $O(1/(C\sqrt{N}\ell))$, $O(1/(C^2 k\chi^2\ell^3))$, and $O(1/(C\ell^2))$, respectively.

Next change the output columns in this hidden-weight block, keeping the current gates fixed. Each neuron contributes a transposed block of norm at most $\|\mathbf{X}\| \, R_1$, so the squared norm divided by μ is at most $2mNR_1^2/\mu \lesssim Nkh^2/\lambda$, also small under (17). Finally the output-weight Jacobian has norm at most $\big\|\sigma(\mathbf{W}^\top \mathbf{X})\big\|_{\mathrm{F}} \le \sqrt{N} \, \|\mathbf{W}\|_{\mathrm{F}} \lesssim c_2\sqrt{mNd}$ throughout the ball. After division by $\sqrt{\mu}$ this is $O(\sqrt{Nd}/(qb\sqrt{\lambda}))$, controlled by the ratio in (17). Summing the three bounds proves the claim. The ball is convex, and output differences on a segment equal its integrated Jacobian applied to the displacement: the output is piecewise polynomial there, and an identically zero preactivation has zero directional change. The integral thus obeys the same bound. □

In particular, on the effective space the full training kernel $\mathbf{H}_t = \mathcal{J}(\vartheta_t)\mathcal{J}(\vartheta_t)^\top$ satisfies

$$\Big\| \mathbf{K}_0^{-1/2} (\mathbf{H}_t - \mathbf{K}_0) \mathbf{K}_0^{-1/2} \Big\| \le 2\gamma + \gamma^2$$

whenever the weights remain in the ball. Indeed, $\Big\| \mathbf{K}_0^{-1/2} \mathcal{J}_* \Big\| = 1$ and the normalized Jacobian difference has norm at most γ by Lemma H.7.

For the weight-radius comparisons below, write $C_V = \sqrt{N}(R_V + R_1)$ and $C_W = \sqrt{N}(R_W + R_2)$ as in Lemma H.4. To ensure that $MBC_V/\mu \le R_2$ and $MBC_W/\mu \le R_1$, for any fixed absolute M, it suffices that

$$mh\,p \gtrsim \frac{Na\sqrt{k+\ell}}{\lambda b^2}, \qquad mh\,pq^2 \gtrsim \frac{N\sqrt{d+\ell}}{\lambda b^3\sqrt{k}}, \tag{20}$$

with constants depending on M. Indeed, substitute (18), the initialization bounds above, and $B \le \sqrt{N}$. The first inequality follows from $mh = \sqrt{N}\,\ell^2$ and the product in (17); the second follows from $pq^2 \gtrsim N^{3/2}\sqrt{k}\,ad/(\lambda^2 b^4)$, since $a/b \ge 1$ and $\lambda \le 1/2$. The events used in these estimates depend only on the standardized initial draws and hold simultaneously for every choice of positive scales satisfying (17).

Proof of Theorem 3.2. Throughout this proof, write $L = \lambda_1(\mathbf{K}_0)$ and $\mu = \lambda_{Nn}(\mathbf{K}_0)$, matching the full-rank spectrum in Theorem 3.2. Let $a_t = \|\mathbf{r}_t\|$, $B = \|\mathbf{Y}\|_{\mathrm{F}}$ and $\theta = \sqrt{1-\mu/(2L)}$. Coupled initialization gives $a_0 = B$; if $B = 0$, there is nothing to prove.

Take $(k, a, b) = (n, 1, 1)$ in (17) and (18). The width in the theorem suffices by the log convention in Appendix A. The ratio and product in (5) give $q = c_1/c_2 \gtrsim \sqrt{Nd/\lambda}$ and $p = c_1 c_2 \gtrsim \sqrt{Nn}/\lambda$, since $\kappa(\mathbf{K}_0) \ge 1$. On the event of Lemma H.7, of probability at least $1-\delta$, the initialization bounds of Lemmas D.1 and D.2 give, with $\ell_0 = \log(32m/\delta)$,

$$\tfrac{1}{2}c_1\sqrt{n} \le R_V \le c_1(\sqrt{n} + \sqrt{2\ell_0}), \qquad \tfrac{1}{2}c_2\sqrt{d} \le R_W \le c_2(\sqrt{d} + \sqrt{2\ell_0}),$$
$$\tfrac{1}{2}c_1^2 m\lambda \le \mu \le L \le 9c_1^2 nm\lambda_1(\overline{\mathbf{K}}).$$

Suppose the weights remain within the radii R_1, R_2 and $a_j \le B\theta^j$ through step t. We first bound the displacement to step $t+1$. The gradient of each output column has norm at most $C_W a_j$, and that of each hidden column at most $C_V a_j$, where $C_W = \sqrt{N}(R_W + R_2)$ and $C_V = \sqrt{N}(R_V + R_1)$. Summing the GD updates gives, for every $i \in [2m]$,

$$\|[\mathbf{V}_{t+1} - \mathbf{V}_0]_{:,i}\| \le \frac{\eta B C_W}{1-\theta} \le \frac{4BC_W}{\mu},$$
$$\|[\mathbf{W}_{t+1} - \mathbf{W}_0]_{:,i}\| \le \frac{\eta B C_V}{1-\theta} \le \frac{4BC_V}{\mu}.$$

Here $\eta = 1/L$ and $1-\theta = \mu/[2L(1+\theta)]$. By (20), these bounds are at most R_1 and R_2, respectively: the hidden-column requirement is $mh\,p \gtrsim N\sqrt{n+\ell}/\lambda$, and the output-column requirement is $mh\,pq^2 \gtrsim N\sqrt{d+\ell}/(\lambda\sqrt{n})$. Both follow from $mh = \sqrt{N}\,\ell^2$ and the above bounds on p, q.

The segment to the next iterate therefore stays in the ball. Write $D_t = \mathcal{J}(\vartheta_t)/\sqrt{L}$, and let C_t be the averaged Jacobian on this segment divided by $\sqrt{L}$. With $\mathbf{v} = D_t^\top \mathbf{r}_t$, the GD update gives $\mathbf{r}_{t+1} = \mathbf{r}_t - C_t\mathbf{v}$. Set $s = \sqrt{\mu/L}$. By Lemma H.7 with $\gamma = 1/1024$, $\|\mathbf{v}\| \ge (1-\gamma)sa_t$, $\|C_t - D_t\| \le 2\gamma s$ and $\|C_t\| \le 1+\gamma$. Since $\mathbf{r}_t^\top D_t\mathbf{v} = \|\mathbf{v}\|^2$, expanding the squared residual gives

$$a_{t+1}^2 \le a_t^2 - \left[2\left(1 - \frac{2\gamma}{1-\gamma}\right) - (1+\gamma)^2\right]\|\mathbf{v}\|^2 \le \left(1 - \frac{\mu}{2L}\right)a_t^2.$$

Thus $a_{t+1} \le B\theta^{t+1}$ and $\mathcal{L}_t/\mathcal{L}_0 \le (1-\mu/(2L))^t$. The convergence result follows by induction, with rate $\theta \le 1 - \frac{\Theta(1)}{n\kappa}$ by Lemma D.2. □

H.7 Proof of Theorem 3.3

For a reference Jacobian $\mathcal{J}_*$ on the output space under consideration, let L and μ be the largest and smallest eigenvalues of $\mathcal{J}_*\mathcal{J}_*^\top$, with $\mu > 0$. Let $s = \sqrt{\mu/L}$ and $\eta = 1/L$, and set $\beta = (1-s)/(1+s)$, $\rho = 1-s$ and $\theta = 1 - s/4$. We combine the residual with the scaled momentum

$$\mathbf{m}_t = \sqrt{L}\big(\vartheta_t - \vartheta_{t-1} + \eta\mathbf{g}_{t-1}\big), \qquad \mathbf{g}_t = \mathcal{J}(\vartheta_t)^\top \mathbf{r}_t, qquad \mathrm{y}_t = (\mathbf{r}_t, \mathbf{m}_t).$$

Write $D = \mathcal{J}_*/\sqrt{L}$, $D_t = \mathcal{J}(\vartheta_t)/\sqrt{L}$, and C_t for the averaged Jacobian from ϑ_t to ϑ_{t+1} divided by $\sqrt{L}$. With $\mathbf{u}_t = \sqrt{L}(\vartheta_{t+1} - \vartheta_t)$, (3) gives

$$\mathbf{u}_t = \beta\mathbf{m}_t - (1+\beta)D_t^\top \mathbf{r}_t, \qquad \mathbf{m}_{t+1} = \beta\mathbf{m}_t - \beta D_t^\top \mathbf{r}_t, \qquad \mathbf{r}_{t+1} = \mathbf{r}_t + C_t\mathbf{u}_t.$$

Replacing C_t, D_t by D defines a fixed transition T for y_t. When both iterates lie in the ball (18), Lemma H.7 gives $\|D_t - D\|, \|C_t - D\| \le \gamma s$. Thus $\mathsf{y}_{t+1} = (T + E_t)\mathsf{y}_t$ with $\|E_t\| \le 8\gamma s$: the nonzero blocks of E_t are $-(1+\beta)(C_t D_t^\top - DD^\top)$, $\beta(C_t - D)$, and $-\beta(D_t - D)^\top$.

Lemma H.8 (Joint-state stability)**.** *For $j \ge 0$, $\left\|T^j\right\| \le (4 + 2js)\rho^j$ (with $0^0 = 1$). Consequently,*

$$\sup_{j\ge 0} \theta^{-j} \left\|T^j\right\| \le 8, \qquad \sum_{j\ge 0} \theta^{-j-1} \left\|T^j\right\| \le 9/s.$$

Proof. Use a singular value decomposition of D. Null parameter directions have multiplier $\beta \le \rho$; each singular value $\sigma \in [s, 1]$, with $x = \sigma^2$, gives the block

$$T_\sigma = \begin{pmatrix} 1 - (1+\beta)x & \beta\sqrt{x} \\ -\beta\sqrt{x} & \beta \end{pmatrix}.$$

For $s < 1$, put $c = \sqrt{(1-x)/(1-s^2)}$, $w = \sqrt{1-c^2}$ and $R = \rho c$. The half trace and determinant are Rc and R^2, respectively, and

$$T_\sigma - Rc\mathbf{I}_2 = \frac{1}{1+s}\begin{pmatrix} s - x & \rho\sqrt{x} \\ -\rho\sqrt{x} & x - s \end{pmatrix}, \qquad \|T_\sigma - Rc\mathbf{I}_2\| \le \rho(2s + 3w),$$

using $x = s^2 + (1-s^2)w^2$. The characteristic equation gives, for $j \ge 1$, $T_\sigma^j = R^j T_j(c)\mathbf{I}_2 + R^{j-1}U_{j-1}(c)(T_\sigma - Rc\mathbf{I}_2)$, where T_j, U_j are Chebyshev polynomials. Their trigonometric representations give $|T_j(c)| \le 1$ and $|U_{j-1}(c)| \le \min(j, 1/w)$. Since $R \le \rho$, the asserted power bound follows, including $c = 0, 1$ by continuity. If $s = 1$, all blocks vanish. For the weighted bounds use $\rho/\theta \le e^{-3s/4}$ and sum the geometric series: $4/(\theta - \rho) + 2s\rho/(\theta - \rho)^2 < 9/s$. □

Proof of Theorem 3.3. We prove the theorem by induction. Throughout this proof, write $L = \lambda_1(\mathbf{K}_0)$ and $\mu = \lambda_{Nn}(\mathbf{K}_0)$, matching the full-rank spectrum in Theorem 3.3. Set $(k, a, b) = (n, 1, 1)$. The width and the two bounds

$$\frac{c_1}{c_2} \gtrsim \frac{\sqrt{Nd}}{\sqrt{\lambda}}, \qquad c_1 c_2 \gtrsim \frac{\sqrt{Nn}}{\lambda}$$

in (6) give (17). We work on the initialization event of Lemma H.7, with $\gamma = 1/1024$; by the NAG ledger in Appendix A, this event has probability at least $1 - \delta$ and holds uniformly over the radius ball.

Take $C_1 = 40$ and write $B = \|\mathbf{Y}\|_{\mathrm{F}}$. At the 0-th step, $\|\mathbf{r}_0\| = B$ and the weight changes are zero. Since $\vartheta_{-1} = \vartheta_0$ and $\mathbf{g}_{-1} = \mathbf{g}_0$, the initial momentum variable is $\mathbf{m}_0 = D_0^\top \mathbf{r}_0$. Thus $\|\mathsf{y}_0\| \le 3B$, using $\|D_0\| \le 1 + \gamma s$.

Now suppose $\|\mathsf{y}_j\| \le C_1 B\theta^j$ and the weight radii hold from step 0 to t. Our choice of η and β satisfies $\beta \le 1 - s \le \theta^2 < \theta < 1$. To extend the induction to step $t+1$, we first verify the two weight-radius requirements in Lemma H.4:

$$16C_1C_W \|\mathbf{Y}\|_{\mathrm{F}} \le \mu R_1, \qquad 16C_1C_V \|\mathbf{Y}\|_{\mathrm{F}} \le \mu R_2,$$

Since $\eta = 1/L$ and $(1-\theta)^2 = 1/(16\kappa(\mathbf{K}_0)) = \mu/(16L)$, these inequalities imply $\eta C_1 C_W \|\mathbf{Y}\|_{\mathrm{F}} \le (1-\theta)^2 R_1$ and $\eta C_1 C_V \|\mathbf{Y}\|_{\mathrm{F}} \le (1-\theta)^2 R_2$. To check them, use (20) with $(k, a, b) = (n, 1, 1)$: the product condition gives $mhp \gtrsim N\sqrt{n+\ell}/\lambda$, and the ratio and product give $mhpq^2 \gtrsim N\sqrt{d+\ell}/(\lambda\sqrt{n})$. Hence the next weights remain in the ball.

We may now apply Lemma H.7 on the segment ending at step $t+1$. Unrolling the recursion for y_{t+1} and using Lemma H.8 gives

$$\theta^{-(t+1)} \|\mathsf{y}_{t+1}\| \le 8 \|\mathsf{y}_0\| + 72\gamma \max_{0\le j\le t} \theta^{-j} \|\mathsf{y}_j\| \le (24 + 72\gamma C_1)B \le C_1 B.$$

Thus the induction conditions for the $t+1$ step hold. In particular, $\mathcal{L}_t \le C_1^2\theta^{2t}\mathcal{L}_0 \le C_1^2 e^{-st/2}\mathcal{L}_0$. By Lemma D.2, $\theta = 1 - \frac{1}{4\sqrt{\kappa(\mathbf{K}_0)}} \le 1 - \frac{\Theta(1)}{\sqrt{n\kappa}}$. This proves the stated iteration bound (rounded up); the $\log(1/\epsilon)$ order is as $\epsilon \downarrow 0$, or one may use $\log(2/\epsilon)$ uniformly for $0 < \epsilon < 1$. □

H.8 Proof of Theorem 4.1

Proof of Theorem 4.1. We prove the theorem by induction on the invariant subspace $\mathcal{E} = \mathbb{R}^N \otimes \mathrm{col}(\mathbf{Y})$. Throughout this proof, write $L = \lambda_1(\mathbf{K}_0)$ and $\mu = \lambda_{Nr}(\mathbf{K}_0)$, matching the nonzero-spectrum quantities in Theorem 4.1. By Lemma 4.7, the residual and error terms stay in $\mathcal{E}$. The columns of $\mathbf{V}_0$ lie in $\mathrm{col}(\mathbf{Y})$, and the output-weight gradient and extrapolation in (3) preserve this space. We can therefore write the output and output weights in orthonormal coordinates on $\mathrm{col}(\mathbf{Y})$ without changing the update or the loss.

Set $k = r$, $a = \|\mathbf{Y}\|$ and $b = \sigma_r(\mathbf{Y}) > 0$, so $\chi = a/b = \kappa(\mathbf{Y})$. In these coordinates, the width and the scale bounds in (9) give

$$m \gtrsim \frac{Nr\chi^2}{\lambda}\ell^4, \qquad q \gtrsim \frac{\sqrt{Nd}}{\sqrt{\lambda}\, b}, \qquad p \gtrsim \frac{\sqrt{Nr}\,\kappa(\mathbf{Y})}{\lambda\sigma_r(\mathbf{Y})} = \frac{a\sqrt{Nr}}{\lambda b^2}.$$

Thus (17) holds, with the logarithms accounted for in Appendix A. When invoking Lemma 4.3 here, we only consume its R_V, R_W bounds summarized in Corollary 4.4; these are the first two bounds from the proof of Lemma 4.3, which do not involve τ. The probability accounting is the low-rank ledger in Appendix A: $\delta/8$ is allocated to this τ-free radius event and $\delta/4$ each to the nonzero-spectrum, Frobenius-norm and joint-margin bounds. On their intersection, Lemma H.7 holds with $\gamma = 1/1024$, with no further union over iterations.

Let $s = \sqrt{\mu/L}$, $\theta = 1 - s/4$, $C_1 = 40$ and $B = \|\mathbf{Y}\|_{\mathrm{F}}$. Form the residual and momentum state $\mathsf{y}_t = (\mathbf{r}_t, \mathbf{m}_t)$ in the reduced coordinates as in Appendix H.7. At the 0-th step, $\|\mathbf{r}_0\| = B$ and $\mathbf{m}_0 = D_0^\top \mathbf{r}_0$, so $\|\mathsf{y}_0\| \le 3B$; both weight changes are zero. Now suppose the weight radii and $\|\mathsf{y}_j\| \le C_1 B\theta^j$ hold from step 0 to t. To extend the induction to step $t+1$, the two weight-radius requirements are again $16C_1C_W B \le \mu R_1$ and $16C_1C_V B \le \mu R_2$, since $\beta \le \theta^2$ and $(1-\theta)^2 = \mu/(16L)$. In the present case (20) reads

$$mh\,p \gtrsim \frac{Na\sqrt{r+\ell}}{\lambda b^2}, \qquad mh\,pq^2 \gtrsim \frac{N\sqrt{d+\ell}}{\lambda b^3\sqrt{r}}.$$

The first follows from $mh = \sqrt{N}\ell^2$ and the displayed bound on p. The second follows from $pq^2 \gtrsim N^{3/2}\sqrt{r}\,ad/(\lambda^2 b^4)$, since $a/b \ge 1$ and $\lambda \le 1/2$. By Lemma H.4 the two radius bounds therefore hold at step $t+1$.

On this segment, the reduced Jacobian differs from its reference by at most $\gamma\sqrt{\mu}$. By Lemma 4.5, the positive spectrum satisfies

$$\tfrac{1}{2}c_1^2 m\lambda\sigma_r^2(\mathbf{Y}) \le \mu \le L \le 9c_1^2 mr\lambda_1(\overline{\mathbf{K}})\,\|\mathbf{Y}\|^2.$$

The normalized reference Jacobian has singular values in $[s, 1]$, so Lemma H.8 applies. Unrolling the reduced state recursion gives $\theta^{-(t+1)}\,\|\mathsf{y}_{t+1}\| \le 8\,\|\mathsf{y}_0\| + 72\gamma C_1 B \le C_1 B$. Thus the induction conditions for the $t+1$ step hold, and $\mathcal{L}_t \le C_1^2\theta^{2t}\mathcal{L}_0$. The claimed convergence rate follows from the induction, with contraction factor θ. By Lemma 4.5, $\theta = 1 - \frac{1}{4\sqrt{\kappa(\mathbf{K}_0)}} \le 1 - \frac{\Theta(1)}{\sqrt{r\kappa}\kappa(\mathbf{Y})}$. □

I Additional CK/NTK Diagnostics

This diagnostic uses the same $N = 100, d = 20, n = 10, m = 1000, T = 100$, ratios $\{0.1, 0.3, 1, 3, 10\}$, fixed $c_1 c_2 = 1$, and seeds $0, \ldots, 4$ as Figure 3. The shift curves report checkpoint maxima over $t \in \{0, 10, \ldots, T\}$ for the total empirical kernel $\max_t \|\mathbf{H}_t - \mathbf{H}_0\|_{\mathrm{F}} / \|\mathbf{H}_0\|_{\mathrm{F}}$ and its hidden-layer NTK and output-layer/CK components $\max_t \|\mathbf{K}_t - \mathbf{K}_0\|_{\mathrm{F}} / \|\mathbf{K}_0\|_{\mathrm{F}}$ and $\max_t \|\mathbf{G}_t - \mathbf{G}_0\|_{\mathrm{F}} / \|\mathbf{G}_0\|_{\mathrm{F}}$. The componentwise denominators $\|\mathbf{K}_0\|_{\mathrm{F}}$ and $\|\mathbf{G}_0\|_{\mathrm{F}}$ vary across the ratio sweep, so these relative shifts are intended as smallness and trend diagnostics rather than absolute cross-ratio magnitude comparisons.

For activation states $S_{i,j,t} = \mathbb{1}\{[\mathbf{W}_t]_{:,i}^{\top} \mathbf{x}_j \geq 0\}$, the flip curves report the average and worst-sample fractions

$$\max_{0 \leq t \leq T} \frac{1}{2mN} \sum_{i=1}^{2m} \sum_{j=1}^{N} \mathbb{1}\{S_{i,j,t} \neq S_{i,j,0}\}, \qquad \max_{0 \leq t \leq T} \max_{j \in [N]} \frac{1}{2m} \sum_{i=1}^{2m} \mathbb{1}\{S_{i,j,t} \neq S_{i,j,0}\}.$$

Both count all $2m$ neurons at every integer step; each run's maximum is taken before averaging over seeds. These are empirical diagnostics, not theorem-condition certificates: the proof controls time-uniform radius-ball events and operator norms, whereas the shift curves are Frobenius checkpoint proxies.

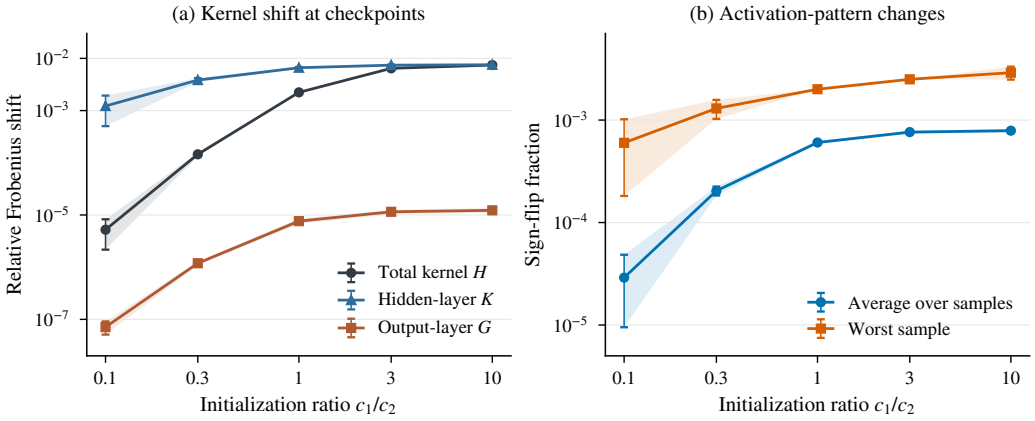


Figure 4: Left: maximum relative Frobenius shifts of $\mathbf{H}$, $\mathbf{K}$, and $\mathbf{G}$ over checkpoints spaced ten steps apart. Right: maximum over all steps of the average and worst-sample activation-change fractions. Points show five-seed means; bands and error bars show ± 1 sample standard deviation.

Across the ratio sweep, the mean checkpoint-max shifts range from $5.22 \cdot 10^{-6}$ to $7.46 \cdot 10^{-3}$ for $\mathbf{H}$, from $1.22 \cdot 10^{-3}$ to $7.55 \cdot 10^{-3}$ for $\mathbf{K}$, and from $7.12 \cdot 10^{-8}$ to $1.23 \cdot 10^{-5}$ for $\mathbf{G}$; the largest mean max-over-time average and worst-sample flip fractions are $7.88 \cdot 10^{-4}$ and $2.90 \cdot 10^{-3}$, respectively.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction state the near-linear-width, acceleration, and low-rank-adaptivity claims, and the formal theorem statements specify the assumptions and scope.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The theorem statements, assumptions, and remarks describe the two-layer ReLU setting, the NTK-dominated regime, exact low-rank requirement, and the approximate-low-rank limitation in Remark 4.2.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The paper states the assumptions for each theorem and provides complete proofs in the appendix, including proof-dependency ledgers and the restricted-subspace argument for the low-rank case.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The numerical experiments are synthetic and the paper specifies the data distributions, dimensions, initialization choices, optimizer parameters, tolerance, and number of independent trials in Section B.

   Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No]

   Justification: We do not provide an anonymized code package at submission time. The experimental setup, including synthetic data generation and hyperparameters, is fully specified in Section B to support reproducibility.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Section B specifies the synthetic data generation, network dimensions, initialization parameters, optimizer step sizes and momentum parameters, width choices, iteration/tolerance criteria, and trial count.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: The convergence-rate plot reports averages over five independent runs with shaded ± 1 standard-deviation bands, and the low-rank experiment reports averages over five independent runs.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: Section B reports the hardware used for the experiments, the approximate GPU runtime per script, and the elapsed time and peak process memory for the CPU diagnostics. Peak GPU memory and total compute across all experiments were not recorded.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work is theoretical and uses synthetic numerical experiments; we are not aware of any aspect that conflicts with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: The paper studies foundational optimization theory for two-layer ReLU networks and does not propose a deployed system, dataset, or application with a direct societal-impact pathway.

    Guidelines:

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release data, models, or systems with high misuse risk; the experiments use synthetic data generated from specified distributions.

    Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [N/A]

Justification: The experiments use synthetic data generated from distributions described in the paper and do not rely on external datasets or proprietary assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [N/A]

Justification: The paper does not introduce or release a new dataset, benchmark, model artifact, or software asset in the current submission.

Guidelines:

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

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

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

Answer: [N/A]

Justification: The research does not involve crowdsourcing, data collection from human participants, or human-subject experiments.

Guidelines:

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

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

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

Answer: [N/A]

Justification: The research does not involve human subjects, so IRB or equivalent human-subject approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: The core methodology and scientific claims do not use LLMs as an important, original, or non-standard component; any editorial assistance would not affect the research method or results.

Guidelines:

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