

The Gradient Independence Approach to GD Convergence

Working Notes — v3

Jeff Edmonds — York University

October 2026

This was written by Claude and not completely read by Jeff

1 Overview

We seek to prove that gradient descent drives training error to zero for a one-hidden-layer network with fixed random f_1 and f_3 , training only the middle weight matrix $\mathbf{W}$.

The key matrix governing convergence is $\mathbf{R} = (\mathbf{G}^T \mathbf{G}) \odot (\mathbf{X}^T \mathbf{X})$, where $\odot$ is the Hadamard (entrywise) product. If $\lambda_{\min}(\mathbf{R}_t)$ stays bounded below along the GD trajectory, convergence follows by standard descent arguments.

We identify a single property of the random function f_3 —*gradient independence*—that guarantees this eigenvalue bound. The chain of logic is:

$$\boxed{\text{Gradient Independence of } f_3} \xrightarrow{\S 4} \boxed{\lambda_{\min}(\mathbf{R}_t) \geq c \alpha_{\min}} \xrightarrow{\S 6, \S 7} \boxed{\mathcal{E}_t \rightarrow 0}$$

Step	Status	Width needed
Independence $\Rightarrow$ eigenvalue bound (static)	PROVED (Hanson–Wright; §4)	$MM' \geq \Omega(D)$
Eigenvalue bound $\Rightarrow$ convergence (fixed floor)	PROVED (standard descent; §6)	(any width)
Static bound $\Rightarrow$ floor holds along the whole trajectory	PROVED, non-circular (§7)	$MM' \gtrsim D^{3/2} \sqrt{\log D}$ (proved); plausibly $\Theta(D)$
Architecture: effective rank	PROVED (§5)	$\min(F, M') \cdot M \geq D$
Architecture: DC-ratio γ bounded away from 1	PROVED, activation-dependent (§5)	activation-dependent
Architecture: decorrelation (Criterion C)	CONJECTURED (§5)	—
Ball conjecture impossibility	PROVED (Edmonds; §8)	cell-based strategies cannot escape lazy regime
Constrained adversary ($M < D$)	PARTIAL (§8.3)	$O(\sqrt{D})$ corrupted pairs; damage $O(\alpha_{\max}/\sqrt{D})$

What is new in v3. Three additions from the October 2026 analysis:

1. **Ball conjecture impossibility (§8).** A clean trichotomy—due to Edmonds—shows that *no* cell-based partition of $\mathbb{R}^{M'}$ (cubes, Voronoi cells, smooth partitions of unity) can prove the ball conjecture

outside the lazy regime ($L \ll \sigma_W$). The argument extends to correlated random fields with any Lipschitz constant. This is a genuine negative result that rules out a natural class of proof strategies.

2. **Constrained adversary at $M < D$ (§8.3).** When $M < D$, the adversary cannot steer each $\Delta \mathbf{z}_d$ independently—it has $MM' < DM'$ degrees of freedom. Combined with eigenvector delocalisation, this limits the damage from corrupted gradient pairs to $O(\alpha_{\max}/\sqrt{D}) \rightarrow 0$. The regime $MM' = \Theta(D)$ with $M < D$ is the most promising for the ball conjecture.
3. **Song et al. (NeurIPS 2021) comparison (§9).** The current frontier for two-layer, all-layer, discrete GD is $\tilde{O}(D^{3/2})$ total parameters (Song, Ramezani-Kebrya, Pethick, Eftekhari, Cevher 2021), not Arora et al.’s D^8 . Our Lipschitz bootstrap at worst-case L_R already gives $D^{3/2}\sqrt{\log D}$ —comparable.

What is new in this version. A detailed review (Sept. 2026) sharpened both halves of the chain above and, importantly, *corrected* the paper’s own diagnosis of why the sigmoid architecture fails. Three additions:

1. **Architecture analysis (§5).** Gradient independence splits into three genuinely separate conditions on the activation law: (A) an *effective-rank* condition $\min(F, M') \cdot M \geq D$; (B) a bounded *DC ratio* γ , identified exactly with the squared degree-1 Hermite coefficient of the activation divided by its total Hermite mass; and (C) *decorrelation* of the centered gradient field, controlled by the activation’s higher Hermite content. **Key correction:** the sigmoid architecture’s shared, collinear gradient direction is *provably harmless* (Lemma 9)—removing it changes $\lambda_{\min}(\mathbf{R})$ by only 2%. Sigmoid fails instead because its DC ratio $\gamma \approx 0.95$ is too close to 1, *and* because of an independent decorrelation failure.
2. **Trajectory bootstrap (§7).** The earlier ball-uniformity reduction is circular and, worse, is unsatisfiable at the paper’s target width. We replace it with a self-consistent, non-circular ODE argument (Theorem 32) that is $3.8\times$ sharper, together with its discrete-GD analogue and a sharp telescoping movement budget. We also show precisely that limiting what an adversary *knows* about f_3 buys nothing (Proposition 36); only limiting how *far* $\mathbf{W}$ moves helps.
3. **Realizability.** Numerical evidence shows $\lambda_{\min}(\mathbf{R}_t)$ is essentially conserved along GD when the target is realizable, and collapses primarily under *unrealizable* targets. Conjecture C4 (§10) is restated under a realizability hypothesis.

Key distinction (incoherence vs. independence). An earlier version used *gradient incoherence* (bounded pairwise magnitudes $|\hat{\mathbf{g}}_d^T \hat{\mathbf{g}}_{d'}| \leq \mu$). Gershgorin applied to incoherence alone gives $MM' \geq \Omega(D^2)$ (§3). Achieving the target $\Omega(D)$ requires *independence*: the off-diagonal entries of $\mathbf{R}$ must have approximately random signs that cancel, not merely small magnitudes.¹

Scope: $M \geq D$ is already handled. When $M \geq D$ (with random f_1), $\text{rank}(\mathbf{X}) = D$ with probability one, and the existing landscape theorem already guarantees every critical point is a global minimum. The gradient independence approach specifically targets the $M < D$ regime, where the output width M' must compensate to achieve $MM' \geq \Omega(D)$.

Why rank preservation is insufficient. Full rank of $\mathbf{R}_t$ ($\det(\mathbf{R}_t) \neq 0$) is preserved almost surely along the GD trajectory for analytic activations, because the set of weight matrices where $\mathbf{R}$ drops rank has measure zero. But this says nothing about the *size* of $\lambda_{\min}(\mathbf{R}_t)$: eigenvalues

¹If the conjecture is ultimately proved only for incoherence (magnitudes) rather than full independence (magnitudes + cancellation), the width condition strengthens to $MM' \geq \Omega(D^2)$. The convergence chain still holds; only the required width changes.

can approach zero without reaching it. What convergence needs is a *quantitative eigenvalue floor* $\lambda_{\min}(\mathbf{R}_t) \geq c > 0$ throughout training. This is the genuine mathematical challenge, and the spectral floor is the central open problem (Item 0) in the main paper.

2 Setup and Notation

Network: $\hat{y}_d = f_3(\mathbf{W}f_1(\mathbf{x}_d^{\text{raw}}))$ with f_1 and f_3 fixed, training only $\mathbf{W}$.

- $\mathbf{x}_d \stackrel{\text{def}}{=} f_1(\mathbf{x}_d^{\text{raw}}) \in \mathbb{R}^M$, $\|\mathbf{x}_d\| = 1$, $d \in [D]$ (post- f_1 features).
- $\mathbf{W} \in \mathbb{R}^{M' \times M}$ (trainable).
- $\mathbf{z}_d \stackrel{\text{def}}{=} \mathbf{W}\mathbf{x}_d \in \mathbb{R}^{M'}$ (pre- f_3 activations).
- $\mathbf{g}_d \stackrel{\text{def}}{=} \nabla_{\mathbf{z}} f_3(\mathbf{z}_d) \in \mathbb{R}^{M'}$, $\alpha_d \stackrel{\text{def}}{=} \|\mathbf{g}_d\|^2$ (sensitivity).
- $\hat{\mathbf{g}}_d \stackrel{\text{def}}{=} \mathbf{g}_d / \|\mathbf{g}_d\|$ (unit gradient direction).
- $e_d \stackrel{\text{def}}{=} f_3(\mathbf{z}_d) - y_d^*$, $\mathcal{E} \stackrel{\text{def}}{=} \sum_d e_d^2$.
- $\mathbf{S} \stackrel{\text{def}}{=} \mathbf{X}^T \mathbf{X} \in \mathbb{R}^{D \times D}$ (data Gram matrix); $S_{dd} = 1$.
- $\mathbf{R} \stackrel{\text{def}}{=} (\mathbf{G}^T \mathbf{G}) \odot \mathbf{S} \in \mathbb{R}^{D \times D}$, with entries $R_{dd} = \alpha_d$, $R_{dd'} = (\mathbf{g}_d^T \mathbf{g}_{d'}) (\mathbf{x}_d^T \mathbf{x}_{d'})$.
- Khatri–Rao columns: $\mathbf{y}_d \stackrel{\text{def}}{=} \mathbf{x}_d \otimes \mathbf{g}_d \in \mathbb{R}^{MM'}$, so $\mathbf{R} = \mathbf{Y}^T \mathbf{Y}$. Since $\mathbf{Y} \in \mathbb{R}^{MM' \times D}$,

$$\text{rank}(\mathbf{R}) = \text{rank}(\mathbf{Y}^T \mathbf{Y}) \leq \min(MM', D) \stackrel{\text{def}}{=} D',$$

an elementary rank cap that will be sharpened in §5 for concrete architectures.

GD update: $\Delta \mathbf{W} = 2\eta \sum_d e_d \mathbf{g}_d \mathbf{x}_d^T$ (ascent), then $\mathbf{W} \leftarrow \mathbf{W} - \Delta \mathbf{W}$.

3 Gradient Incoherence (Pairwise Magnitudes)

We begin with the simpler pairwise condition, which gives $MM' \geq \Omega(D^2)$. The sharper independence condition appears in §4.

Definition 1 (Coherence). *The gradient coherence at weight matrix $\mathbf{W}$ is*

$$\mu_G(\mathbf{W}) \stackrel{\text{def}}{=} \max_{d \neq d'} |\hat{\mathbf{g}}_d^T \hat{\mathbf{g}}_{d'}| = \max_{d \neq d'} \frac{|\mathbf{g}_d^T \mathbf{g}_{d'}|}{\sqrt{\alpha_d \alpha_{d'}}}, \quad \text{where } \mathbf{g}_d = \nabla_{\mathbf{z}} f_3(\mathbf{W}\mathbf{x}_d).$$

The data coherence is $\mu_X \stackrel{\text{def}}{=} \max_{d \neq d'} |\mathbf{x}_d^T \mathbf{x}_{d'}|$.

Lemma 2 (Eigenvalue Bound from Incoherence). *Suppose $\|\mathbf{x}_d\| = 1$ for all d . Then*

$$\lambda_{\min}(\mathbf{R}) \geq \alpha_{\min}(1 - (D-1)\mu_G\mu_X).$$

In particular, $\lambda_{\min}(\mathbf{R}) > 0$ whenever $\mu_G \cdot \mu_X < 1/(D-1)$.

Proof. Factor $\mathbf{R} = \mathbf{A} \mathbf{C} \mathbf{A}$ with $\mathbf{A} = \text{diag}(\sqrt{\alpha_d})$ and $\mathbf{C} = (\hat{\mathbf{G}}^T \hat{\mathbf{G}}) \odot \mathbf{S}$. Then $C_{dd} = 1$ and $|C_{dd'}| \leq \mu_G \mu_X$. For unit $\mathbf{v}$, set $\mathbf{u} = \mathbf{A} \mathbf{v}$; $\|\mathbf{u}\|^2 \geq \alpha_{\min}$, and

$$\mathbf{v}^T \mathbf{R} \mathbf{v} = \mathbf{u}^T \mathbf{C} \mathbf{u} \geq \lambda_{\min}(\mathbf{C}) \alpha_{\min}.$$

By Gershgorin, $\lambda_{\min}(\mathbf{C}) \geq 1 - (D-1) \mu_G \mu_X$. (When $\lambda_{\min}(\mathbf{C}) < 0$ the bound is vacuous; but $\mathbf{R} = \mathbf{Y}^T \mathbf{Y} \succeq 0$ always.) $\square$

Remark 3 (Width from incoherence alone). For random data: $\mu_X = O(1/\sqrt{M})$. For random f_3 : $\mu_G = O(1/\sqrt{M'})$. The positivity condition becomes $(D-1)/\sqrt{MM'} < 1$, i.e., $MM' \geq \Omega(D^2)$. This exceeds the paper’s target $\Omega(D)$. The gap arises because Gershgorin sums $D-1$ *magnitudes* $|C_{dd'}|$ per row. Even if these magnitudes are small on average, they are all positive, so D of them accumulate to $O(D/\sqrt{MM'})$.

Why max $\rightarrow$ average does not help. One might hope that replacing max coherence with average coherence closes the gap. It does not. Gershgorin controls $\lambda_{\min}(\mathbf{C})$ via the row sum $\sum_{d' \neq d} |C_{dd'}|$. Whether each $|C_{dd'}|$ is bounded by the max or the average coherence, the sum is still over $D-1$ positive terms. No pairwise magnitude condition—max, average, or otherwise—can beat $\Omega(D^2)$ through Gershgorin, because Gershgorin always sums D positive terms.

Incoherence vs. Independence: the D^2 -to- D table.

Gradient property		Spectral tool	Width needed	Key mechanism
Incoherence (magnitudes only)		Gershgorin	$MM' \geq \Omega(D^2)$	Sums D positive terms
Independence (magnitudes + signs)		Matrix concentration	$MM' \geq \Omega(D)$	$\sqrt{D}$ cancellation from random signs

The entire gap between $\Omega(D^2)$ and $\Omega(D)$ comes from exploiting sign cancellation rather than relying on magnitude control alone.

4 Gradient Independence (Signed Cancellation)

The key to $\Omega(D)$ width is that for random f_3 , the off-diagonal entries $R_{dd'} = (\mathbf{g}_d^T \mathbf{g}_{d'}) (\mathbf{x}_d^T \mathbf{x}_{d'})$ have not only small magnitudes but approximately *random signs*—positive and negative with roughly equal probability. When D terms with random signs are summed, the typical magnitude is $O(\sqrt{D})$ rather than $O(D)$:

$$\text{magnitudes only (Gershgorin): } \sum_{d'} |C_{dd'}| = O\left(\frac{D}{\sqrt{MM'}}\right),$$

$$\text{with cancellation: } \left| \sum_{d'} C_{dd'} v_{d'} \right| = O\left(\sqrt{\frac{D}{MM'}}\right).$$

This $D \rightarrow \sqrt{D}$ improvement is exactly the factor that separates $\Omega(D^2)$ from $\Omega(D)$. §5 shows that this “independence” hypothesis is not a single monolithic condition on f_3 but decomposes into three separate, checkable conditions on the activation law.

Definition 4 (Approximate Gradient Independence). *We say the gradients $\mathbf{g}_1, \dots, \mathbf{g}_D$ are (σ_g, κ) -approximately independent if:*

- (I1) *Each $\mathbf{g}_d$ is sub-Gaussian in $\mathbb{R}^{M'}$ with $\mathbb{E}[\mathbf{g}_d \mathbf{g}_d^T] = \sigma_g^2 \mathbf{I}_{M'}$ (up to a data-dependent scaling α_d).*
- (I2) *The cross-covariance satisfies $\|\mathbb{E}[\mathbf{g}_d \mathbf{g}_{d'}^T]\|_{\text{op}} \leq \kappa \sigma_g^2$ for $d \neq d'$, with $\kappa \ll 1$.*
- (I3) *Higher-order cumulants are bounded: $\mathbf{g}_d$ conditional on $\{\mathbf{g}_{d'}\}_{d' \neq d}$ remains sub-Gaussian with comparable parameters.*

Theorem 5 (Eigenvalue Bound from Independence). *Suppose $\|\mathbf{x}_d\| = 1$, $\mu_X \leq c/\sqrt{M}$, and the gradients are (σ_g, κ) -approximately independent with $\alpha_d = \|\mathbf{g}_d\|^2$. Then with high probability,*

$$\lambda_{\min}(\mathbf{R}) \geq \alpha_{\min} \left(1 - C \sqrt{\frac{D}{MM'}} - C' \kappa D \right)$$

for universal constants C, C' . In particular, if $\kappa = O(1/D)$ and $MM' \geq \Omega((\bar{\alpha}/\alpha_{\min})^2 \cdot D)$ with $M' \geq \Omega(\log D)$, then $\lambda_{\min}(\mathbf{R}) \geq c \alpha_{\min} > 0$.

Proof sketch. For any unit vector $\mathbf{v} \in \mathbb{R}^D$,

$$\mathbf{v}^T \mathbf{R} \mathbf{v} = \sum_d \sum_{d'} v_d v_{d'} (\mathbf{g}_d^T \mathbf{g}_{d'}) (\mathbf{x}_d^T \mathbf{x}_{d'}) = \left\| \sum_d v_d (\mathbf{x}_d \otimes \mathbf{g}_d) \right\|^2 = \left\| \mathbf{G} \text{diag}(\mathbf{v}) \mathbf{X}^T \right\|_F^2 = \sum_{m'=1}^{M'} \left\| \sum_d v_d g_{d,m'} \mathbf{x}_d \right\|^2.$$

Each inner sum is a weighted combination of the fixed vectors $\mathbf{x}_d$ with random coefficients $v_d g_{d,m'}$. Under approximate independence (I1–I3), the Hanson–Wright inequality applied to $\|\sum_d v_d g_{d,m'} \mathbf{x}_d\|^2 = \mathbf{v}^T \text{diag}(\mathbf{g}^{(m')}) \mathbf{S} \text{diag}(\mathbf{g}^{(m')}) \mathbf{v}$ gives concentration of $\mathbf{v}^T \mathbf{R} \mathbf{v}$ around its expectation, uniformly over $\mathbf{v}$ via an ε -net argument that peels over sparsity levels of $\mathbf{v}$.

The expected Gram matrix satisfies $\mathbb{E}[R_{dd}] = \mathbb{E}[\alpha_d]$ on the diagonal and $|\mathbb{E}[R_{dd'}]| \leq \kappa \sigma_g^2 |\mathbf{x}_d^T \mathbf{x}_{d'}|$ off-diagonal. For κ small, $\lambda_{\min}(\mathbb{E}[\mathbf{R}]) \geq \alpha_{\min}(1 - O(\kappa D))$. Combining gives the result with $MM' \geq \Omega(D)$. $\square$

Remark 6. With $\kappa = 0$ (exact independence) and $\bar{\alpha}/\alpha_{\min} = O(1)$, the condition simplifies to $MM' \geq \Omega(D)$ and $M' \geq \Omega(\log D)$. The $\log D$ factor in Theorem 5’s statement is an artifact of Matrix Bernstein; the Hanson–Wright approach above removes it from the main width condition. Numerical experiments confirm that $\lambda_{\min}(\mathbf{R})/\bar{\alpha}$ tracks the Marchenko–Pastur hard edge $\text{MP} \stackrel{\text{def}}{=} (1 - \sqrt{D/(MM')})^2$ at fixed MM'/D with no degradation as D grows.

Why independence gives cancellation. The operator norm $\|\mathbf{Y}^T \mathbf{Y} - \mathbb{E}[\mathbf{Y}^T \mathbf{Y}]\|_{\text{op}}$ controls $\max_{\|\mathbf{v}\|=1} |\sum_{d \neq d'} R_{dd'} v_d v_{d'}|$. When the $R_{dd'}$ have random signs (which follows from independence of the $\mathbf{g}_d$, since $\mathbf{g}_d^T \mathbf{g}_{d'}$ is a sum of independent products with mean zero), the D^2 off-diagonal terms cancel to $O(\sqrt{D})$ rather than accumulating to $O(D)$. This is the same mechanism as in random matrix theory: the eigenvalues of a Wigner matrix with i.i.d. entries of magnitude $1/\sqrt{D}$ concentrate near $[-2, 2]$, not $[-D, D]$.

5 Architecture Analysis: Which f_3 Achieve Gradient Independence

The phrase “random function with sufficient complexity” used in earlier drafts is too vague to check. This section replaces it with three precise, checkable conditions on the activation law of a

one-hidden-layer random readout, corrects the paper’s diagnosis of the sigmoid failure, and identifies the random Fourier feature (RFF) readout as the architecture that best satisfies all three.

Honesty conventions. Statements below are tagged [THEOREM] (complete proof), [HEURISTIC] (argument valid in a stated limit, not fully proved), or [NUMERICAL] (measured, not proved). All numerical values are reproducible from the scripts listed at the end of §13.

5.1 Setup: the one-hidden-layer random readout

Throughout, the *one-hidden-layer random readout* means

$$f_3(\mathbf{z}) = \sum_{j=1}^F \beta_j \varphi_j(\mathbf{w}_j^T \mathbf{z} + b_j), \quad \mathbf{w}_j \stackrel{iid}{\sim} \mathcal{N}(\mathbf{0}, \sigma_w^2 \mathbf{I}_{M'}), \quad \mathbb{E}[\beta_j] = 0.$$

Write $\mathbf{W}' = [\mathbf{w}_1, \dots, \mathbf{w}_F]^T \in \mathbb{R}^{F \times M'}$ for the (fixed, random) hidden-layer weight matrix, $u_{dj} = \mathbf{w}_j^T \mathbf{z}_d + b_j$ for the pre-activation of node j on input d , and

$$\tau^2 \stackrel{\text{def}}{=} \text{Var}(u_{dj}) = \sigma_w^2 \|\mathbf{z}_d\|^2 + \sigma_b^2 \quad (\text{the operative pre-activation scale}).$$

The GP readout of §5.10 and the RFF readout of §5.9 are both special/limiting cases of this family.

5.2 Effective rank

Lemma 7 (Gradient Field Factorisation, Lemma I.1). *For the one-hidden-layer random readout,*

$$\mathbf{g}_d = \sum_{j=1}^F \underbrace{\beta_j \varphi'_j(u_{dj})}_{=: c_{dj}} \mathbf{w}_j = \mathbf{W}'^T \mathbf{c}_d, \quad \text{hence} \quad \mathbf{G} = \mathbf{W}'^T \mathbf{C}^T, \quad \mathbf{C} \in \mathbb{R}^{D \times F},$$

where $\mathbf{c}_d = (c_{d1}, \dots, c_{dF})^T$ and $\mathbf{C}$ has rows $\mathbf{c}_d^T$. Consequently

$$\text{rank}(\mathbf{G}) \leq \min(F, M', D), \quad \text{rank}(\mathbf{R}) = \text{rank}((\mathbf{G}^T \mathbf{G}) \odot \mathbf{S}) \leq \min(F, M') \cdot \min(M, D).$$

Proof. The first display is the chain rule: $\partial f_3 / \partial \mathbf{z} = \sum_j \beta_j \varphi'_j(u_{dj}) \partial u_{dj} / \partial \mathbf{z} = \sum_j \beta_j \varphi'_j(u_{dj}) \mathbf{w}_j$. Stacking over d gives $\mathbf{G} = \mathbf{W}'^T \mathbf{C}^T$, so $\text{rank}(\mathbf{G}) \leq \min(\text{rank}(\mathbf{W}'), \text{rank}(\mathbf{C})) \leq \min(F, M', D)$. For the second claim, use the Schur-product rank bound $\text{rank}(\mathbf{A} \odot \mathbf{B}) \leq \text{rank}(\mathbf{A}) \cdot \text{rank}(\mathbf{B})$ (Horn–Johnson, *Topics in Matrix Analysis*, §5.5) with $\mathbf{A} = \mathbf{G}^T \mathbf{G}$ ($\text{rank} \leq \text{rank}(\mathbf{G}) \leq \min(F, M')$) and $\mathbf{B} = \mathbf{S}$ ($\text{rank} = \min(M, D)$). $\square$

Corollary 8 (Effective-rank obstruction, Corollary I.2). $\lambda_{\min}(\mathbf{R}) > 0$ requires $\min(F, M') \cdot M \geq D$.

This is the quantitative form of the observation that a readout of the form $f_3(\mathbf{z}) = h(\mathbf{A}\mathbf{z})$ has $\text{rank}(\mathbf{R}) \leq rM$ for $r = \text{rank}(\mathbf{A})$: the operative width is $\min(F, M')$, **not** M' . A readout with few hidden units is rank-limited no matter how large M' is.

5.3 The shared direction is provably harmless

This is the most important correction to the previous narrative: the sign-definiteness diagnosis for why sigmoid-type networks fail is **wrong**.

F	$\min(F, M')$	M	$\text{rank}(\mathbf{R})$	$\lambda_{\min}/\bar{\alpha}$
33		396	396	-0.000000
64		768	400	0.0108
120		1440	400	0.0512
1200		1440	400	0.1765

Table 1: [NUMERICAL] $M = 12$, $M' = 120$, $D = 400$ (so $\min(F, M') \geq 34$ is needed). The rank transition occurs at exactly the predicted F .

Lemma 9 (Projection Lemma, Lemma I.3). *Let $\mathbf{\Pi}$ be **any** orthogonal projection on $\mathbb{R}^{M'}$. Then*

$$\lambda_{\min}((\mathbf{G}^T \mathbf{G}) \odot \mathbf{S}) \geq \lambda_{\min}((\mathbf{\Pi} \mathbf{G})^T (\mathbf{\Pi} \mathbf{G})) \odot \mathbf{S}).$$

Proof. For unit $\mathbf{v} \in \mathbb{R}^D$,

$$\mathbf{v}^T [(\mathbf{G}^T \mathbf{G}) \odot \mathbf{S}] \mathbf{v} = \sum_{d,d'} v_d v_{d'} (\mathbf{g}_d^T \mathbf{g}_{d'}) (\mathbf{x}_d^T \mathbf{x}_{d'}) = \left\| \sum_d v_d (\mathbf{x}_d \otimes \mathbf{g}_d) \right\|^2.$$

$\mathbf{I}_M \otimes \mathbf{\Pi}$ is an orthogonal projection on $\mathbb{R}^{MM'}$, hence a contraction, and $(\mathbf{I}_M \otimes \mathbf{\Pi})(\mathbf{x}_d \otimes \mathbf{g}_d) = \mathbf{x}_d \otimes \mathbf{\Pi} \mathbf{g}_d$. Therefore $\left\| \sum_d v_d \mathbf{x}_d \otimes \mathbf{g}_d \right\|^2 \geq \left\| \sum_d v_d \mathbf{x}_d \otimes \mathbf{\Pi} \mathbf{g}_d \right\|^2 = \mathbf{v}^T [(\mathbf{\Pi} \mathbf{G})^T (\mathbf{\Pi} \mathbf{G})) \odot \mathbf{S}] \mathbf{v}$. Take the minimum over $\mathbf{v}$. $\square$

Corollary 10 (Shared Components Are Free, Corollary I.4). *If $\mathbf{g}_d = \bar{\mathbf{v}} + \mathbf{h}_d$ with $\bar{\mathbf{v}}$ **independent of** d (more generally, the d -independent part lies in a fixed subspace V with $\dim V = r$), then with $\mathbf{\Pi} = \mathbf{\Pi}_{V^\perp}$ and $\mathbf{h}_{d,\perp} = \mathbf{\Pi}_{V^\perp} \mathbf{h}_d$,*

$$\lambda_{\min}(\mathbf{R}) \geq \lambda_{\min}((\mathbf{H}_\perp^T \mathbf{H}_\perp) \odot \mathbf{S}).$$

Remark 11 (Interpretation). A collinear/shared gradient component can **never** be the cause of a small $\lambda_{\min}(\mathbf{R})$: it can be projected away at the cost of r dimensions of M' . What it *does* do is inflate $\bar{\alpha} = \|\bar{\mathbf{v}}\|^2 + \mathbb{E}\|\mathbf{h}\|^2$, which degrades the *ratio* $\lambda_{\min}/\bar{\alpha}$ —and therefore the convergence rate relative to the natural scale—by the factor $1 - \gamma$ of §5.4.

architecture	$k = 0$	$k = 1$	$k = 2$	$k = 4$	$k = 8$	top-4 s.v. share
sigmoid, $L=1$	0.00081	0.00079	0.00074	0.00071	0.00061	0.954, 0.004, 0.003, 0.003
tanh, $L=1$	0.00976	0.00956	0.00923	0.00860	0.00741	0.806, ...
$ u $, $L=1$	0.04642	0.04491	0.04384	0.04090	0.03384	0.150, ...

Table 2: [NUMERICAL] $\lambda_{\min}(\mathbf{R})$ after projecting out the top k left singular directions of $\mathbf{G}$. Monotone in k , as Lemma 9 requires; the 95.4% DC direction of sigmoid costs only 2% of $\lambda_{\min}$.

5.4 The Hermite DC ratio

Lemma 12 (Exact DC Decomposition, Lemma I.5). *Let $m_j \stackrel{\text{def}}{=} \mathbb{E}_U[\varphi'_j(U)]$ with $U \sim \mathcal{N}(0, \tau^2)$, and $\psi_j \stackrel{\text{def}}{=} \varphi'_j - m_j$. Then*

$$\mathbf{g}_d = \bar{\mathbf{v}} + \mathbf{h}_d, \quad \bar{\mathbf{v}} = \sum_j \beta_j m_j \mathbf{w}_j \quad (d\text{-independent}), \quad \mathbf{h}_d = \sum_j \beta_j \psi_j(u_{dj}) \mathbf{w}_j.$$

Proof. Immediate from Lemma 7: write $c_{dj} = \beta_j \varphi'_j(u_{dj}) = \beta_j m_j + \beta_j \psi_j(u_{dj})$ and sum the two pieces separately against $\mathbf{w}_j$. $\square$

Lemma 13 (Stein/Hermite Identity, Lemma I.6). *Expand $\varphi = \sum_{k \geq 0} a_k h_k(\cdot/\tau)$ in normalised probabilists’ Hermite polynomials $h_k = \text{He}_k/\sqrt{k!}$ (orthonormal for the standard Gaussian measure). Then, using $h'_k = \sqrt{k} h_{k-1}$,*

$$\varphi'(u) = \tau^{-1} \sum_{k \geq 1} a_k \sqrt{k} h_{k-1}(u/\tau), \quad m = \mathbb{E}[\varphi'(U)] = \frac{a_1}{\tau} = \frac{\mathbb{E}[U \varphi(U)]}{\tau^2}.$$

Moreover, if (U, U') are jointly Gaussian, mean zero, variance τ^2 , correlation ρ , then by Mehler’s formula $\mathbb{E}[h_j(U/\tau)h_k(U'/\tau)] = \delta_{jk}\rho^j$, so

$$K_g(\rho) \stackrel{\text{def}}{=} \mathbb{E}[\varphi'(U)\varphi'(U')] = \tau^{-2} \sum_{k \geq 1} k a_k^2 \rho^{k-1}.$$

Definition 14 (The DC Ratio, Corollary I.7).

$$\gamma \stackrel{\text{def}}{=} \frac{\mathbb{E}\|\bar{\mathbf{v}}\|^2}{\mathbb{E}\|\mathbf{g}_d\|^2} = \frac{\mathbb{E}_\varphi[m^2]}{\mathbb{E}_\varphi\mathbb{E}[\varphi'(U)^2]} \stackrel{\text{single } \varphi}{=} \frac{a_1^2}{\sum_k k a_k^2} = \frac{K_g(0)}{K_g(1)}.$$

(The middle equality uses $\mathbb{E}[\mathbf{w}_j^T \mathbf{w}_k] = 0$ for $j \neq k$, so all cross terms in $\mathbb{E}\|\bar{\mathbf{v}}\|^2$ and $\mathbb{E}\|\mathbf{g}_d\|^2$ vanish.)

Remark 15. The shared gradient direction is generated exactly by the degree-1 Hermite component of the activation—equivalently, it is the gradient of the *linear part* of f_3 , whose gradient field is constant. This links the program to the information-exponent/“leap” literature (Ben Arous–Gheissari–Jagannath; Damian–Lee–Soltanolkotabi; Abbe–Boix–Adserà–Misiakiewicz), where the first non-zero Hermite coefficient is the governing quantity—a well-understood, citable object, which is a real gain over the ad hoc “sufficient random complexity.”

activation	$\tau=0.25$	0.5	1	2	4	8	large- τ law
Sig(u)	0.9996	0.9943	0.9522	0.7900	0.5198	0.2876	$2.394/\tau$
tanh(u)	0.9943	0.9522	0.7900	0.5198	0.2876	0.1481	$1.197/\tau$
softplus	0.9851	0.9471	0.8521	0.7172	0.6115	0.5539	$\rightarrow 1/2$
ReLU	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	$\frac{1}{2}$ exactly
GELU	0.8761	0.7029	0.5484	0.4940	0.4895	0.4935	$\rightarrow 1/2$
$ u $	0	0	0	0	0	0	0 exactly
cos(u), rand. phase	—	0.7807	0.3651	0.0176	10^{-5}	—	$e^{-\tau^2}$
cos($3u$), rand. phase	—	0.1058	0.0001	0	0	—	$e^{-9\tau^2}$

Table 3: [NUMERICAL] $\gamma_\varphi(\tau) = (\mathbb{E}\varphi')^2/\mathbb{E}[\varphi'^2]$. Stein identity check ($\tau = 2$): $\mathbb{E}[\varphi']$ vs. a_1/τ agree to 6 decimals for every activation tested.

Proposition 16 (Sign-Blindness, Proposition I.8). γ depends on the activation law only through $\mathbb{E}_\varphi[m^2] \geq 0$. Hence replacing $\varphi_j \mapsto \varepsilon_j \varphi_j$ with random signs $\varepsilon_j = \pm 1$ leaves γ **unchanged**, and more generally no randomisation over a family of activations with $|m|$ bounded below can reduce γ below $\min_\varphi m^2 / \max_\varphi \mathbb{E}[\varphi'^2]$.

Proof. Each hidden node contributes the PSD rank-one piece $\beta_j^2 m_j^2 \mathbf{1}\mathbf{1}^T$ to $\mathbb{E}[\mathbf{G}^T \mathbf{G}]$'s constant part (Lemma 12). Randomising the sign of φ_j multiplies m_j by ε_j and hence multiplies $\beta_j^2 m_j^2$ by $\varepsilon_j^2 = 1$: it is invisible to γ , which depends only on m_j^2 . (It is also invisible to the model itself, since ε_j is absorbed into β_j .) There are no signs of $m_j^2 \geq 0$ to cancel, so randomising over any family with $|m| \geq m_{\min} > 0$ cannot push γ below $m_{\min}^2 / \max \mathbb{E}[\varphi'^2]$. $\square$

random family	$\tau=0.5$	1	2	4
random- sign $\varepsilon \text{Sig}(u)$	0.9943	0.9522	0.7900	0.5198
plain $\text{Sig}(u)$ (reference)	0.9943	0.9522	0.7900	0.5198
random- shift $\text{Sig}(u - b)$, $b \sim \mathcal{N}(0, 9)$	0.9577	0.8639	0.6796	0.4593
random- gain $\tanh(gu)$, $g \sim U[.5, 3]$	0.7738	0.5405	0.3134	0.1665
random- phase $\cos(3u + \xi)$	0.1058	0.0001	0	0

Table 4: [NUMERICAL] Random signs are **identical to five decimals**. Random shifts help by 7%; random gains help only by mixing τ scales; random phases help exponentially.

Random piecewise-linear activations. [THEOREM/NUMERICAL] Let φ' be ± 1 on K bins of Gaussian probabilities $P_1, \dots, P_K$ with i.i.d. random signs. Then $\mathbb{E}_s[m^2] = \sum_k P_k^2$ and $\mathbb{E}[\varphi'^2] = 1$, so $\gamma = \sum_k P_k^2 \stackrel{\text{equal mass}}{=} 1/K$. Measured, against $1/K$ (exact match):

K	1	2	4	8	16	32	64
γ	1.000	0.510	0.263	0.127	0.0615	0.0296	0.0142

Remark 17 (Unified reading). γ^{-1} is, up to constants, *the number of sign changes of φ' across the operative pre-activation range*, or $e^{+\#}$ for smooth oscillation. This is the honest replacement for “ φ' must take both signs.”

5.5 γ is necessary but not sufficient: the three-part criterion

Proposition 18 (Counterexample: $\gamma = 0$ with $\lambda_{\min} = 0$, Counterexample I.9). *Take $\varphi = \text{He}_2$, so $\varphi'(u) = \sqrt{2}u$ is **linear** (hence $a_1 = 0$ and $\gamma = 0$ exactly). Then*

$$\mathbf{g}_d = \sqrt{2} \sum_j \beta_j (\mathbf{w}_j^T \mathbf{z}_d + b_j) \mathbf{w}_j = \underbrace{\sqrt{2} \left(\sum_j \beta_j \mathbf{w}_j \mathbf{w}_j^T \right) \mathbf{z}_d}_{=: \mathbf{A}} + \underbrace{\sqrt{2} \sum_j \beta_j b_j \mathbf{w}_j}_{=: \mathbf{c}},$$

an **affine** function of $\mathbf{z}_d = \mathbf{W} \mathbf{x}_d$, hence an affine function of $\mathbf{x}_d$ with linear part of dimension $\leq M$. Thus $\text{rank}(\mathbf{G}) \leq M + 1$ and $\text{rank}(\mathbf{R}) \leq (M + 1)M$. Measured at $M = 12, D = 400$: $\lambda_{\min}(\mathbf{R}) = 0$ to machine precision while $1 - \gamma = 0.84$.

So the criterion is genuinely three-part:

[Criterion A] Effective rank. $\min(F, M') \cdot M \geq D$ (Corollary 8), and more sharply, the Hermite mass of the activation must not sit entirely at degree ≤ 1 (Proposition 18).

[Criterion B] No shared direction. $\gamma = a_1^2 / \sum_k k a_k^2$ bounded away from 1. (Failing this costs a factor $1 - \gamma$ in $\lambda_{\min} / \bar{\alpha}$, *not* a rank deficiency—Lemma 9.)

[Criterion C] Decorrelation. $K_g(\rho_{dd'})/K_g(1) = \sum_k k a_k^2 \rho^{k-1} / \sum_k k a_k^2$ small at the operative correlation $\rho_{dd'}$. This requires **high effective Hermite degree**, and it is the binding condition in practice.

activation	$\lambda_{\min}/\bar{\alpha}$	/MP	$f_1=1-\gamma$	f_2/MP	$f_1 f_2/\text{MP}$	$\bar{\alpha}/\alpha_{\min}$
sigmoid	0.03497	0.156	0.325	0.473	0.154	1.48
tanh	0.10295	0.460	0.633	0.714	0.452	1.57
ReLU	0.03041	0.136	0.442	0.295	0.130	1.42
GELU	0.02692	0.120	0.463	0.253	0.117	1.38
mixed monotone (4 kinds/node)	0.03935	0.176	0.502	0.339	0.170	1.57
random-sign sigmoid	0.03533	0.158	0.361	0.429	0.155	1.61
$ u $	0.06171	0.276	0.890	0.304	0.271	1.51
He ₂	-0.00000	0.000	0.841	0.000	0.000	1.86
He ₄	0.02079	0.093	0.862	0.106	0.091	4.87
zig-zag $K=8$	0.08504	0.380	0.711	0.517	0.367	1.54
zig-zag $K=32$	0.12173	0.544	0.813	0.646	0.525	1.50
RFF $\cos(u+\xi)$	0.17879	0.799	0.971	0.799	0.776	1.58
RFF $\cos(3u+\xi)$	0.18865	0.843	0.972	0.839	0.815	1.59

Table 5: [NUMERICAL] $M=12$, $M'=120$, $D=400$, $F=1200$, $\tau=3$, $\text{MP}=0.2237$. $f_1 \cdot f_2$ reproduces the measured value to three digits (the projection inequality of Lemma 9 is essentially tight). Mixing four monotone activations per node gains 13% over plain sigmoid. $|u|$ has $f_1=0.89$ but $f_2/\text{MP}=0.30$: killing γ alone is not enough. He₄ has $\gamma=0$ and good f_1 but is wrecked by $\bar{\alpha}/\alpha_{\min}=4.9$.

5.6 Scale sensitivity: the sharpest architectural distinction

Proposition 19 (Homogeneous Activations Are Scale-Frozen, Proposition I.10). *If φ is positively homogeneous of degree p ($\varphi(cu) = c^p \varphi(u)$, $c > 0$ —ReLU, $|u|$, leaky ReLU, and pure monomials), then φ' is homogeneous of degree $p-1$, so for $\sigma_b=0$ the normalised gradient field $\hat{\mathbf{g}}_d$ is **invariant** under $\mathbf{W} \mapsto c\mathbf{W}$. Hence $\lambda_{\min}(\mathbf{R})/\bar{\alpha}$ does not depend on the initialisation scale at all.*

Proof. $\mathbf{g}_d(c\mathbf{W}) = c^{p-1} \mathbf{g}_d(\mathbf{W})$ entrywise, so $\mathbf{G} \mapsto c^{p-1} \mathbf{G}$ and both $\mathbf{R}$ and $\bar{\alpha}$ scale by $c^{2(p-1)}$. $\square$

Proposition 20 (Dot-Product vs. Stationary Kernels, Proposition I.11). *For a readout with no random phase, the pre-activation correlation is*

$$\rho_{dd'} = \frac{\sigma_w^2 \mathbf{z}_d^T \mathbf{z}_{d'} + \sigma_b^2}{\sigma_w^2 \|\mathbf{z}_d\|^2 + \sigma_b^2} \xrightarrow{\sigma_w \rightarrow \infty} \frac{\mathbf{z}_d^T \mathbf{z}_{d'}}{\|\mathbf{z}_d\|^2} \approx \mathbf{x}_d^T \mathbf{x}_{d'} = S_{dd'},$$

which **saturates at the data coherence** and cannot be reduced by scaling $\mathbf{W}$. For a random-phase (Fourier) readout the gradient kernel is stationary,

$$K_g \propto e^{-\sigma_w^2 \|\mathbf{z}_d - \mathbf{z}_{d'}\|^2 / 2},$$

which $\rightarrow 0$ as σ_w or $\|\mathbf{W}_0\|$ grows. (The first “ $\rightarrow$ ” uses the standard concentration $\mathbf{z}_d^T \mathbf{z}_{d'} \approx \sigma_W^2 M' \mathbf{x}_d^T \mathbf{x}_{d'}$ for random $\mathbf{W}_0$ —[HEURISTIC]; the stationary kernel identity is [THEOREM].)

Remark 21. This, not sign-definiteness, is the deepest architectural distinction. A dot-product readout can only decorrelate by raising its *Hermite degree* (ρ^{k-1}); a Fourier readout can decorrelate by raising its *bandwidth*, which the designer controls.

activation	$\tau=1$	2	4	8	16	32
sigmoid	0.005	0.060	0.267	0.527	0.673	0.703
ReLU	0.148	0.148	0.148	0.148	0.148	0.148
$ u $	0.287	0.287	0.287	0.287	0.287	0.287
He ₄	0.220	0.115	0.095	0.091	0.090	0.089
RFF $\cos(u+\xi)$	0.048	0.447	0.826	0.847	0.849	0.862
zig-zag $K=32$	0.808	0.677	0.472	0.327	0.236	0.182

Table 6: [NUMERICAL] $\lambda_{\min}(\mathbf{R})/(\bar{\alpha} \cdot \text{MP})$ as pre-activation scale τ is swept. Sigmoid improves with τ (its effective Hermite degree grows), ReLU/ $|u|$ are frozen (Proposition 19), fixed-bin zig-zags *worsen* with τ (their bins stop resolving the range), and the Fourier readout improves monotonically—the designer’s bandwidth knob at work.

5.7 Depth (brief remark)

Write $f_3(\mathbf{z}) = \boldsymbol{\omega}^T \varphi_L(\mathbf{V}_L \cdots \varphi_1(\mathbf{V}_1 \mathbf{z} + \mathbf{c}_1) \cdots)$, so

$$\mathbf{g}_d = \mathbf{V}_1^T \mathbf{D}_1^{(d)} \mathbf{V}_2^T \mathbf{D}_2^{(d)} \cdots \mathbf{V}_L^T \mathbf{D}_L^{(d)} \boldsymbol{\omega}, \quad \mathbf{D}_l^{(d)} = \text{diag}(\varphi_l'(\text{pre}_l^{(d)})).$$

Proposition 22 (Depth Multiplies γ , Proposition I.12). [HEURISTIC, MEAN-FIELD.] Write $\mathbf{D}_l^{(d)} = m_l \mathbf{I} + \boldsymbol{\Pi}_l^{(d)}$ with $\mathbb{E}[\boldsymbol{\Pi}_l^{(d)}] = 0$. In the infinite-width, uncorrelated-layer limit, the only term in $\mathbf{g}_d$ independent of d is $(\prod_l m_l) \mathbf{V}_1^T \cdots \mathbf{V}_L^T \boldsymbol{\omega}$, giving

$$\gamma_L = \prod_{l=1}^L \gamma_l.$$

If true, sigmoid ($\gamma \approx 0.95$) would need $L \approx 25 \ln D$ layers for $\gamma_L = 1/D$, while ReLU ($\gamma = 1/2$ exactly) needs only $L = \log_2 D$. [NUMERICAL] confirms the exponential decay of the top-singular-value share for $L \leq 3$, but $\lambda_{\min}/\text{MP}$ is **non-monotone in L** : it peaks at $L \approx 3$ (tanh, moderate gain) or $L \approx 6-8$ ($|u|$, ReLU) and then *degrades*, because depth also inflates $\bar{\alpha}/\alpha_{\min}$ and collapses forward representations. At high gain, depth is monotonically harmful. Depth never beats a one-layer Fourier readout (0.78–0.84 vs. a best depth result of ≈ 0.71). **Conclusion: depth helps exponentially in γ but is a worse engineering answer than a Fourier readout, and it introduces a second conditioning problem. We keep it as a remark and do not build the paper on it.**

5.8 Correcting the sigmoid failure diagnosis

Definition 23 (Two-Layer Sigmoid Network). Consider the architecture

$$f_3(\mathbf{z}_d) = \boldsymbol{\omega}^T \text{Sig}(\mathbf{W}' \text{Sig}(\mathbf{z}_d) + \boldsymbol{\omega}'),$$

with random $\mathbf{W}' \in \mathbb{R}^{M' \times M'}$, $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \mathbb{R}^{M'}$. The gradient is $\mathbf{g}_d = \text{Sig}'(\mathbf{z}_d) \odot ((\mathbf{W}')^T \mathbf{p}_d)$, where $\mathbf{p}_d = \text{Sig}'(\mathbf{h}_d) \odot \boldsymbol{\omega}$ is the backpropagated signal.

Old (incorrect) diagnosis. An earlier version of these notes claimed sigmoid fails because $\text{Sig}' > 0$ everywhere forces every $\mathbf{p}_d = \mathbf{D}_d \boldsymbol{\omega}$ ($\mathbf{D}_d \succ 0$ diagonal) to be a positive rescaling of the shared vector $\boldsymbol{\omega}$, so all $(\mathbf{W}')^T \mathbf{p}_d$ are nearly parallel, and concluded this collinearity is the failure mechanism.

Corrected diagnosis. Lemma 9 shows this reasoning is wrong: *any* shared/collinear component of the gradient field can be projected away at negligible cost. Table 9 (§5.3) confirms it directly: removing sigmoid’s dominant (95.4%-of-energy) shared direction changes $\lambda_{\min}(\mathbf{R})$ by only 2%. The real failure has two, separate, quantifiable causes (Table 5, sigmoid row):

1. **The DC ratio** $\gamma \approx 0.95$ **at the operative scale** costs a factor $f_1 = 1 - \gamma \approx 0.33$ in $\bar{\alpha}_\perp/\bar{\alpha}$ (Definition 14, Table 1).
2. **An independent decorrelation failure:** even after removing the shared direction, the centered field is poorly conditioned, $f_2/\text{MP} \approx 0.47$ (Criterion C, §5.5).

The product $f_1 f_2 \approx 0.15$ matches the measured $\lambda_{\min}/(\bar{\alpha} \cdot \text{MP}) = 0.156$ to three digits. This failure is architectural—it degrades gradually with τ (Table 5.6, sigmoid row) rather than being a fixed collinearity defect, and persists (in weaker form) for tanh and other monotone, saturating activations, all of which have γ close to 1 at moderate τ .

5.9 The RFF readout, with the rank condition

Definition 24 (RFF Readout). *Let $\omega_1, \dots, \omega_F \in \mathbb{R}^{M'}$ be i.i.d. from a distribution with density p (e.g., $\mathcal{N}(\mathbf{0}, \sigma^{-2}\mathbf{I})$), $\phi_j \sim \text{Unif}[0, 2\pi]$, and $\beta_j \sim \mathcal{N}(0, 1/F)$, all independent. Define*

$$f_3(\mathbf{z}) = \sqrt{\frac{2}{F}} \sum_{j=1}^F \beta_j \cos(\omega_j^T \mathbf{z} + \phi_j).$$

The RFF readout is an instance of the one-hidden-layer random readout of §5 with $\varphi_j(u) = \cos(u + \phi_j)$, a random-phase family. By Criterion A (Corollary 8) it requires

$$\min(F, M') \cdot M \geq D$$

as a necessary condition—absent from earlier drafts. By Criterion B (Table 1), random phase makes $\mathbb{E}[\varphi'_j] = -\sin(\phi_j)$ zero *in expectation* over ϕ_j , so γ decays exponentially with τ ($\gamma = e^{-\tau^2}$ for the base frequency, faster for higher-frequency variants); this is why RFF has no shared-direction problem at all. By Criterion C, the gradient kernel is stationary (Proposition 20) and decorrelates with bandwidth, not merely with data separation. Concretely:

- **GP limit.** As $F \rightarrow \infty$, f_3 converges to a GP with the kernel $k(\mathbf{z}, \mathbf{z}') = \int \cos(\omega^T(\mathbf{z} - \mathbf{z}')) p(\omega) d\omega$. For $\omega \sim \mathcal{N}(\mathbf{0}, \sigma^{-2}\mathbf{I})$, this recovers the SE kernel of §5.10, and Proposition 26 carries over in the limit.
- **Finite F suffices,** provided Criterion A holds: in simulations with $F = O(M')$ and $\min(F, M')M \geq D$, $\lambda_{\min}(\mathbf{R})/\bar{\alpha}$ matches the Marchenko–Pastur ideal to 2–3 significant figures (Table 5, last two rows).

Why quadratic f_3 fails. This is now a special case of Proposition 18: a quadratic readout has activation of pure Hermite degree ≤ 2 , so its gradient field is exactly affine in $\mathbf{z}$, giving $\text{rank}(\mathbf{R}) \leq (M+1)M < D$ for D large enough—regardless of how the quadratic form’s coefficients are randomised. A single query exhausts its randomness. A valid law needs Hermite mass at high enough degree that both γ and the decorrelation kernel $K_g(\rho)/K_g(1)$ are small—which the GP and RFF architectures both provide.

5.10 Gaussian process limit

Definition 25 (Gaussian Process Readout). A Gaussian process (GP) readout is a random function $f_3 : \mathbb{R}^{M'} \rightarrow \mathbb{R}$ such that for any finite set of inputs $\mathbf{z}_1, \dots, \mathbf{z}_n \in \mathbb{R}^{M'}$, the outputs $(f_3(\mathbf{z}_1), \dots, f_3(\mathbf{z}_n))$ are jointly Gaussian with

$$\begin{aligned}\mathbb{E}[f_3(\mathbf{z})] &= 0, \\ \text{Cov}(f_3(\mathbf{z}), f_3(\mathbf{z}')) &= k(\mathbf{z}, \mathbf{z}').\end{aligned}$$

A GP is stationary if $k(\mathbf{z}, \mathbf{z}') = \phi(\|\mathbf{z} - \mathbf{z}'\|)$ for some function $\phi : [0, \infty) \rightarrow \mathbb{R}$ with $\phi(0) = 1$ (after normalization). The correlation length $\sigma > 0$ is the scale at which correlations decay: $k(\mathbf{z}, \mathbf{z}') \approx 0$ when $\|\mathbf{z} - \mathbf{z}'\| \gg \sigma$.

Common kernels.

- *Squared exponential (SE)*: $k(\mathbf{z}, \mathbf{z}') = \exp(-\|\mathbf{z} - \mathbf{z}'\|^2/(2\sigma^2))$. Infinitely differentiable sample paths. The gradient process ∇f_3 is also a GP.
- *Matérn- ν* : $k(\mathbf{z}, \mathbf{z}') = \frac{2^{1-\nu}}{\Gamma(\nu)} (\sqrt{2\nu} r/\sigma)^\nu K_\nu(\sqrt{2\nu} r/\sigma)$ with $r = \|\mathbf{z} - \mathbf{z}'\|$. Sample paths are $\lceil \nu \rceil - 1$ times differentiable. Requires $\nu > 1$ for gradients to exist.

Gradient covariance structure. Since f_3 is Gaussian and differentiable, $\nabla_{\mathbf{z}} f_3$ is also a Gaussian process with covariance

$$\text{Cov}(\nabla f_3(\mathbf{z}), \nabla f_3(\mathbf{z}')) = \nabla_{\mathbf{z}} \nabla_{\mathbf{z}'}^T k(\mathbf{z}, \mathbf{z}') \in \mathbb{R}^{M' \times M'}. \quad (1)$$

For the SE kernel, writing $\boldsymbol{\delta} = \mathbf{z} - \mathbf{z}'$ and $r^2 = \|\boldsymbol{\delta}\|^2$:

$$\text{Cov}(\mathbf{g}_d, \mathbf{g}_{d'}) = \frac{e^{-r^2/(2\sigma^2)}}{\sigma^2} \left(\mathbf{I}_{M'} - \frac{\boldsymbol{\delta} \boldsymbol{\delta}^T}{\sigma^2} \right). \quad (2)$$

Key properties.

1. **Marginal distribution:** At any single input, $\mathbf{g}_d \sim \mathcal{N}(\mathbf{0}, \sigma^{-2} \mathbf{I}_{M'})$, so $\alpha_d = \|\mathbf{g}_d\|^2 \sim \sigma^{-2} \chi_{M'}^2$ with $\mathbb{E}[\alpha_d] = M'/\sigma^2$. In particular, $\alpha_d > 0$ almost surely, and $\alpha_d \approx M'/\sigma^2$ concentrates for large M' .
2. **Well-separated inputs** ($\|\mathbf{z}_d - \mathbf{z}_{d'}\| \gg \sigma$): $k(\mathbf{z}_d, \mathbf{z}_{d'}) \approx 0$, so $\text{Cov}(\mathbf{g}_d, \mathbf{g}_{d'}) \approx \mathbf{0}$. The gradients are approximately *independent* Gaussian vectors. Conditions I1–I3 hold with $\kappa = \max_{d \neq d'} e^{-\|\mathbf{z}_d - \mathbf{z}_{d'}\|^2/(2\sigma^2)} \approx 0$.
3. **Nearby inputs** ($\|\mathbf{z}_d - \mathbf{z}_{d'}\| \ll \sigma$): $k \approx 1$, so $\text{Cov}(\mathbf{g}_d, \mathbf{g}_{d'}) \approx \sigma^{-2} \mathbf{I}_{M'}$, meaning $\mathbf{g}_d \approx \mathbf{g}_{d'}$. Gradient independence fails.
4. **Intermediate regime:** the covariance interpolates smoothly. Gradient independence holds whenever $\Delta_{\min} \stackrel{\text{def}}{=} \min_{d \neq d'} \|\mathbf{z}_d - \mathbf{z}_{d'}\|$ satisfies $\Delta_{\min} \gg \sigma$.

Proposition 26 (GP Gradient Independence). Let f_3 be a GP with SE kernel and correlation length σ . If $\Delta_{\min} \geq c_0 \sigma \sqrt{\log D}$ for a sufficiently large constant c_0 , then the gradients $\mathbf{g}_1, \dots, \mathbf{g}_D$ satisfy conditions I1–I3 of Definition 4 with $\sigma_g = 1/\sigma$ and $\kappa = O(D^{-c_0^2/2})$.

Proof sketch. The cross-covariance (2) has operator norm $\|\text{Cov}(\mathbf{g}_d, \mathbf{g}_{d'})\|_{\text{op}} \leq \sigma^{-2} e^{-\Delta_{\min}^2/(2\sigma^2)}$. With $\Delta_{\min} = c_0 \sigma \sqrt{\log D}$: $e^{-c_0^2 \log D/2} = D^{-c_0^2/2}$. For $c_0 \geq 2$, this gives $\kappa = O(1/D)$, and $\kappa D = O(1)$.

Condition I3 (conditional sub-Gaussianity) follows from the Gaussian structure: conditional distributions of a GP remain Gaussian with covariance bounded by the marginal covariance. $\square$

What remains for the GP case. Proposition 26 establishes gradient independence at a *single* $\mathbf{W}$. The full conjecture requires uniformity over the GD trajectory, which §7 now handles directly (rather than via a ball around $\mathbf{W}_0$). It reduces to showing that the minimum separation $\Delta_{\min}(\mathbf{W}) = \min_{d \neq d'} \|\mathbf{W}\mathbf{x}_d - \mathbf{W}\mathbf{x}_{d'}\|$ stays $\gg \sigma$ along the trajectory. For random $\mathbf{W}_0$ with initialization scale $\sigma_{\mathbf{W}}$,

$$\Delta_{\min}(\mathbf{W}_0) \approx \sigma_{\mathbf{W}} \sqrt{M'/M} \min_{d \neq d'} \|\mathbf{x}_d - \mathbf{x}_{d'}\|,$$

and movement of total size B changes separations by at most $B \max_{d \neq d'} \|\mathbf{x}_d - \mathbf{x}_{d'}\|$.

5.11 Q1 summary: the binding gaps and the best path

What is *not* yet proved:

- [GAP 1] Proposition 22 ($\gamma_L = \prod_l \gamma_l$) at finite width, which needs a Gaussian-equivalence theorem for the backward pass of a deep random net (plausibly available from Tensor Programs-type results, not verified here in citable form).
- [GAP 2] A lower bound on f_2 (Criterion C): the sharp fixed- $\mathbf{W}$ Khatri–Rao theorem (Hanson–Wright + sparsity-peeling, §4) for a gradient field with kernel $K_g(\rho)$ rather than exactly independent Gaussians. **This is the single missing static lemma.**
- [GAP 3] The crowding/coherence tension: for a dot-product readout, $\rho_{dd'} = S_{dd'}$ is fixed by the data, so the Welch bound caps conditioning unless the Hermite degree is $\gtrsim \log D / \log(1/\mu_X)$; for a stationary readout the escape is bandwidth instead. Not yet a theorem.

Recommended path: keep the RFF readout as the target architecture and state Criteria A–C as hypotheses of Conjecture C4 (§10); drop “different activation per node” (measured 13% gain only); keep depth as a remark, not a route.

6 Proved: Eigenvalue Bound Implies GD Convergence

Theorem 27 (Gradient Flow). *Along gradient flow $\dot{\mathbf{W}} = -\nabla_{\mathbf{W}} \mathcal{E}$, the total error satisfies $\dot{\mathcal{E}} = -4 \mathbf{e}^T \mathbf{R} \mathbf{e}$. If $\lambda_{\min}(\mathbf{R}(t)) \geq \ell(t) > 0$ throughout $[0, T]$, then*

$$\mathcal{E}(t) \leq \mathcal{E}(0) \exp\left(-4 \int_0^t \ell(s) ds\right).$$

A uniform floor $\ell(t) \geq \mu > 0$ gives $\mathcal{E}(t) \leq \mathcal{E}(0) e^{-4\mu t}$.

Theorem 28 (Finite-Step GD, corrected smoothness constant). *If $\mathbf{R}_t \succeq \mu \mathbf{I}$ and $\eta \leq 1/\beta_t$ at every iterate, where*

$$\beta_t = 2\lambda_{\max}(\mathbf{R}_t) + 2 \sum_d |e_{dt}| \|\nabla^2 f_3(\mathbf{z}_{dt})\|_{\text{op}}$$

is the (corrected) local smoothness constant, then

$$\mathcal{E}_{t+1} \leq (1 - 2\eta\mu) \mathcal{E}_t.$$

Hence $\mathcal{E}_T \leq (1 - 2\eta\mu)^T \mathcal{E}_0 \rightarrow 0$.

Proof sketch. The first-order error update is $\mathbf{e}_{t+1} = (\mathbf{I} - 2\eta\mathbf{R}_t)\mathbf{e}_t + O(\eta^2)$, with the $O(\eta^2)$ remainder controlled by $\sum_d |e_{dt}| \|\nabla^2 f_3(\mathbf{z}_{dt})\|_{\text{op}}$. The descent lemma with this smoothness constant gives $\mathcal{E}_{t+1} \leq \mathcal{E}_t - \frac{\eta}{2} \|\nabla \mathcal{E}_t\|_F^2$ for $\eta \leq 1/\beta_t$ (numerically, $\eta \leq 1/(2\lambda_{\max}(\mathbf{R}_t))$) alone—dropping the curvature factor of 2—is *false*; the factor of 2 on both terms of β_t is needed). Since $\|\nabla \mathcal{E}\|_F^2 = 4\mathbf{e}^T \mathbf{R} \mathbf{e} \geq 4\mu \mathcal{E}$, the result follows. $\square$

Remark 29. The smoothness constant β_t includes a curvature term that depends on the current residuals $|e_d|$. Near convergence ($\mathbf{e} \rightarrow 0$), $\beta_t \rightarrow 2\lambda_{\max}(\mathbf{R})$, and the step-size condition simplifies. In the gradient-flow limit ($\eta \rightarrow 0$), this issue vanishes and Theorem 27 applies directly.

Remark 30. Both results require the eigenvalue bound *at every iterate*, not just at initialization. §7 shows how to obtain this—non-circularly, and without an unsatisfiable ball hypothesis.

7 Along the GD Trajectory

GD produces $\mathbf{W}_0, \mathbf{W}_1, \dots$ with $\mathbf{R}_t$ depending on $\mathbf{W}_t$. Theorems 27 and 28 need the eigenvalue floor at *every* iterate. Two approaches are conceivable: (A) bound how much $\mathbf{R}$ can change as $\mathbf{W}$ moves (a Lipschitz/bootstrap argument), or (B) treat GD as an adversary constrained by what it can *know* about f_3 at each step, and bound its worst case. We show (B) collapses *exactly* onto (A) (Proposition 36: limiting knowledge buys nothing), and then give the sharp, non-circular version of (A).

7.1 Spectral Lipschitz bound

Standing assumptions. (A1) $f_3 \in C^2$ with $H \stackrel{\text{def}}{=} \sup_{\mathbf{z}} \|\nabla^2 f_3(\mathbf{z})\|_{\text{op}} < \infty$. (A2) $\|\mathbf{x}_d\| = 1$; $\mathbf{S} = \mathbf{X}^T \mathbf{X}$. (A3) $\Gamma \stackrel{\text{def}}{=} \sup_{t \geq 0} \|\mathbf{G}_t\|_{\text{op}} < \infty$. Define

$$L_R \stackrel{\text{def}}{=} 2 H \Gamma \|\mathbf{X}\|_{\text{op}}.$$

Lemma 31 (Spectral Lipschitz Bound, Lemma II.1). *For all $\mathbf{W}, \mathbf{W}'$ with $\|\mathbf{G}(\mathbf{W})\|_{\text{op}}, \|\mathbf{G}(\mathbf{W}')\|_{\text{op}} \leq \Gamma$,*

$$\|\mathbf{R}(\mathbf{W}) - \mathbf{R}(\mathbf{W}')\|_{\text{op}} \leq L_R \|\mathbf{W} - \mathbf{W}'\|_F.$$

Proof. Since $\mathbf{S} \succeq 0$ with $S_{dd} = 1$, the Schur product bound $\|\mathbf{A} \odot \mathbf{S}\|_{\text{op}} \leq \|\mathbf{A}\|_{\text{op}} \max_d S_{dd}$ (Horn–Johnson, *Topics in Matrix Analysis*, Thm. 5.5.18) gives

$$\|\mathbf{R} - \mathbf{R}'\|_{\text{op}} \leq \|\mathbf{G}^T \mathbf{G} - \mathbf{G}'^T \mathbf{G}'\|_{\text{op}} \leq \|\Delta \mathbf{G}\|_{\text{op}} (\|\mathbf{G}\|_{\text{op}} + \|\mathbf{G}'\|_{\text{op}}) \leq 2\Gamma \|\Delta \mathbf{G}\|_{\text{op}}$$

with $\Delta \mathbf{G} = \mathbf{G} - \mathbf{G}'$. And

$$\|\Delta \mathbf{G}\|_{\text{op}} \leq \|\Delta \mathbf{G}\|_F = \left(\sum_d \|\nabla f_3(\mathbf{W} \mathbf{x}_d) - \nabla f_3(\mathbf{W}' \mathbf{x}_d)\|^2 \right)^{1/2} \leq H \|(\mathbf{W} - \mathbf{W}') \mathbf{X}\|_F \leq H \|\mathbf{W} - \mathbf{W}'\|_F \|\mathbf{X}\|_{\text{op}}.$$

$\square$

7.2 Trajectory bootstrap: gradient flow

Theorem 32 (Trajectory Bootstrap, Gradient Flow, Theorem II.2). *Let $\dot{\mathbf{W}} = -\nabla_{\mathbf{W}} \mathcal{E}$, $\ell_0 \stackrel{\text{def}}{=} \lambda_{\min}(\mathbf{R}_0)$. If*

$$\ell_0^{3/2} > \frac{3}{2} L_R \sqrt{\mathcal{E}_0},$$

then, with $\ell_\infty \stackrel{\text{def}}{=} (\ell_0^{3/2} - \frac{3}{2} L_R \sqrt{\mathcal{E}_0})^{2/3} > 0$:

- (i) $\lambda_{\min}(\mathbf{R}_t) \geq \ell_\infty$ for all $t \geq 0$;
- (ii) $\mathcal{E}(t) \leq \mathcal{E}_0 e^{-4\ell_\infty t}$;
- (iii) $\|\mathbf{W}_t - \mathbf{W}_0\|_F \leq \sqrt{\mathcal{E}_0/\ell_\infty}$ for all t .

Proof. Let $T^* \stackrel{\text{def}}{=} \inf\{t : \lambda_{\min}(\mathbf{R}_t) < \ell_\infty\}$ and work on $[0, T^*)$. Three identities: $\|\nabla \mathbf{W} \mathcal{E}\|_F^2 = 4 \mathbf{e}^T \mathbf{R} \mathbf{e}$ (Theorem 27); $\dot{\mathcal{E}} = -\|\nabla \mathcal{E}\|_F^2 = -4 \mathbf{e}^T \mathbf{R} \mathbf{e}$; hence $\|\dot{\mathbf{W}}\|_F = \sqrt{-\dot{\mathcal{E}}}$. Let $m(t) \stackrel{\text{def}}{=} \int_0^t \|\dot{\mathbf{W}}\|_F ds$. By Lemma 31,

$$\lambda_{\min}(\mathbf{R}_t) \geq \ell_0 - L_R m(t) =: \lambda(t), \quad -\dot{\mathcal{E}} \geq 4\lambda \mathcal{E}.$$

On $[0, T^*)$, $\mathcal{E}$ is strictly decreasing, so reparametrise by $\mathcal{E}$:

$$\left| \frac{dm}{d\mathcal{E}} \right| = \frac{1}{\sqrt{-\dot{\mathcal{E}}}} \leq \frac{1}{2\sqrt{\lambda \mathcal{E}}} \implies \frac{d\lambda}{d\mathcal{E}} \leq \frac{L_R}{2\sqrt{\lambda \mathcal{E}}}.$$

Separating, $\sqrt{\lambda} d\lambda \leq \frac{1}{2} L_R \mathcal{E}^{-1/2} d\mathcal{E}$, and integrating from $\mathcal{E}$ up to $\mathcal{E}_0$ (where $\lambda = \ell_0$):

$$\frac{2}{3}(\ell_0^{3/2} - \lambda(\mathcal{E})^{3/2}) \leq L_R(\sqrt{\mathcal{E}_0} - \sqrt{\mathcal{E}}) \implies \lambda(\mathcal{E})^{3/2} \geq \ell_0^{3/2} - \frac{3}{2} L_R \sqrt{\mathcal{E}_0} = \ell_\infty^{3/2}.$$

So $\lambda_{\min}(\mathbf{R}_t) \geq \lambda(t) \geq \ell_\infty$ throughout $[0, T^*)$, with strict inequality while $\mathcal{E} > 0$; by continuity $T^* = \infty$. Part (ii) is Grönwall on $\dot{\mathcal{E}} \leq -4\ell_\infty \mathcal{E}$. Part (iii):

$$m(\infty) = \int_0^{\mathcal{E}_0} \frac{d\mathcal{E}}{\sqrt{-\dot{\mathcal{E}}}} \leq \int_0^{\mathcal{E}_0} \frac{d\mathcal{E}}{2\sqrt{\ell_\infty \mathcal{E}}} = \sqrt{\mathcal{E}_0/\ell_\infty}. \quad \square$$

Remark 33 (Comparison with the old bootstrap). Setting $\mu = \ell_0/2$, the old (circular) bootstrap's condition $L_R B \leq \mu$ with the loose budget $B = 2\sqrt{D\bar{\alpha}\mathcal{E}_0}/\mu$ required $L_R \sqrt{\mathcal{E}_0} \lesssim 0.05 \ell_0^{3/2}/\sqrt{D\bar{\alpha}}$, and even with the sharp budget $B = 2\sqrt{\mathcal{E}_0/\mu}$ it required $L_R \sqrt{\mathcal{E}_0} \leq \ell_0^{3/2}/(4\sqrt{2}) = 0.177 \ell_0^{3/2}$. Theorem 32 requires only $L_R \sqrt{\mathcal{E}_0} < \frac{2}{3} \ell_0^{3/2} = 0.667 \ell_0^{3/2}$: **a factor 3.8 better, with no circularity and no ball hypothesis to satisfy.**

7.3 Discrete GD

Proposition 34 (Telescoping Movement Budget). *Under the hypotheses of Theorem 28 with $\mathbf{R}_t \succeq \mu \mathbf{I}$ and $\eta_t \leq 1/\beta_t$, the descent inequality $\mathcal{E}_t - \mathcal{E}_{t+1} \geq \frac{\eta}{2} \|\nabla \mathcal{E}_t\|_F^2$ combined with $\|\nabla \mathcal{E}_t\|_F \geq 2\sqrt{\mu \mathcal{E}_t}$ gives, via $(a-b)/\sqrt{a} \leq 2(\sqrt{a} - \sqrt{b})$,*

$$\|\Delta \mathbf{W}_t\|_F = \eta \|\nabla \mathcal{E}_t\|_F \leq \frac{\eta \|\nabla \mathcal{E}_t\|_F^2}{2\sqrt{\mu \mathcal{E}_t}} \leq \frac{\mathcal{E}_t - \mathcal{E}_{t+1}}{\sqrt{\mu \mathcal{E}_t}} \leq \frac{2}{\sqrt{\mu}} (\sqrt{\mathcal{E}_t} - \sqrt{\mathcal{E}_{t+1}}),$$

which telescopes to

$$\boxed{\sum_t \|\Delta \mathbf{W}_t\|_F \leq 2\sqrt{\mathcal{E}_0/\mu} \stackrel{\text{def}}{=} B.}$$

This replaces the earlier triangle-inequality budget $B = 2\sqrt{D\bar{\alpha}\mathcal{E}_0}/\mu$, which was $\sim 77\times$ looser at the paper's typical parameters (the earlier proof bounded $\|\Delta \mathbf{W}_t\|_F$ termwise via Cauchy-Schwarz rather than telescoping the descent inequality).

Theorem 35 (Discrete Trajectory Bootstrap, Theorem II.3). *Under the corrected smoothness constant of Theorem 28, if*

$$\ell_0^{3/2} > 3 L_R \sqrt{\mathcal{E}_0}$$

then the same ODE-comparison argument as Theorem 32 (with the stopping time T^* now over iterates, and Proposition 34 in place of the flow’s arc-length bound) gives $\lambda_{\min}(\mathbf{R}_t) \geq \ell_\infty > 0$ at every iterate t , hence $\mathcal{E}_t \leq \mathcal{E}_0(1 - 2\eta\ell_\infty)^t \rightarrow 0$ and $\sum_t \|\Delta \mathbf{W}_t\|_F \leq B = 2\sqrt{\mathcal{E}_0/\ell_\infty}$. The constant 3 (vs. $\frac{3}{2}$ for flow) is the price of discretisation.

7.4 Width accounting

Normalise $\bar{\alpha} = \Theta(1)$ and $f_3 = O(1)$, so for a GP/RFF readout of correlation length σ : $\bar{\alpha} \approx M'/\sigma^2 = 1 \Rightarrow \sigma = \sqrt{M'}$, and $H \approx \sqrt{M'}/\sigma^2 = c_H/\sqrt{M'}$. With $\|\mathbf{X}\|_{\text{op}} \approx \sqrt{D/M}$, $\Gamma \approx \sqrt{\bar{\alpha}D/M'}$, $\mathcal{E}_0 \approx c_E D$, $\ell_0 = \text{MP} \cdot \bar{\alpha}$:

$$L_R \approx \frac{2c_H D}{M'\sqrt{M}}, \quad \ell_0^{3/2} > \frac{3}{2} L_R \sqrt{\mathcal{E}_0} \iff \boxed{M'^2 M \gtrsim C D^3 / \text{MP}^3}.$$

Minimising MM' subject to $M'^2 M \geq C D^3$ gives $MM' \geq \sqrt{C M_{\min}} D^{3/2}$, so with the anti-crowding requirement $M \gtrsim \log D$:

$$\textbf{Worst-case } L_R : \quad MM' \gtrsim D^{3/2} \sqrt{\log D}.$$

At a fixed aspect ratio $M' \propto M$ it degrades to $MM' \gtrsim D^2$.

[NUMERICAL] Fitting the margin $\ell_0^{3/2}/(\frac{3}{2}L_R\sqrt{\mathcal{E}_0})$ over 13 configurations ($M \in [8, 32]$, $M' = 10M$, $D \in [50, 400]$): $\text{margin}_{L_R} \propto (MM')^{1.28} D^{-2.07} \Rightarrow \text{margin} = 1$ at $MM' \approx 62 D^{1.62}$, consistent with the analytic $D^{3/2}$ – D^2 bracket. **This already beats Du et al.’s D^4 – D^8 by a wide margin, and it is unconditional given the static bound of §4.**

7.5 Adversary equivalence: limiting knowledge buys nothing

Formalise three adversary classes with total budget $B = \sum_t \|\Delta \mathbf{W}_t\|_F$:

- **(B1) Oblivious-uniform:** bound $\sup_{\|\mathbf{W} - \mathbf{W}_0\|_F \leq B}$ of the failure event.
- **(B2) $\Delta \mathbf{W}$ -adaptive:** $\Delta \mathbf{W}_t$ is $\mathcal{F}_t$ -measurable, $\mathcal{F}_t = \sigma(\mathbf{W}_0, \mathbf{X}, \Delta \mathbf{W}_0, \dots, \Delta \mathbf{W}_{t-1})$.
- **(B3) Full information:** $\Delta \mathbf{W}_t$ may depend on f_3 arbitrarily.

Proposition 36 (Adversary Equivalence, Proposition II.4). *The set of $\mathbf{W}_T$ reachable by a $\Delta \mathbf{W}$ -adaptive adversary with budget B is exactly $\{\mathbf{W} : \|\mathbf{W} - \mathbf{W}_0\|_F \leq B\}$ —take $T = 1$ and $\Delta \mathbf{W}_0$ any matrix of norm $\leq B$, which is trivially $\mathcal{F}_0$ -measurable. Hence*

$$\sup_{(B2) \text{ strategies}} \mathbb{P}[\text{failure}] = \sup_{(B1)} \mathbb{P}[\text{failure}],$$

and any critical-point-of- f_3 attack applies verbatim to (B2).

Remark 37. This is the direct answer to the question of whether restricting what the adversary *knows* about f_3 , step by step, can help. It cannot: the intuition that “ $\Delta \mathbf{W}_t$ is a scrambled sum, so the adversary learns little about individual $\mathbf{g}_d$ ” is true (a data-processing inequality) but does no work, because the adversary does not need to know anything about f_3 to fail—it only has to

walk far enough. (Note: every smooth f_3 has critical points at spacing $\sim \sigma$, where $\alpha_d = 0$; but the bound $\lambda_{\min}(\mathbf{R}) \geq c\alpha_{\min}$ is *trivially satisfied* at such points since $\alpha_{\min} = 0$ makes it vacuous, not false. The paper assumes throughout that the operator maintains $\alpha_{\min} > 0$.) **The obstruction is geometric, not informational.** Restricting the adversary’s σ -algebra cannot help; only restricting B can. In particular the earlier claim—“if no adversary with $(\mathbf{W}_0, \mathbf{X}, \Delta\mathbf{W}_{0:T-1})$ -information can cause gradient alignment, then GD certainly cannot”—is valid as an implication (GD *is* such a strategy) but its premise requires careful analysis: the reachable set $\|\mathbf{W} - \mathbf{W}_0\|_F \leq B$ does not contain $\mathbf{W} = 0$ when $B < \|\mathbf{W}_0\|_F$, and the ball conjecture remains open—see §8 for a precise analysis of why cell-based approaches nevertheless fail. (When $B < \|\mathbf{W}_0\|_F \approx \sqrt{M'}$, the point $\mathbf{W} = 0$ lies outside the ball, so the earlier claim that “ $\mathbf{W}_T = 0$ kills the ball conjecture” was incorrect.) The correct replacement for that paragraph is exactly Theorem 35, with the budget attached.

7.6 When the ball formulation works

An oblivious-uniform bound over a ball is only as good as the ball is small relative to the correlation length σ . With $\bar{\alpha} \approx M'/\sigma^2$, $\mu = c\bar{\alpha}$, $\mathcal{E}_0 \approx c_E D$, the sharp budget of Proposition 34 gives

$$B = 2\sqrt{\mathcal{E}_0/\mu} = 2\sigma\sqrt{\frac{c_E D}{c M'}} < \sigma \iff \boxed{M' > \frac{4c_E}{c} D.}$$

The ball formulation is not dead in general—it is dead exactly when $M' = o(D)$. For $M' = \Omega(D)$, the entire trajectory is confined inside one correlation length and a uniform-over-a-ball statement is valid; this is precisely the “lazy regime” in which Du et al. and Arora et al. operate. Since the paper’s target regime is $MM' = \Theta(D)$ with $M' \ll D$, the ball formulation must be replaced by the trajectory bootstrap of Theorems 32–35—but stating the boundary quantitatively, rather than declaring the ball approach simply dead, is more useful and tells us exactly what “lazy” means in this problem. [NUMERICAL]: measured ($M=5, M'=50, D=120, \sigma = 2.36$), GD travelled 1.10–1.28 correlation lengths—just outside the ball’s safe zone, consistent with $M' \ll D$ here.

7.7 The binding gap in L_R

[NUMERICAL] The worst-case $L_R = 2H\Gamma\|\mathbf{X}\|_{\text{op}}$ used above is loose. Auditing the inequality chain

$$|\dot{\lambda}_{\min}| \leq \|\dot{\mathbf{R}}\|_{\text{op}} \leq \|\dot{\mathbf{G}}^T \mathbf{G} + \mathbf{G}^T \dot{\mathbf{G}}\|_{\text{op}} \leq 2\|\dot{\mathbf{G}}\|_{\text{op}} \|\mathbf{G}\|_{\text{op}} \leq 2\|\dot{\mathbf{G}}\|_F \|\mathbf{G}\|_{\text{op}} \leq 2H\|\dot{\mathbf{W}}\mathbf{X}\|_F \|\mathbf{G}\|_{\text{op}} \leq L_R \|\dot{\mathbf{W}}\|_F$$

step by step ($M=8, M'=80, D=200$, RFF $\tau = 3$; $L_R = 262.7$) shows a cumulative slack of $\times 4800$ to $\times 20700$ depending on direction, while a direct search for the worst local Lipschitz direction attains only $L_{\text{att}} = 0.730$ —so L_R is loose by a factor between 40 (identified, individually fixable, without any search) and 360 (from the search, an optimistic upper bound on what any sharp argument can deliver). The three largest, individually addressable sources of slack, in order, are:

- (a) $\|\Delta\mathbf{G}\|_{\text{op}}$ vs. $\|\Delta\mathbf{G}\|_F$ ($\times 4.8$ measured)—needs a matrix-concentration/decoupling argument exploiting near-independence of the Hessians $\mathbf{H}_d = \nabla^2 f_3(\mathbf{z}_d)$ at well-separated $\mathbf{z}_d$;
- (b) a bilinear Schur bound $\|(\mathbf{A}^T \mathbf{B}) \odot \mathbf{S}\|_{\text{op}}$ using $\|\mathbf{S}\|_{\text{op}} \approx D/M$ rather than the crude $\|\mathbf{A}^T \mathbf{B}\|_{\text{op}}$ ($\times 3.3$ measured);
- (c) **bottom-eigenvector delocalisation** ($\times 2.6$ measured, $\times \sqrt{M/\log D}$ asymptotically): with $\mathbf{u}$ the bottom eigenvector of $\mathbf{R}_t$ and $\mathbf{N} = (\mathbf{u}\mathbf{u}^T) \odot \mathbf{S}$, the exact identity $\dot{\lambda}_{\min} = 2\langle \dot{\mathbf{G}}, \mathbf{G}\mathbf{N} \rangle_F$ gives a refined constant $L_R^{\text{ref}} = L_R \cdot \|\mathbf{N}\|_F$; a k -sparse $\mathbf{u}$ would force $\mathbf{u}^T \mathbf{R} \mathbf{u} \geq \lambda_{\min}(\mathbf{R}_{TT})$ on the

corresponding principal submatrix, which is well-conditioned when those $\mathbf{z}_d$ are separated—so the same sparsity-peeling technique needed for the static bound (Gap 2 of §5) would prove (c) as well.

Closing (a)(b)(c) alone (factor ≈ 40 , no search required) moves the worst-case width from $MM' \approx 62 D^{1.62}$ to $\approx 4 D^{1.62}$; the optimistic limit L_{att} moves it to $\approx 108 D^{0.90}$. **The upper end of that bracket is $\Theta(D)$, the paper’s target, reached by a purely deterministic matrix inequality—no covering argument, no freshness, no probability. Narrowing this bracket is the single highest-value calculation in the program.**

Remark 38 (Two routes not pursued). **Conditional freshness** (querying f_3 only where it is unqueried) requires step size $\varepsilon \gtrsim \sigma$ for the Lipschitz-safe step $\varepsilon = \mu/(2L_R)$, which works out to $MM' = \Omega(D^2)$ —worse than the bootstrap above, because GD’s queries move continuously and are never truly fresh. A **martingale refinement** using the drift structure of $\dot{\lambda}_{\min}$ for stationary kernels gives at best a $D^{1/8}$ gain at the target widths. Neither is worth building the proof on.

8 Ball Conjecture: Fundamental Limits

The *ball conjecture* asks: for all $\Delta \mathbf{W}$ with $\|\Delta \mathbf{W}\|_F \leq L$, do the gradients $\mathbf{g}_d(\mathbf{W}_0 + \Delta \mathbf{W})$ remain decorrelated? We show that no cell-based proof strategy can establish this outside the lazy regime. The analysis is due to Edmonds.

8.1 The cube construction

Consider f_3 defined as follows: partition $\mathbb{R}^{M'}$ into axis-aligned cubes of side r . For each cube C , draw an independent random gradient $\mathbf{g}_C \in \mathbb{R}^{M'}$. For $\mathbf{z} \in C$, set $f_3(\mathbf{z})$ to a linear function with gradient $\mathbf{g}_C$.

At initialization ($\mathbf{W} = \mathbf{W}_0$), each $\mathbf{z}_d = \mathbf{W}_0 \mathbf{x}_d$ lands in a random cube. Three cases arise, depending on the cube side r :

Case Big (r too large): Two data points $\mathbf{z}_d, \mathbf{z}_{d'}$ land in the same cube, forcing $\mathbf{g}_d = \mathbf{g}_{d'}$. By a birthday argument in M' dimensions, this happens when $r \gtrsim C_1 \sigma_W / D^{2/M'}$, where $\sigma_W = 1/\sqrt{M}$ is the initialization scale.

Case Small (r too small): At initialization, $\mathbf{z}_d$ and $\mathbf{z}_{d'}$ are far apart—different cubes, independent gradients. But moving $\mathbf{W}$ by $\Delta \mathbf{W}$ can move $\mathbf{z}_d$ through many cubes. The adversary tries all $(2L/r)^{2M'}$ pairs of destination cubes for $(\mathbf{z}_d, \mathbf{z}_{d'})$, each with an independent random gradient. With enough pairs, one will have $\mathbf{g}_d \approx \mathbf{g}_{d'}$.

Matrix concentration (Theorem 5) tolerates at most $q \leq e^{cM'/D}$ configurations before the union bound fails. So we need $(2L/r + 1)^{M'} < e^{cM'/D}$, giving $r > \Omega(DL)$.

Case Middle (the attempted sweet spot): We want cubes small enough that each $\mathbf{z}_d$ starts in a different cube (Case Big avoided) but large enough that no $\mathbf{z}_d$ leaves its cube under any $\|\Delta \mathbf{W}\| \leq L$ (Case Small avoided).

Theorem 39 (Cube Impossibility). *For any cube side r , the constraints from Cases Big and Small are*

$$\underbrace{\Omega(DL)}_{\text{Case Small}} < r < \underbrace{O(\sigma_W)}_{\text{Case Big}},$$

which requires $L < O(\sigma_W/D) = O(1/(\sqrt{M}D))$. The GD movement budget is $L \approx 2\sqrt{\mathcal{E}_0/\alpha_{\min}} = \Theta(\sqrt{D})$, independent of M and M' . Hence no cube side r makes the construction work outside the lazy regime $L \ll \sigma_W$.

Remark 40. The factor D in the denominator comes from the union bound over D data points consuming the matrix concentration exponent $e^{-cM'}$: both the concentration exponent and the configuration count $(2L/r)^{M'}$ scale as $e^{\Theta(M')}$, so M' cancels, leaving the D -dependent constraint $\ln(2L/r) < c/D$. Increasing M' does not help.

8.2 Correlated cubes

A natural modification: instead of independent gradients per cube, impose a Lipschitz condition. If $\|\mathbf{z}_1 - \mathbf{z}_2\| \leq \ell$ (the correlation length), then $\|\mathbf{g}_1 - \mathbf{g}_2\| \leq \ell'(\ell)$ with $\lim_{\ell \rightarrow 0} \ell'(\ell) = 0$.

This kills Case Small: moving to a neighboring cube gives a *correlated* gradient, not an independent one. The adversary gets no fresh lottery tickets from small movements.

Theorem 41 (Correlated Field Impossibility). *For any Lipschitz random field on $\mathbb{R}^{M'}$ with correlation length ℓ_c (gradients approximately independent beyond distance ℓ_c , approximately equal within distance ℓ_c):*

- Decorrelation at initialization requires $\ell_c \ll \|\mathbf{z}_d - \mathbf{z}_{d'}\|_{\text{init}} \approx \sqrt{2M'/M}$.
- The adversary can bring $\mathbf{z}_d$ and $\mathbf{z}_{d'}$ within distance ℓ_c by spending $\|\Delta\mathbf{W}\| \approx \ell_c\sqrt{M/M'}$ (aligning along $\mathbf{x}_d - \mathbf{x}_{d'}$).
- Prevention requires $L < \ell_c\sqrt{M/M'} < \sqrt{2}$, i.e., $L = O(1)$: the lazy regime again.

Proof sketch. When $M \geq D$, the map $\Delta\mathbf{W} \mapsto (\Delta\mathbf{z}_1, \dots, \Delta\mathbf{z}_D)$ has full row rank, so the adversary can steer each $\Delta\mathbf{z}_d$ essentially independently with budget $L/\sqrt{D}$ per pair. To bring one pair within ℓ_c : cost $\approx \sqrt{M'/M} \cdot \ell_c$. With L total budget, the adversary concentrates $O(L\sqrt{M}/(M'\ell_c))$ pairs. This exceeds D when $L \gg \ell_c\sqrt{M'D/M}$, at which point all gradients are correlated. $\square$

The deep geometric reason. Both the cube and correlated-field impossibilities reduce to the same barrier: in the $M \geq D$ regime, the adversary has more degrees of freedom in $\Delta\mathbf{W}$ (MM' entries) than there are data constraints (D points), so it can rearrange the pre-activations $\mathbf{z}_d$ nearly at will. Any property of f_3 alone—cubes, smooth fields, Gaussian processes—cannot prevent this rearrangement. The protection must come from either:

1. $M < D$ (the adversary is *constrained*—see §8.3), or
2. The trajectory (GD has no reason to concentrate pre-activations—see Theorems 32/35).

8.3 Constrained adversary at $M < D$

When $M < D$ and $MM' = \Theta(D)$, the adversary's degrees of freedom are limited: $\Delta\mathbf{W} \in \mathbb{R}^{M' \times M}$ has $MM' = O(D)$ entries, but steering each $\Delta\mathbf{z}_d \in \mathbb{R}^{M'}$ independently requires DM' degrees of freedom. The movements are **coupled**.

Lemma 42 (Constrained Pair Budget). *With total budget $\|\Delta\mathbf{W}\|_F \leq L$ and $M < D$, the adversary can bring at most $k = O(L\sqrt{M/M'})$ pairs of pre-activations within correlation length ℓ_c of each other. When $M = M' = \sqrt{D}$: $k = O(L)$.*

Proof sketch. Each pair requires aligning $\Delta \mathbf{z}_d - \Delta \mathbf{z}_{d'} = \Delta \mathbf{W}(\mathbf{x}_d - \mathbf{x}_{d'})$ along an M' -dimensional direction. The M -dimensional constraint on $\Delta \mathbf{W}$ limits how many such alignments can be achieved simultaneously. $\square$

The damage from these corrupted pairs is controlled by eigenvector delocalisation:

Lemma 43 (Delocalised Damage Bound). *Let $\mathbf{R}$ have delocalised bottom eigenvector: $\max_d |v_{d,\min}|^2 = O(1/D)$. Then corrupting k gradient pairs (making $\mathbf{g}_d \approx \mathbf{g}_{d'}$) changes $\lambda_{\min}(\mathbf{R})$ by at most*

$$|\Delta \lambda_{\min}| \leq k \cdot O\left(\frac{\alpha_{\max}}{D}\right).$$

Proof sketch. By first-order perturbation theory (Davis–Kahan), $|\Delta \lambda_{\min}| \leq \|v_{\min}\|_S^2 \cdot \|\Delta \mathbf{R}\|_{\text{op}}$ restricted to the support of the perturbation. Each corrupted pair (d, d') modifies two rows/columns. Delocalisation bounds $v_{d,\min}^2 + v_{d',\min}^2 = O(1/D)$. $\square$

Corollary 44 (Ball Conjecture at $M < D$). *If $k = O(\sqrt{D})$ pairs are corrupted and eigenvectors are delocalised, then the total damage is*

$$|\Delta \lambda_{\min}| \leq O\left(\frac{\alpha_{\max}}{\sqrt{D}}\right) \rightarrow 0 \quad \text{as } D \rightarrow \infty.$$

The remaining $D - O(\sqrt{D})$ uncorrupted pairs contribute the standard concentration floor $\lambda_{\min} \geq c \alpha_{\min}$, which dominates.

Remark 45. This chain—constrained adversary + delocalised eigenvectors—connects three of the paper’s ideas: the ball conjecture, the constrained movement budget, and the “Quantum Reality” delocalisation phenomenon. They were always the same theorem; this analysis makes the connection explicit.

Status. Lemmas 42 and 43 are proof sketches, not formal theorems. Formalising the constrained pair budget requires a careful analysis of the rank of the map $\Delta \mathbf{W} \mapsto (\Delta \mathbf{z}_1, \dots, \Delta \mathbf{z}_D)$, and the delocalisation bound requires proving that Khatri–Rao eigenvectors are indeed delocalised (related to the paper’s existing conjecture about randomly rotated eigenvectors). Both are concrete, attackable problems.

9 Related Work and Literature Positioning

The current frontier. The best existing width condition for two-layer, all-layer, discrete GD convergence is due to Song, Ramezani-Kebrya, Pethick, Eftekhari, and Cevher (NeurIPS 2021, arXiv:2111.01875): $\tilde{O}(D^{3/2})$ total parameters (width $\tilde{O}(D^{3/2}/d)$ where d is input dimension), for smooth activations. They tighten the standard NTK perturbation argument by bounding the spectral norm of the NTK perturbation directly, rather than entry-by-entry.

Other key results.

- **Montanari–Zhong (2022):** Width $\tilde{O}(D/d)$, total parameters $\tilde{O}(D)$. But trains only the first layer (second fixed to ± 1), requires gradient flow (not discrete GD), and spherical data. Does not apply to our setting.

- **Bombari, Amani, Mondelli (NeurIPS 2022):** Width $\tilde{O}(\sqrt{D})$ per layer, total $\tilde{O}(D)$. But requires depth $L \geq 3$. Does not apply to two-layer networks.
- **Arora, Du, Hu, Li, Wang (ICML 2019):** Width $\Omega(D^4)$ to $\Omega(D^8)$. Now superseded by Song et al.
- **Banerjee et al. (UAI 2023):** NTK positivity at initialization with $\tilde{O}(D)$ width, any constant depth. Static result only—no trajectory proof.

Our positioning.

Method	Total params	Status	Limitations
Song et al. 2021	$\tilde{O}(D^{3/2})$	Proved	Smooth act., large input dim
Our Lipschitz bootstrap (worst-case L_R)	$O(D^{3/2}\sqrt{\log D})$	Proved	Requires static independence
Our Lipschitz bootstrap (random f_3)	$O(D^2)$	Estimated	L_R tightening unproved
Our Lipschitz bootstrap (with cancellation)	$\Theta(D)$	Conjectured	Independence assumption fails at boundary

The genuinely open problem—two-layer, all-layer, discrete GD at $\Theta(D)$ width—remains unproved by anyone. Our framework identifies the exact bottleneck (the Lipschitz constant L_R , or equivalently Criterion C of the gradient independence property) and provides a concrete path through it.

10 The Gradient Independence Conjecture

Conjecture 46 (Gradient Independence Along the GD Trajectory (C4)). *Let $f_3 : \mathbb{R}^{M'} \rightarrow \mathbb{R}$ be drawn from a law satisfying Criteria A–C of §5.5 (e.g., a GP with bandwidth σ as in Definition 25, or an RFF readout as in Definition 24), with*

$$\min(F, M') \cdot M \geq D$$

(Corollary 8). *Let $\mathbf{X} \in \mathbb{R}^{M \times D}$ with $\|\mathbf{x}_d\| = 1$ and $\|\mathbf{X}\mathbf{X}^T\|_{\text{op}} \leq C_X D/M$. Assume the target is **realizable**: there exists $\mathbf{W}^*$ with $f_3(\mathbf{W}^* \mathbf{x}_d) = y_d^*$ for all d (or, more generally, $\mathcal{E}_{\min} \stackrel{\text{def}}{=} \min_{\mathbf{W}} \mathcal{E}(\mathbf{W})$ is small).*

Then with high probability over f_3 and $\mathbf{W}_0$: the gradients $\mathbf{g}_{dt} = \nabla_{\mathbf{z}} f_3(\mathbf{W}_t \mathbf{x}_d)$ satisfy (σ_g, κ_t) -approximate independence (Definition 4) with $\kappa_t = O(1/D)$ at every GD iterate $\mathbf{W}_t$, provided $MM' \geq C \frac{\bar{\alpha}}{\alpha_{\min}} D$.

Remark 47. In the main paper this is **Conjecture C4** (the Gradient Independence Conjecture). Conjectures C1–C3 concern the Marchenko–Pastur eigenvalue distribution, per-point rate persistence, and ghost absorption, respectively. C4 is the proposed path to the spectral floor $\lambda_{\min}(\mathbf{R}_t) \geq c > 0$ that the other conjectures implicitly require.

Consequence. Under this conjecture, Theorem 5 gives: for $MM' \geq C \frac{\bar{\alpha}}{\alpha_{\min}} D$,

$$\lambda_{\min}(\mathbf{R}) \geq \frac{1}{2} \alpha_{\min} > 0 \quad \text{at every iterate,}$$

and by Theorems 32/35, this floor is self-sustaining along the entire trajectory whenever $\ell_0^{3/2} > 3L_R \sqrt{\mathcal{E}_0}$, so GD converges: $\mathcal{E}_T \leq (1 - \eta \alpha_{\min})^T \mathcal{E}_0 \rightarrow 0$.

Evidence for.

1. *GP case*: Proposition 26 proves independence at a single $\mathbf{W}$ under the separation condition $\Delta_{\min} \geq c\sigma\sqrt{\log D}$.
2. *RFF readout case*: the RFF architecture (Definition 24, §5.9) achieves the Marchenko–Pastur ideal $\lambda_{\min}(\mathbf{R})/\bar{\alpha} \approx 0.80\text{--}0.84 \times \text{MP}$ in simulations—the best of any architecture tested (Table 5).
3. *Architecture criteria hold constructively*: Criterion A (Corollary 8) and Criterion B (Proposition 16, Table 1) are proved theorems for the RFF/GP family; only Criterion C (decorrelation) remains conjectural (§5, Gap 2).
4. *Realizable-target conservation*: [NUMERICAL] with **realizable** targets, $\lambda_{\min}(\mathbf{R}_t)/\bar{\alpha}_t$ is essentially *conserved* along GD: +0.4% to +5% for travel up to 0.84 correlation lengths, and −8% to −35% at 1.1–1.3 lengths, with clean convergence throughout ($MM'/D = 2.1$ in the tested configuration).
5. *Initialization*: at $\mathbf{W}_0$, the $\mathbf{z}_d = \mathbf{W}_0\mathbf{x}_d$ are generically distinct, and for complex f_3 , gradients at distinct inputs carry approximately independent directional information.

Evidence against / cautions.

1. **Monotone-activation networks fail on the DC ratio and decorrelation, not on collinearity.** The sigmoid architecture (Definition 23, corrected in §5.8) has $\gamma \approx 0.95$ at moderate τ and a separate decorrelation failure ($f_2/\text{MP} \approx 0.47$); tanh, softplus, and other monotone, saturating activations fail similarly but less severely.
2. **Quadratic/low-Hermite-degree f_3 fails exactly** (Proposition 18): Hermite mass at degree ≤ 1 gives an affine gradient field and $\text{rank}(\mathbf{R}) \leq (M+1)M$, regardless of randomisation.
3. **Crowding in the low- M regime.** The $\mathbf{z}_d$ lie in an M -dimensional subspace, so for large D/M , some pair of $\mathbf{z}_d$ must be close, and Lipschitz gradients force alignment. The conjecture implicitly requires M large enough that separations are maintained (§5, Gap 3).
4. **Trajectory drift is real but bounded, not sign-definite.** In simulations, $\lambda_{\min}(\mathbf{R}_t)/\bar{\alpha}$ changes by +5% to −35% over the course of training (see above); this is handled by the self-consistent bootstrap of Theorem 32, not by a claim that the drift is bounded below by zero (which is false).
5. **Unrealizable targets cause genuine collapse.** With targets outside the range of f_3 (e.g., scale 8 against an $O(1)$ -range RFF readout), residuals never decay, $\mathbf{W}$ runs away, and $\lambda_{\min}(\mathbf{R}_t)/\bar{\alpha}$ has been observed to collapse by 98.9% alongside genuine training stalls. This is an approximation failure masquerading as a conditioning failure, not a counterexample to C4 under realizability—hence the realizability hypothesis above.
6. **Ball-based uniformity is unsatisfiable at the target width.** An earlier version required independence uniformly over $\|\mathbf{W} - \mathbf{W}_0\| \leq B$ for arbitrary B ; §7 shows this is valid only when $M' = \Omega(D)$, which is precisely the regime the paper is trying to avoid. The trajectory formulation (Theorems 32/35) is required for $M' = o(D)$.

11 What This Approach Proves and Does Not Prove

Proves (given gradient independence at every iterate, or the Lipschitz bootstrap condition of §7):

1. Collective exponential decay: $\mathcal{E}_t \leq \mathcal{E}_0 (1 - 2\eta\mu)^t$ with $\mu = \Omega(\alpha_{\min})$, self-sustaining along the whole trajectory, non-circularly (Theorems 32, 35).
2. Bounded total parameter movement, $\sum_t \|\Delta \mathbf{W}_t\|_F \leq 2\sqrt{\mathcal{E}_0/\mu}$ (Proposition 34).
3. All eigenvalues of $\mathbf{R}_t$ stay bounded below.
4. Two of the three architectural criteria for gradient independence (effective rank, DC ratio) as clean theorems for the one-hidden-layer random readout family (§5).
5. That restricting an adversary’s knowledge of f_3 , as opposed to its movement budget, cannot improve any of the above bounds (Proposition 36).

Does NOT prove:

1. **Per-point rates.** Collective decay does not imply each e_d decays at rate α_d . Off-diagonal coupling can cause individual errors to temporarily increase.
2. **Managed-noise coupling.** Gradient independence does not directly imply the managed-noise hypothesis $|\xi_d| \leq \sqrt{\rho} S$. These are related but genuinely different conditions.
3. **Generalization.** Zero training error says nothing about test performance.
4. **Criterion C (decorrelation) as a theorem**, for either the static bound or its dynamic (Lipschitz) analogue—both remain conjectural pending the sharp Khatri–Rao lemma of §5, Gap 2.
5. $\Theta(D)$ **width**, unconditionally. The best unconditional result given the static bound is $MM' \gtrsim D^{3/2}\sqrt{\log D}$ (§7.4); $\Theta(D)$ is plausible but requires closing the Lipschitz-constant gap of §7.

12 Open Problems

Open Problem (1. Prove the Conjecture for the RFF Readout, in Hermite Language). Replace the earlier “sufficient random complexity” with a precise, checkable statement: the activation’s Hermite mass must lie at degree ≥ 2 (kills γ , Criterion B) *and* at degree $\gtrsim \log D / \log(1/\mu_X)$ (kills the pairwise correlation, Criterion C). This connects the problem to the information-exponent/leap literature, and subsumes both the quadratic- f_3 counterexample and the ridge/He₂ counterexample (Proposition 18) as the degree- ≤ 1 and rank-limited cases.

For the RFF architecture (Definition 24), prove that with high probability over the random frequencies ω_j , phases ϕ_j , and coefficients β_j : the gradients along the GD trajectory satisfy Criterion C (the missing static lemma, Gap 2 of §5). The RFF case avoids the forward–backward coupling that plagues monotone architectures (§5.8), and reduces to the GP case (Proposition 26) as $F \rightarrow \infty$. An end-to-end proof in the lazy regime ($M' = \Omega(D)$, §7) appears within reach and would already improve on Du et al.’s D^4 – D^8 width.

Open Problem (2. Prove the Sharp Spectral Lipschitz Constant). The trajectory bootstrap (Theorems 32, 35) is loose in L_R by a factor of 40–360 (§7, “the binding gap”). Find L^* within $O(1)$ of the attainable rate by proving the three recoverable pieces identified there: (a) a matrix-concentration bound on $\|\Delta \mathbf{G}\|_{\text{op}}$ exploiting near-independent Hessians at separated points; (b) a bilinear Schur bound using $\|\mathbf{S}\|_{\text{op}} \approx D/M$; and (c) bottom-eigenvector delocalisation via the same sparsity-peeling technique needed for the static bound (Open Problem 1). Closing this gap moves the provable width from $MM' \approx 62 D^{1.62}$ toward $MM' \approx 4 D^{1.62} - 108 D^{0.90}$, i.e. plausibly $\Theta(D)$ —and it is a purely deterministic matrix inequality, not a probabilistic statement.

Open Problem (3. Connection to Paper’s Conjectures). • **C1 (MP eigenvalue distribution):** gradient independence would *explain* the MP distribution of $\mathbf{R}$ ’s eigenvalues, since $\mathbf{Y}^T \mathbf{Y}$ with independent columns has MP spectrum. Criterion C directly controls the deviation from this ideal (Table 5).

- **C2 (per-point rate persistence):** gradient independence gives collective convergence. Per-point rates require additional diagonal dominance.
- **C3 (ghost absorption):** gradient independence addresses non-ghost eigenvalues. Ghost error ($k > 0$) requires separate treatment.

Open Problem (4. An Unconditional Intermediate Theorem at $MM' = O(D^{3/2})$). Given the static bound of §4 (Theorem 5), Theorem 35 plus the width accounting of §7.4 already prove, **unconditionally on the Lipschitz gap of Open Problem 2:** for an RFF readout satisfying Criteria A–B (proved) and Criterion C (assumed as the static bound), gradient flow and discrete GD converge at

$$MM' = O(D^{3/2} \sqrt{\log D})$$

(or $O(D^{3/2})$ at the natural aspect ratio $M \approx M' \approx D^{3/4}$, using the refined Lipschitz constant of §7(c)). This should be written up and published now, independently of whether $\Theta(D)$ is ever reached: it already substantially improves on Du et al.’s and Arora et al.’s $D^4 - D^8$.

13 Proposed Simulations

1. **Eigenvalue ratio along GD.** Run GD with $f_3 = \text{RFF readout}$ (Definition 24). At each iterate, compute $\lambda_{\min}(\mathbf{R}_t)/\bar{\alpha}_t$. Compare with the Marchenko–Pastur prediction MP. Track the monotone decrease and stabilization.
2. **Off-diagonal cancellation.** Track $\|\mathbf{G}_t^T \mathbf{G}_t - \text{diag}(\alpha_{1t}, \dots, \alpha_{Dt})\|_{\text{op}}/\bar{\alpha}_t$ along GD. This directly measures the off-diagonal contribution to $\mathbf{R}$. (Note: $\|\hat{\mathbf{G}}^T \hat{\mathbf{G}} - \mathbf{I}\|_{\text{op}} \geq 1$ identically when $D > M'$, so the normalized-gradient version is vacuous.)
3. **Eigenvalue tracking.** Plot $\lambda_{\min}(\mathbf{R}_t)$ and $\alpha_{\min}(t)$ vs. t .
4. **Width scaling.** Fix D , vary MM' . Plot the transition in $\lambda_{\min}(\mathbf{R}_0)$ vs. MM'/D .
5. **DC ratio and decorrelation sweep.** For each candidate activation, measure $\gamma_\varphi(\tau)$ (Table 1) and the two-factor decomposition f_1, f_2 (Table 5) across τ , and compare against the Hermite-coefficient predictions of §5.4.
6. **Realizability sweep.** For a fixed architecture, vary $\|\mathbf{W}^* - \mathbf{W}_0\|_F$ (realizable) and target scale (unrealizable) and track $\lambda_{\min}(\mathbf{R}_t)/\bar{\alpha}_t$ and $\mathcal{E}_t/\mathcal{E}_0$, replicating the reconciliation of §10.

7. **Lipschitz-slack audit.** Reproduce the step-by-step audit of §7 (the binding gap) at several widths, and track the attainable rate L_{att} found by local search against L_R .

Reproducibility. All numerical claims in §5 and §7 are reproducible from the scripts `q1a_gamma.py`, `q1b_arch.py`, `q1c_depth.py`, `q1d_decomp.py`, `q2_trajectory.py`, `q2b_slack.py`, `q2c_scaling.py`, `q2h_teacher.py` (Python 3.12, numpy 2.5.3, scipy 1.18.1; Gauss–Hermite quadrature uses a 40 001-point fine grid, not `hermegauss` above order ~ 100 , which overflows).

14 Summary

$$\begin{array}{ccc}
 \underbrace{f_3 \text{ satisfies Criteria A–C}}_{\text{Conj. 46}} & \Rightarrow \underbrace{\mathbf{g}_d \text{ approx. indep.}}_{\text{Grad. Indep.}} & \xrightarrow{\text{Thm 5}} \underbrace{\lambda_{\min}(\mathbf{R}_0) \geq c \alpha_{\min}}_{\text{Static Floor}} \\
 & \Downarrow \text{Thm 32/35} & \\
 \underbrace{\lambda_{\min}(\mathbf{R}_t) \geq \ell_{\infty} > 0 \ \forall t}_{\text{Trajectory Floor}} & \xrightarrow{\text{Thm 28}} & \underbrace{\mathcal{E}_t \rightarrow 0}_{\text{Convergence}}
 \end{array}$$

Width needed: proved unconditionally (Open Problem 4) at $MM' \geq O(D^{3/2}\sqrt{\log D})$; target $MM' \geq \Omega(D)$ is plausible pending Open Problem 2 (the Lipschitz-constant gap).

The gradient independence property is:

- Not a monolithic property of f_3 , but three separable conditions on its activation law—effective rank, DC ratio, and decorrelation (§5)—two of which are now theorems.
- Sufficient for convergence at width $\Omega(D)$ (rather than $\Omega(D^2)$ from incoherence alone), *if* the third condition (decorrelation) holds.
- Provable along the *entire GD trajectory*, non-circularly, via a self-consistent Lipschitz bootstrap (§7), not just at initialization.
- Testable in simulations, and empirically robust under realizable targets.
- Provable for GPs with well-separated queries, and best achieved in practice by RFF readouts; sigmoid and other monotone-activation architectures fail on the DC ratio and decorrelation, *not* on gradient collinearity.
- The ball conjecture is fundamentally limited by cell-based impossibility (§8): any partition-based proof strategy is confined to the lazy regime when $M \geq D$. The constrained adversary regime $M < D$ (§8.3) offers a potential escape.
- The current literature frontier is $\tilde{O}(D^{3/2})$ (Song et al. 2021); our Lipschitz bootstrap is competitive at $O(D^{3/2}\sqrt{\log D})$ and targets $\Theta(D)$.

The central research question: prove Criterion C (decorrelation) for an RFF readout (Definition 24) along the GD trajectory, and prove the sharp spectral Lipschitz constant L^* of §7—together these would establish gradient independence, and hence GD convergence, at width $MM' = \Theta(D)$.

A Summary of What Is Proved, Assumed, and Refuted

#	Statement	Status
I.1	$\mathbf{G} = \mathbf{W}'^T \mathbf{C}^T$; $\text{rank}(\mathbf{R}) \leq \min(F, M')M$	THEOREM
I.2	Necessary: $\min(F, M')M \geq D$	THEOREM (verified exactly)
I.3	Projection lemma $\lambda_{\min}(\mathbf{R}) \geq \lambda_{\min}((\mathbf{\Pi G})^T(\mathbf{\Pi G}) \odot \mathbf{S})$	THEOREM
I.4	Shared gradient components are free	THEOREM
I.5	$\mathbf{g}_d = \bar{\mathbf{v}} + \mathbf{h}_d$, $\bar{\mathbf{v}} = \sum_j \beta_j m_j \mathbf{w}_j$	THEOREM
I.6	$m = a_1/\tau$; $K_g(\rho) = \tau^{-2} \sum_k k a_k^2 \rho^{k-1}$	THEOREM
I.7	$\gamma = a_1^2 / \sum_k k a_k^2$	THEOREM
I.8	Randomising activation signs cannot reduce γ	THEOREM (5 d.p.)
I.9	$\gamma = 0 \not\Rightarrow \lambda_{\min} > 0$ (He ₂)	THEOREM (counterex.)
I.10	Homogeneous activations are scale-frozen	THEOREM
I.11	Dot-product kernels saturate at μ_X ; stationary do not	THEOREM+HEURISTIC
I.12	$\gamma_L = \prod_l \gamma_l$	HEURISTIC (mean-field; verified $L \leq 3$)
—	Different activation per node helps	FALSE (13% measured)
—	Sigmoid fails because gradients are collinear	FALSE (costs 2%)
II.1	$\ \mathbf{R}(\mathbf{W}) - \mathbf{R}(\mathbf{W}')\ \leq 2H\Gamma\ \mathbf{X}\ _{\text{op}}\ \mathbf{W} - \mathbf{W}'\ _F$	THEOREM
II.2	Flow bootstrap: $\ell_0^{3/2} > \frac{3}{2}L_R\sqrt{\mathcal{E}_0}$	THEOREM
II.3	Discrete bootstrap + telescoping budget $2\sqrt{\mathcal{E}_0/\mu}$	THEOREM
II.4	$\Delta\mathbf{W}$ -adaptive adversary = ball adversary	THEOREM
II.5	Ball formulation valid $\iff M' = \Omega(D)$	THEOREM (given scalings)
—	Drift of $\lambda_{\min}$ bounded below by 0	FALSE (−35% measured)
—	L_R tight	FALSE ($\geq 40\times\text{--}360\times$ loose)
—	Freshness route reaches $\Theta(D)$	FALSE ($\Omega(D^2)$)
III.1	Cube impossibility: no cube side r allows non-lazy ball proof	THEOREM (Edmonds)
III.2	Correlated field impossibility: Lipschitz correlation doesn't help when $M \geq D$	THEOREM
III.3	Constrained adversary at $M < D$: $O(\sqrt{D})$ corrupted pairs	SKETCH
III.4	Delocalised damage: corrupted pairs do $O(\alpha_{\max}/D)$ damage each	SKETCH
—	Cell-based ball proof at $\Theta(D)$ when $M \geq D$	IMPOSSIBLE
—	$\mathbf{W} = 0$ kills the ball conjecture	FALSE ($\mathbf{W}=0$ outside ball)
—	$\alpha_d = 0$ at a critical point falsifies $\lambda_{\min} \geq c\alpha_{\min}$	FALSE (bound is trivially true)