

Training the Middle Layer: The Eigenvalue Floor Along Gradient Descent

Notes, October 2026

Jeff Edmonds, York University*

1 Introduction and Summary

Model. We train only the middle layer of

$$\mathbf{a}_d \mapsto \mathbf{x}_d \stackrel{\text{def}}{=} f_1(\mathbf{a}_d) \in \mathbb{R}^M \mapsto \mathbf{z}_d \stackrel{\text{def}}{=} \mathbf{W}\mathbf{x}_d \in \mathbb{R}^{M'} \mapsto f_3(\mathbf{z}_d) \in \mathbb{R},$$

where f_1 and f_3 are fixed random maps and $\mathbf{W} \in \mathbb{R}^{M' \times M}$ is trained by gradient descent (GD) on $\mathcal{E} \stackrel{\text{def}}{=} \sum_{d=1}^D e_d^2$, with $e_d \stackrel{\text{def}}{=} f_3(\mathbf{z}_d) - y_d^*$. Write $\mathbf{g}_d \stackrel{\text{def}}{=} \nabla_{\mathbf{z}} f_3(\mathbf{z}_d) \in \mathbb{R}^{M'}$ and $\alpha_d \stackrel{\text{def}}{=} \|\mathbf{g}_d\|^2$ (the *sensitivity*), and normalize $\|\mathbf{x}_d\| = 1$.

The rate matrix. Since $\nabla_{\mathbf{W}} e_d = \mathbf{g}_d \mathbf{x}_d^T$,

$$\|\nabla_{\mathbf{W}} \mathcal{E}\|_F^2 = 4 \mathbf{e}^T \mathbf{R} \mathbf{e}, \quad \mathbf{R} \stackrel{\text{def}}{=} (\mathbf{G}^T \mathbf{G}) \odot (\mathbf{X}^T \mathbf{X}), \quad R_{dd'} = (\mathbf{g}_d^T \mathbf{g}_{d'}) (\mathbf{x}_d^T \mathbf{x}_{d'}).$$

$\mathbf{R}$ is the empirical NTK restricted to the $\mathbf{W}$ -block. It is the Gram matrix of the Khatri–Rao vectors $\mathbf{y}_d \stackrel{\text{def}}{=} \mathbf{x}_d \otimes \mathbf{g}_d \in \mathbb{R}^{MM'}$, so $\text{rank}(\mathbf{R}) \leq \min(MM', D)$.

Why a floor is everything. Along gradient flow, $\dot{\mathcal{E}} = -4 \mathbf{e}^T \mathbf{R} \mathbf{e}$, so a floor $\lambda_{\min}(\mathbf{R}_t) \geq \mu > 0$ for all t gives $\mathcal{E}_t \leq \mathcal{E}_0 e^{-4\mu t}$ (Lemma 2).

The easy case and the hard case.

- *Wide*, $M \geq D$. If the $\mathbf{x}_d$ are linearly independent, Schur’s bound (Lemma 3) gives, at every time,

$$\lambda_{\min}(\mathbf{R}_t) \geq \lambda_{\min}(\mathbf{X}^T \mathbf{X}) \alpha_{\min}(t),$$

a floor for free, as long as the sensitivities stay bounded below. (We always assume this; keeping activations responsive is the practitioner’s job.)

- $M < D \leq MM'$. Now $\mathbf{X}^T \mathbf{X}$ is singular, and the gradients $\mathbf{g}_d$ must supply the missing dimensions. Everything depends on how the $\mathbf{g}_d$ evolve during training.

*Drafted with substantial help from Claude (an AI assistant). Every statement carries a status label; see the end of Section 1.

Conjecture 1 (Eigenvalue floor along GD). *Let f_1 , $\mathbf{W}_0$ and f_3 be random, with f_3 drawn from a sufficiently rich smooth family (e.g. a Gaussian process, or random cosine features). Let $M < D$, $MM' \geq C D$, and let the targets be realizable. Then with high probability $\lambda_{\min}(\mathbf{R}_t) \geq c \alpha_{\min}(t)$ for all t . Here $C > 1$ and $c > 0$ are constants that do not depend on D (Remark 7); the strongest form would allow every $C > 1$, with c depending on C .*

Statement	Status	§
A floor on $\lambda_{\min}(\mathbf{R}_t)$ gives geometric convergence	Proved	2
Wide case $M \geq D$: $\lambda_{\min}(\mathbf{R}_t) \geq \lambda_{\min}(\mathbf{X}^T \mathbf{X}) \alpha_{\min}(t)$	Proved	2
f_3 depending on k directions: $\text{rank}(\mathbf{R}) \leq kM$; quadratic f_3 : $\text{rank}(\mathbf{R}) \leq M(M+3)/2$	Proved	3.1
Independent Gaussian $\mathbf{g}_d$: $\lambda_{\min}(\mathbf{R}) \geq c \alpha_{\min}$ once $MM' \geq CD$	Sketch	3.2
Gaussian-process f_3 , well-separated $\mathbf{z}_d$: the $\mathbf{g}_d$ are nearly independent	Sketch	3.3
$\text{rank}(\mathbf{R}_t)$ is preserved along GD (prob. 1), but rank alone is not enough	Proved	4.1
Restricting what an adversary knows about f_3 does not help	Proved	4.2
Lipschitz bootstrap: the floor persists if $\ell_0^{3/2} > \frac{3}{2} L_R \sqrt{\mathcal{E}_0}$	Proved	4.3
... which closes only for $MM' \gtrsim D^2$, $M' \gtrsim D$ (lazy regime)	Heuristic	4.4
Piecewise-constant gradients: no union-bound proof outside the lazy regime	Heuristic	5
Floor along GD for cosine f_3 , $D = 51$, MM' from $1.1D$ to $51D$	Numerical	6
Conjecture 1 at $MM' = \Theta(D)$, for any f_3	Open	7

Labels. [PROVED]: complete proof here or in the main paper. [SKETCH]: an argument we believe can be completed but have not written out. [HEURISTIC]: an order-of-magnitude estimate. [NUMERICAL]: small simulations ($D = 51$), not evidence about large D .

2 Two Basic Facts

Lemma 2 ([PROVED]). *Along gradient flow $\dot{\mathbf{W}} = -\nabla_{\mathbf{W}} \mathcal{E}$, we have $\dot{\mathcal{E}} = -4 \mathbf{e}^T \mathbf{R} \mathbf{e}$. If $\lambda_{\min}(\mathbf{R}_t) \geq \mu$ on $[0, T]$, then $\mathcal{E}_T \leq \mathcal{E}_0 e^{-4\mu T}$.*

Proof. $\dot{\mathcal{E}} = \langle \nabla_{\mathbf{W}} \mathcal{E}, \dot{\mathbf{W}} \rangle = -\|\nabla_{\mathbf{W}} \mathcal{E}\|_F^2 = -4 \mathbf{e}^T \mathbf{R} \mathbf{e} \leq -4\mu \mathcal{E}$; apply Grönwall. $\square$

Lemma 3 ([PROVED]; wide case). *For positive semidefinite $\mathbf{A}, \mathbf{B}$, $\lambda_{\min}(\mathbf{A} \odot \mathbf{B}) \geq \lambda_{\min}(\mathbf{A}) \min_i B_{ii}$. Hence $\lambda_{\min}(\mathbf{R}) \geq \lambda_{\min}(\mathbf{X}^T \mathbf{X}) \alpha_{\min}$ for every $\mathbf{W}$.*

Proof. $\mathbf{A} - \lambda_{\min}(\mathbf{A}) \mathbf{I} \succeq 0$, so by the Schur product theorem $(\mathbf{A} - \lambda_{\min}(\mathbf{A}) \mathbf{I}) \odot \mathbf{B} \succeq 0$, i.e. $\mathbf{A} \odot \mathbf{B} \succeq \lambda_{\min}(\mathbf{A}) \text{diag}(B_{11}, \dots, B_{DD})$. Take $\mathbf{A} = \mathbf{X}^T \mathbf{X}$ and $\mathbf{B} = \mathbf{G}^T \mathbf{G}$, whose diagonal entries are the α_d . $\square$

For random unit vectors $\mathbf{x}_d$ and $M \geq (1 + \delta)D$, $\lambda_{\min}(\mathbf{X}^T \mathbf{X})$ is bounded below by a constant depending on δ (roughly $(1 - \sqrt{D/M})^2$). So in the wide case GD converges at rate $\gtrsim \alpha$ whenever $\alpha_{\min}(t) \geq \alpha$. When $M < D$, Lemma 3 is useless because $\lambda_{\min}(\mathbf{X}^T \mathbf{X}) = 0$.

3 At a Single Weight Matrix

3.1 When f_3 is too simple

Proposition 4 ([PROVED]). (a) If $f_3(\mathbf{z}) = h(\mathbf{U}\mathbf{z})$ for some $\mathbf{U} \in \mathbb{R}^{k \times M'}$ and differentiable h , then $\text{rank}(\mathbf{R}) \leq kM$ for every $\mathbf{W}$. (b) If f_3 is quadratic, then $\text{rank}(\mathbf{R}) \leq M(M+3)/2$ for every $\mathbf{W}$.

Proof. (a) $\mathbf{g}_d = \mathbf{U}^T \nabla h(\mathbf{U}\mathbf{z}_d)$ lies in the row space V of $\mathbf{U}$, so $\mathbf{y}_d \in \mathbb{R}^M \otimes V$, a space of dimension at most kM . (b) Write $\nabla f_3(\mathbf{z}) = \mathbf{A}\mathbf{z} + \mathbf{b}$. Then $\mathbf{y}_d = (\mathbf{I}_M \otimes \mathbf{A}\mathbf{W})(\mathbf{x}_d \otimes \mathbf{x}_d) + \mathbf{x}_d \otimes \mathbf{b}$. The vectors $\mathbf{x}_d \otimes \mathbf{x}_d$ lie in the symmetric subspace of $\mathbb{R}^M \otimes \mathbb{R}^M$ (dimension $M(M+1)/2$), and the vectors $\mathbf{x}_d \otimes \mathbf{b}$ lie in an M -dimensional space. $\square$

So a linear f_3 , or a single neuron $f_3(\mathbf{z}) = \sigma(\mathbf{a}^T \mathbf{z})$, has $k = 1$ and $\text{rank}(\mathbf{R}) \leq M < D$: then $\lambda_{\min}(\mathbf{R}) = 0$ although $\alpha_{\min} > 0$, and Conjecture 1 fails. A one-hidden-layer readout with F hidden units (including random cosine features with F frequencies) has $\text{rank}(\mathbf{R}) \leq \min(F, M')M$, so $FM \geq D$ is necessary. A quadratic f_3 fails once $M(M+3)/2 < D$.

3.2 Independent gradients

Claim 5 ([SKETCH]; probably known). Let $\mathbf{x}_1, \dots, \mathbf{x}_D$ be independent uniformly random unit vectors in $\mathbb{R}^M$, and let $\mathbf{g}_1, \dots, \mathbf{g}_D$ be independent $N(\mathbf{0}, \mathbf{I}_{M'})$ vectors, independent of the $\mathbf{x}_d$. Then with high probability

$$\lambda_{\min}(\mathbf{R}) \geq M' \left(1 - C\sqrt{D/(MM')} - C\sqrt{\log D/M'} \right).$$

In particular $\lambda_{\min}(\mathbf{R}) \geq c\alpha_{\min}$ once $MM' \geq C'D$ and $M' \geq C'\log D$.

Sketch. For a unit vector $\mathbf{v} \in \mathbb{R}^D$,

$$\left\| \sum_d v_d \mathbf{y}_d \right\|^2 = \sum_{m'=1}^{M'} \mathbf{h}_{m'}^T \mathbf{A}_{\mathbf{v}} \mathbf{h}_{m'},$$

where the vectors $\mathbf{h}_{m'} \stackrel{\text{def}}{=} (g_{1m'}, \dots, g_{Dm'})^T$ are i.i.d. $N(\mathbf{0}, \mathbf{I}_D)$ and $\mathbf{A}_{\mathbf{v}} \stackrel{\text{def}}{=} \text{diag}(\mathbf{v}) \mathbf{X}^T \mathbf{X} \text{diag}(\mathbf{v})$, with $\text{tr} \mathbf{A}_{\mathbf{v}} = 1$. By Hanson–Wright this concentrates around M' at scale $\sqrt{M'} \|\mathbf{A}_{\mathbf{v}}\|_F + \|\mathbf{A}_{\mathbf{v}}\|_{\text{op}}$; a union bound over a net of unit vectors $\mathbf{v}$, organized by how spread out $\mathbf{v}$ is, should give the claim. We have not written this out, and we suspect it is known.

Remark 6 (Near-orthogonality alone gives only D^2). If one only knows $|\mathbf{g}_d^T \mathbf{g}_{d'}| \leq \mu_G \sqrt{\alpha_d \alpha_{d'}}$ and $|\mathbf{x}_d^T \mathbf{x}_{d'}| \leq \mu_X$ for $d \neq d'$, Gershgorin gives $\lambda_{\min}(\mathbf{R}) \geq \alpha_{\min} - (D-1) \alpha_{\max} \mu_G \mu_X$. With $\mu_G \approx 1/\sqrt{M'}$ and $\mu_X \approx 1/\sqrt{M}$ this needs $MM' \gtrsim D^2$. The improvement to $MM' = \Theta(D)$ in Claim 5 comes from the random signs of the off-diagonal entries, which independence provides.

Remark 7 (The constants: why $C > 1$). [HEURISTIC; NUMERICAL.] $MM' \geq D$ already gives full rank, but not a floor. Write $\bar{\alpha} \stackrel{\text{def}}{=} \frac{1}{D} \sum_d \alpha_d$. For independent Gaussian $\mathbf{g}_d$, the Marchenko–Pastur law suggests $\lambda_{\min}(\mathbf{R}) \approx \bar{\alpha} (1 - \sqrt{D/MM'})^2$. With $MM' = CD$ this predicts $c \approx (1 - C^{-1/2})^2$: about 0.002, 0.09, 0.25, 0.44 for $C = 1.1, 2, 4, 9$. At $C = 1$ the floor is only of order $\bar{\alpha}/D^2$, so convergence would take on the order of D^2 iterations rather than logarithmically many. In a quick simulation ($D = 100, 400, 1600$; $\mathbf{x}_d$ uniform on the sphere; $\mathbf{g}_d \sim N(\mathbf{0}, \mathbf{I}_{M'})$; $M = 10$ or $M \approx M'$; a few trials each), $\lambda_{\min}(\mathbf{R})/\bar{\alpha}$ at $D = 1600$ was within 20% of these values for $C \geq 1.1$, and $\lambda_{\min}(\mathbf{R}) D^2/\bar{\alpha}$ stayed between 0.07 and 1 at $C = 1$. The sketch of Claim 5 gives only some large, unspecified C .

3.3 Gaussian-process f_3

[SKETCH.] Let f_3 be a centered Gaussian process with kernel $k(\mathbf{z}, \mathbf{z}') = e^{-\|\mathbf{z}-\mathbf{z}'\|^2/(2\sigma^2)}$, so σ is its *correlation length*. Then ∇f_3 is a Gaussian process with

$$\text{Cov}(\nabla f_3(\mathbf{z}), \nabla f_3(\mathbf{z}')) = \sigma^{-2} e^{-r^2/(2\sigma^2)} (\mathbf{I} - \delta\delta^T/\sigma^2), \quad \delta = \mathbf{z} - \mathbf{z}', \quad r = \|\delta\|.$$

Each $\mathbf{g}_d \sim N(\mathbf{0}, \sigma^{-2} \mathbf{I}_{M'})$, and if $\min_{d \neq d'} \|\mathbf{z}_d - \mathbf{z}_{d'}\| \geq c_0 \sigma \sqrt{\log D}$, every cross-covariance has operator norm at most $\sigma^{-2} c_0^2 \log D \cdot D^{-c_0^2/2}$. So the $\mathbf{g}_d$ are nearly independent, and Claim 5 should transfer by a perturbation argument (not written). At a random $\mathbf{W}_0$ with $N(0, 1/M)$ entries, $\|\mathbf{z}_d - \mathbf{z}_{d'}\| \approx \sqrt{2M'/M}$ (for $M, M' \gtrsim \log D$), so it suffices that $\sigma \lesssim \sqrt{M'/(M \log D)}$.

Caveat. This concerns a single $\mathbf{W}$ chosen independently of f_3 . Along GD, $\mathbf{W}_t$ depends on f_3 , so these distributional statements do not apply directly. That is the whole difficulty.

4 Along the Gradient Descent Path

4.1 Rank is preserved, but rank is not enough

[PROVED IN THE MAIN PAPER.] For analytic activations and random initialization, with probability one GD never enters the set where $\text{rank}(\mathbf{R}) < \min(MM', D)$. This gives no lower bound on $\lambda_{\min}(\mathbf{R}_t)$.

[PROVED.] *Example.* Let $D = 2$, $M = 1$ ($x_{\pm} = \pm 1$), $M' = 2$, $f_3(z_1, z_2) = z_1 + q(z_2)$ with $q(v) = (v^2 - 1)^2$, both targets equal to 2, and $\mathbf{W} = (u, v)^T$. Then $\alpha_d = 1 + q'(v)^2 \geq 1$, $\mathbf{R}$ has eigenvalues 2 and $2q'(v)^2$, and $\mathcal{E} = 2u^2 + 2(q(v) - 2)^2$. From any start with $0 < |v_0| < 1$, GD converges to the strict local minimum $(0, 0)$ with $\mathcal{E} = 2$, while $\mathbf{R}_t$ has full rank at every finite time and $\lambda_{\min}(\mathbf{R}_t) = 2q'(v_t)^2 \rightarrow 0$.

4.2 Limiting knowledge does not help

[PROVED; TRIVIAL.] Every $\mathbf{W}$ with $\|\mathbf{W} - \mathbf{W}_0\|_F \leq B$ can be reached in one step by an adversary that knows nothing about f_3 . So any bound that must hold against all adversaries with movement budget B must hold uniformly over that ball (the *ball version*). The alternative is to use the particular path that GD follows (the *trajectory version*, Conjecture 1).

4.3 A proved trajectory result: the Lipschitz bootstrap

Lemma 8 ([PROVED]). *Let $H \geq \|\nabla^2 f_3(\mathbf{z})\|_{\text{op}}$ for all $\mathbf{z}$ on the segments from $\mathbf{W}\mathbf{x}_d$ to $\mathbf{W}'\mathbf{x}_d$, and $\Gamma \geq \max(\|\mathbf{G}(\mathbf{W})\|_{\text{op}}, \|\mathbf{G}(\mathbf{W}')\|_{\text{op}})$. Then*

$$\|\mathbf{R}(\mathbf{W}) - \mathbf{R}(\mathbf{W}')\|_{\text{op}} \leq 2H\Gamma \|\mathbf{X}\|_{\text{op}} \|\mathbf{W} - \mathbf{W}'\|_F.$$

Proof. Since $\mathbf{X}^T \mathbf{X} \succeq 0$ has unit diagonal, $\|\mathbf{A} \odot \mathbf{X}^T \mathbf{X}\|_{\text{op}} \leq \|\mathbf{A}\|_{\text{op}}$ for every $\mathbf{A}$ (Horn and Johnson, *Topics in Matrix Analysis*, Thm. 5.5.18). Hence $\|\mathbf{R} - \mathbf{R}'\|_{\text{op}} \leq \|\mathbf{G}^T(\mathbf{G} - \mathbf{G}') + (\mathbf{G} - \mathbf{G}')^T \mathbf{G}'\|_{\text{op}} \leq 2\Gamma \|\mathbf{G} - \mathbf{G}'\|_F$, and $\|\mathbf{G} - \mathbf{G}'\|_F^2 = \sum_d \|\nabla f_3(\mathbf{W}\mathbf{x}_d) - \nabla f_3(\mathbf{W}'\mathbf{x}_d)\|^2 \leq H^2 \|(\mathbf{W} - \mathbf{W}')\mathbf{X}\|_F^2 \leq H^2 \|\mathbf{X}\|_{\text{op}}^2 \|\mathbf{W} - \mathbf{W}'\|_F^2$. $\square$

Theorem 9 ([PROVED]; Lipschitz bootstrap, gradient flow). *Let $\ell_0 \stackrel{\text{def}}{=} \lambda_{\min}(\mathbf{R}(\mathbf{W}_0)) > 0$. Suppose $\|\mathbf{R}(\mathbf{W}) - \mathbf{R}(\mathbf{W}')\|_{\text{op}} \leq L_R \|\mathbf{W} - \mathbf{W}'\|_F$ for all $\mathbf{W}, \mathbf{W}'$ in the ball $\mathfrak{B}$ of radius ρ around $\mathbf{W}_0$, and*

$$\ell_0^{3/2} > \frac{3}{2} L_R \sqrt{\mathcal{E}_0}, \quad \ell_{\infty} \stackrel{\text{def}}{=} (\ell_0^{3/2} - \frac{3}{2} L_R \sqrt{\mathcal{E}_0})^{2/3}, \quad \rho \geq \sqrt{\mathcal{E}_0 / \ell_{\infty}}.$$

Then along gradient flow, for all $t \geq 0$: (i) $\lambda_{\min}(\mathbf{R}_t) \geq \ell_\infty$; (ii) $\mathcal{E}_t \leq \mathcal{E}_0 e^{-4\ell_\infty t}$; (iii) the path length up to time t is less than $\sqrt{\mathcal{E}_0/\ell_\infty}$, so $\mathbf{W}_t \in \mathfrak{B}$.

Proof. Let $m(t)$ be the path length up to time t and $\lambda(t) \stackrel{\text{def}}{=} \ell_0 - L_R m(t)$. While $\mathbf{W}_t \in \mathfrak{B}$, Weyl's inequality gives $\lambda_{\min}(\mathbf{R}_t) \geq \lambda(t)$. Let T^* be the first time with $\lambda(t) \leq \ell_\infty/2$ or $m(t) \geq \rho$ ($T^* = \infty$ if none). We may assume $\mathcal{E}_t > 0$ (otherwise the flow has stopped at zero error). On $[0, T^*)$, $\|\dot{\mathbf{W}}\|_F^2 = -\dot{\mathcal{E}} = 4\mathbf{e}^T \mathbf{R} \mathbf{e} \geq 4\lambda\mathcal{E}$, so $\sqrt{\lambda} \leq \sqrt{-\dot{\mathcal{E}}}/(2\sqrt{\mathcal{E}})$ and

$$\frac{d}{dt} \lambda^{3/2} = -\frac{3}{2} L_R \sqrt{\lambda} \sqrt{-\dot{\mathcal{E}}} \geq -\frac{3}{2} L_R \frac{-\dot{\mathcal{E}}}{2\sqrt{\mathcal{E}}} = \frac{3}{2} L_R \frac{d}{dt} \sqrt{\mathcal{E}}.$$

Integrating, $\lambda(t)^{3/2} \geq \ell_0^{3/2} - \frac{3}{2} L_R (\sqrt{\mathcal{E}_0} - \sqrt{\mathcal{E}_t}) \geq \ell_\infty^{3/2}$. Then $\sqrt{-\dot{\mathcal{E}}} \geq 2\sqrt{\ell_\infty \mathcal{E}}$, so $\dot{m} = \sqrt{-\dot{\mathcal{E}}} \leq -\dot{\mathcal{E}}/(2\sqrt{\ell_\infty \mathcal{E}}) = -\ell_\infty^{-1/2} \frac{d}{dt} \sqrt{\mathcal{E}}$ and $m(t) \leq (\sqrt{\mathcal{E}_0} - \sqrt{\mathcal{E}_t})/\sqrt{\ell_\infty} < \rho$. By continuity neither stopping condition is ever met, so $T^* = \infty$, giving (i) and (iii); (ii) follows from Lemma 2. $\square$

A discrete-GD version (constant 3 in place of $\frac{3}{2}$, under a step-size condition) follows the same lines [SKETCH]. The argument is not circular: it needs only ℓ_0 , L_R and $\mathcal{E}_0$.

4.4 What width does the bootstrap give?

[HEURISTIC.] Think of f_3 as Gaussian-process-like with output scale 1 and correlation length σ . Typical sizes (estimates, not proved):

$$\begin{aligned} \alpha_d &\approx M'/\sigma^2, & H &\approx c_H \sqrt{M'}/\sigma^2, & \|\mathbf{G}\|_{\text{op}} &\approx (\sqrt{M'} + \sqrt{D})/\sigma, \\ \|\mathbf{X}\|_{\text{op}} &\approx 1 + \sqrt{D/M}, & \mathcal{E}_0 &\approx c_E D, & \ell_0 &\approx c_0 M'/\sigma^2. \end{aligned}$$

With $L_R = 2H\Gamma\|\mathbf{X}\|_{\text{op}}$ the factors of σ cancel, and the hypothesis of Theorem 9 becomes

$$M' \gtrsim \sqrt{D} (\sqrt{M'} + \sqrt{D}) (1 + \sqrt{D/M}).$$

Checking the cases $M \leq D$ and $M' \leq D$, this forces $M' \gtrsim D$ and $MM' \gtrsim D^2$. In that regime the path length $\sqrt{\mathcal{E}_0/\ell_\infty} \approx \sigma\sqrt{D/M'}$ is at most about one correlation length, so the gradient field barely changes along the path. This is the lazy regime, and we expect the resulting statement to be routine. At $MM' = \Theta(D)$ the bootstrap, with this L_R , gives nothing.

[NUMERICAL.] At $M = 8$, $M' = 80$, $D = 200$ with a cosine-feature f_3 , the bound $L_R = 2H\Gamma\|\mathbf{X}\|_{\text{op}}$ was about 360 times the largest local rate of change of $\lambda_{\min}(\mathbf{R})$ that we found by numerical search. Three loose steps in the proof of Lemma 8 ($\|\Delta\mathbf{G}\|_{\text{op}} \leq \|\Delta\mathbf{G}\|_F$; the Schur bound; ignoring the bottom eigenvector) account for a factor of about 40 at that size. We do not know how this slack scales with D , so we cannot say whether a sharper L_R reaches $MM' = \Theta(D)$.

5 Piecewise-Constant Gradients

[HEURISTIC; A NEGATIVE RESULT FOR ONE PROOF STRATEGY, DUE TO EDMONDS.] Suppose f_3 is linear on each cube of side r of a grid in $\mathbb{R}^{M'}$, with an independent random gradient on each cube. If r is large, two $\mathbf{z}_d$ share a cube already at initialization (birthday bound), so $\mathbf{g}_d = \mathbf{g}_{d'}$. If r is small, an adversary that can steer the $\mathbf{z}_d$ independently (possible when $M \geq D$) visits many cubes and finds pairs with nearly equal gradients, defeating any union bound. The two requirements are compatible only when the movement budget is far below the initial spacing of the $\mathbf{z}_d$ (the lazy regime). This rules out proving “independent gradients throughout a ball” for this class of f_3 . It is not a counterexample to Conjecture 1: when $M \geq D$ the floor holds anyway (Lemma 3). Smooth random f_3 avoid the small- r failure, because nearby points have strongly correlated gradients.

6 Simulations

[NUMERICAL.] Data: $\mathbf{x}_d$ independent uniformly random unit vectors in $\mathbb{R}^M$, $D = 51$. Readout: random cosine features $f_3(\mathbf{z}) = F^{-1/2} \sum_{j=1}^F \beta_j \cos(\boldsymbol{\omega}_j^T \mathbf{z} + \phi_j)$ with $F = 200$, $\boldsymbol{\omega}_j \sim N(\mathbf{0}, \mathbf{I}_{M'})$, $\phi_j \sim U[0, 2\pi]$, $\beta_j \sim N(0, 1)$, so the correlation length is about 1. $\mathbf{W}_0$ and a teacher $\mathbf{W}^*$ have i.i.d. $N(0, 1/M)$ entries; targets $y_d^* = f_3(\mathbf{W}^* \mathbf{x}_d)$ (realizable). Plain GD with $\eta = 0.25/\lambda_{\max}(\mathbf{R}_0)$, up to 20,000 iterations (stopping at $\mathcal{E} < 10^{-20}$); three seeds per width. Ranges are over the three seeds.

(M, M')	$\frac{MM'}{D}$	$\frac{\lambda_{\min}}{\alpha_{\min}}$ at $t=0$	min along GD	final $\mathcal{E}$	iterations	$\max_d \ \Delta \mathbf{z}_d\ $	$\ \mathbf{W}_T - \mathbf{W}_0\ _F$	$\ \mathbf{W}_0\ _F$
(7, 8)	1.1	0.0004–0.0012	0.0002–0.0006	$2 \cdot 10^{-5}$ – $9 \cdot 10^{-4}$	20,000 (cap)	1.6–1.8	2.9–3.4	2.5–2.6
(10, 10)	2.0	0.037–0.19	0.025–0.071	$\approx 10^{-20}$	6,431–9,638	0.9–1.2	1.9–2.9	2.8–3.5
(15, 15)	4.4	0.23–0.30	0.23–0.25	$\approx 10^{-20}$	954–1,278	0.75–1.1	1.8–2.3	3.9–4.0
(25, 25)	12	0.42–0.55	0.39–0.42	$\approx 10^{-20}$	633–886	0.53–0.85	1.5–1.8	5.0–5.3
(51, 51)	51	0.63–0.65	0.57–0.65	$\approx 10^{-20}$	204–237	0.42–0.50	1.1–1.3	7.0–7.3

Observations.

- $\lambda_{\min}(\mathbf{R}_t)/\alpha_{\min}(t)$ never fell below 0.36 times its initial value, and in 10 of the 15 runs it ended higher than it started.
- At $MM' \approx 1.1D$, $\mathbf{R}_0$ is already poorly conditioned (compare the Marchenko–Pastur edge $(1 - \sqrt{D/MM'})^2 \approx 0.002$), and the error (initially 14–50) was still 10^{-5} – 10^{-3} after 20,000 iterations.
- The typical (median) $\mathbf{z}_d$ moved between about 0.2 correlation lengths (at $MM' = 51D$) and 1.2 (at $MM' = 1.1D$); the farthest moved about 1.8. At small MM'/D the weights moved a distance comparable to $\|\mathbf{W}_0\|_F$, although $\|\mathbf{W}_t\|_F$ never fell below $0.93 \|\mathbf{W}_0\|_F$ in any run. So the runs at small MM'/D are not deep in the lazy regime.
- The smallest initial distance between two $\mathbf{z}_d$ was only 0.2–0.95 correlation lengths, so the $\mathbf{g}_d$ were not well separated in the sense of Section 3.3, even at $t = 0$.
- $D = 51$ and three seeds per width: these runs say nothing reliable about large D .

7 Open Problems

1. **The central problem.** Prove Conjecture 1 at $MM' = \Theta(D)$, $M < D$, for any concrete random f_3 (a Gaussian process, or random cosine features), beyond the lazy regime.
2. Is Claim 5 known? Extend it to cosine features with finitely many frequencies, where the $\mathbf{g}_d$ share the same random frequencies and so are not independent.
3. A sharper Lipschitz bound: control $\mathbf{v}^T(\Delta \mathbf{R})\mathbf{v}$ for the bottom eigenvector $\mathbf{v}$ of $\mathbf{R}$, rather than $\|\Delta \mathbf{R}\|_{\text{op}}$.
4. Explain the apparent self-correction in Section 6: why does GD so often raise $\lambda_{\min}(\mathbf{R}_t)/\alpha_{\min}(t)$?