WHEN DOES GRADIENT DESCENT GLOBALLY TRAIN A DEEP LINEAR RESNET?

Anonymous authors
Paper under double-blind review

ABSTRACT

When does gradient descent globally train a deep linear residual network from a nonzero initialization? We separate three aspects of factorization geometry: a lower radial coordinate r_- controlling gradient transmission, an upper radial coordinate r_+ controlling finite-step stiffness, and a centered logarithmic shape variable Θ_{sh} measuring multiplicative inter-layer mismatch. Inside an explicit spectral-balance basin, these coordinates yield a quantitative Polyak–Łojasiewicz inequality. We prove that a nested basin is forward invariant under depth-normalized gradient descent using relative-step smoothness, exact first-order cancellation of discrete imbalance drift, and summable radial motion; consequently, the training loss converges linearly to the global least-squares value. Exact scalar constructions show that the radial exponents are worst-case sharp and that a radially neutral depth-two inverse pair can have a connected one-step stability threshold of order $e^{-\Theta_{\mathrm{sh}}/2}$. We complement this worst-case obstruction with a perturbative result: for zero-target identity regression, an identity end-to-end predictor whose factors are each ε-close to identity has connected normalized threshold $2L + O_L(\varepsilon^2)$, where L is the depth. Thus, shape is needed for uniform worst-case control but is not, by itself, an instance-wise instability score. Matrix experiments reproduce this distinction: generic geometry orders stability, whereas matched same-predictor families can vary substantially in shape while retaining nearly unchanged thresholds.

1 INTRODUCTION

A deep linear residual network can realize the same end-to-end predictor through many different factorizations, yet gradient descent (GD) updates the individual factors rather than their product. For a depth-L model, write

$$P(F) = F_L \cdots F_1, \qquad F_\ell = I + W_\ell.$$

This representation raises a basic finite-step question: *which factorization geometries admit a global training guarantee, and which geometric effects obstruct guarantees that are uniform over such factorizations?* Prior analyses obtain convergence guarantees under several structured initialization or factorization regimes, including near-identity, random, orthogonal, and balanced settings (Hu et al., 2020; Zou et al., 2020; Min et al., 2023; Nguegnang et al., 2024; Chen et al., 2022). Our focus is an explicit factor-space region whose geometry is preserved by positive-step GD.

Three-coordinate geometry. Using the recursive polar gauge $F_\ell = C_\ell S_\ell C_{\ell-1}^\top$ and $Z_\ell = \log S_\ell$, define

$$r_- := \sum_\ell \log \sigma_{\min}(F_\ell), \qquad r_+ := \sum_\ell \log \|F_\ell\|_2, \qquad \Theta_{\mathrm{sh}} := L \sum_{\ell<L} \|Z_{\ell+1} - Z_\ell\|_2.$$

These coordinates separate effects that enter different parts of the finite-step analysis: lower singular-value transmission, upper-scale smoothness, and inter-layer mismatch. The lower radial coordinate controls gradient transmission, the upper radial coordinate controls finite-step stiffness, and Θ_{sh} controls the multiplicative inter-layer mismatch. Inside

$$\mathcal{B}(\rho_-, \rho_+, \theta) = \{r_- \geq -\rho_-,\, r_+ \leq \rho_+,\, \Theta_{\mathrm{sh}} \leq \theta\},$$

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

we prove

$$\sum_{\ell=1}^{L} \|\nabla_{F_\ell} \mathcal{L}\|_F^2 \geq 2L\alpha^2 \exp\left[-\frac{2(L-1)}{L}\rho_- - \frac{2(L-1)}{L^2}\theta\right](\mathcal{L} - \mathcal{L}^\star),$$

where $\alpha > 0$ is the data-dependent conditioning constant defined in Section 3. The main difficulty is not this static PL inequality, but showing that finite-step GD remains in a region where it continues to hold.

Forward invariance. For the depth-normalized step $\eta = \eta_0/L$, we prove that an inner spectral-balance basin remains inside a strictly larger one for all time. The proof combines relative-step smoothness with exact first-order cancellation in the evolution of adjacent-layer Gram imbalance. The remaining second-order drift is converted into logarithmic shape control through a noncommutative additive-to-log comparison, while radial motion is summable. A nested first-exit argument then closes all three coordinates simultaneously. Consequently,

$$\mathcal{E}(F^{(t)}) \leq (1 - \eta_0\mu_1)^t \mathcal{E}(F^{(0)}),$$

so the training loss converges linearly to the global least-squares value.

Worst-case obstructions versus instance-wise severity. Scalar witnesses isolate the three geometric effects. Balanced contraction attains the lower-radial gradient-transmission exponent; balanced expansion yields a stable normalized step of order $e^{-2(L-1)\rho_+/L}$; and a radially neutral depth-two inverse pair has

$$\eta_{0,\mathrm{small}} \sim 4e^{-\Theta_{\mathrm{sh}}/2}.$$

Thus, any uniform finite-step stability guarantee over this family must account for shape. This worst-case dependence, however, is not a monotone instance-wise law. For zero-target identity regression with $P(F) = I$ and $\max_\ell \|F_\ell - I\|_2 \leq \varepsilon$, we prove

$$\eta_{0,\mathrm{conn}}(F) = 2L + O_L(\varepsilon^2),$$

because the product constraint cancels the first-order perturbation of the balanced threshold. Matrix experiments reproduce this distinction: generic factorization geometry orders stability, whereas radially matched same-predictor families can vary substantially in shape while retaining nearly unchanged thresholds.

Contributions. We provide (i) a scale-separated spectral geometry and quantitative gradient transmission bound; (ii) a finite-step forward-invariance theorem for nonzero invertible initializations, yielding global-in-time linear convergence; and (iii) coordinate-isolating obstruction laws together with a complementary second-order robustness result near the balanced identity. Together, these results distinguish the geometry needed for uniform finite-step control from the geometry that predicts the severity of a particular realization. The resulting theory is sufficient rather than necessary: it gives a forward-invariant regime with explicit global guarantees and sharp coordinate-wise obstructions, without claiming a necessary-and-sufficient classification of trainable initializations.

2 Related Work

Deep-linear optimization. Deep linear models isolate nonconvex factorization geometry while retaining nontrivial optimization dynamics. Prior work studies their landscapes, dynamics, alignment, convergence, implicit bias, and low-rank behavior (Saxe et al., 2014; Kawaguchi, 2016; Laurent and von Brecht, 2018; Ji and Telgarsky, 2019; Shamir, 2019; Gunasekar et al., 2017; Arora et al., 2019b; Li et al., 2021; Bah et al., 2021), including the effects of initialization and factorization structure (Eftekhari, 2020; Tarmoun et al., 2021; Bordelon and Pehlevan, 2025; Dominé et al., 2025). Our emphasis is an explicit factor-space region that is preserved by finite-step GD and supports a uniform quantitative rate.

Balancedness and discrete dynamics. Balancedness and Gram-difference invariants are central to gradient-flow and discrete analyses of deep matrix factorizations (Du et al., 2018; Gidel et al., 2019; Arora et al., 2019a; Min et al., 2021; 2023; Nguegnang et al., 2024), with exact and relaxed balanced

classes further clarifying product-space dynamics (Dominé et al., 2025; Marcotte et al., 2026). We use exact first-order cancellation of discrete imbalance drift together with a noncommutative additive-to-log comparison and recursive-polar geometry to control radial and shape coordinates within a nested first-exit argument.

Residual parameterizations and finite-step stability. Initialization and residual/identity parameterizations can yield provable optimization guarantees for deep linear models (Hardt and Ma, 2017; Bartlett et al., 2018; Hu et al., 2020; Zou et al., 2020; Barboni et al., 2022); nonlocal convergence is also known under specific balance, rank, or architectural conditions (Chen et al., 2022; Min et al., 2023; Nguegnang et al., 2024). Our main result instead gives an explicit discrete-time invariant basin for invertible effective residual factors, separating lower radial scale, upper radial scale, and multiplicative shape. The obstruction and perturbative results further distinguish uniform worst-case finite-step control from instance-wise stability. Appendix L gives a theorem-level comparison with closely related results.

3 Setup and Spectral-Balance Geometry

Model and objective. Consider a depth-L linear residual network with effective factors $F_\ell = I + W_\ell$ and

$$P(F) = F_L \cdots F_1, \qquad \widehat{Y}(F) = BP(F)AX, \qquad \mathcal{L}(F) = \frac{1}{2}\|BP(F)AX - Y\|_F^2.$$

We measure optimization error relative to the unconstrained least-squares value

$$\mathcal{E}(F) = \mathcal{L}(F) - \mathcal{L}^\star, \qquad \mathcal{L}^\star = \min_\Theta \frac{1}{2}\|\Theta X - Y\|_F^2.$$

Assume that A has full column rank, B has full row rank, and $r = \operatorname{rank}(X) \geq 1$, and define

$$\alpha = \sigma_{\min}(B)\sigma_{\min}(A)\sigma_r(X), \quad \beta = \|B\|_2\|AX\|_2, \quad \beta_F = \|B\|_2\|AX\|_F.$$

For

$$Q_\ell = F_L \cdots F_{\ell+1}, \qquad R_\ell = F_{\ell-1} \cdots F_1,$$

with empty products equal to the identity, the layer gradient is

$$G_\ell = Q_\ell^\top B^\top (BP(F)AX - Y)X^\top A^\top R_\ell^\top.$$

Spectral-balance geometry. For invertible factors, use the recursive polar gauge

$$F_\ell = C_\ell S_\ell C_{\ell-1}^\top, \qquad C_\ell^\top C_\ell = I, \qquad S_\ell \succ 0, \qquad Z_\ell = \log S_\ell.$$

Define

$$r_- = \sum_{\ell=1}^{L} \log \sigma_{\min}(F_\ell), \qquad r_+ = \sum_{\ell=1}^{L} \log \|F_\ell\|_2,$$

and

$$\Theta_{\mathrm{sh}} = L \sum_{\ell=1}^{L-1} \|Z_{\ell+1} - Z_\ell\|_2.$$

Here, r_- and r_+ capture lower and upper radial scale, while Θ_{sh} measures multiplicative interlayer mismatch. Define $\mathcal{B}(\rho_-, \rho_+, \theta) = \{F : r_- \geq -\rho_-, \quad r_+ \leq \rho_+, \quad \Theta_{\mathrm{sh}} \leq \theta\}$. For every $F \in \mathcal{B}(\rho_-, \rho_+, \theta)$ and every layer,

$$\sigma_{\min}(Q_\ell)\sigma_{\min}(R_\ell) \geq e^{-A_-}, \qquad \|Q_\ell\|_2\|R_\ell\|_2 \leq e^{A_+}, \tag{1}$$

where

$$A_\pm = \frac{L-1}{L}\rho_\pm + \frac{L-1}{L^2}\theta.$$

These gauge-independent bounds follow from centering the logarithmic polar path; see Appendix A.

Additive versus logarithmic balance. The discrete dynamics naturally control the adjacent Gram imbalance

$$D_\ell = F_{\ell+1}^\top F_{\ell+1} - F_\ell F_\ell^\top .$$

The following noncommutative comparison connects this additive defect to the logarithmic shape coordinate.

Proposition 3.1 (Additive and logarithmic balance)**.** *Let $S, T \succ 0$. If $S, T \succeq mI$, then*

$$\|\log S - \log T\|_2 \le \frac{1}{2m^2}\|S^2 - T^2\|_2. \tag{2}$$

If $S, T \preceq MI$, then

$$\|S^2 - T^2\|_2 \le 2M^2\|\log S - \log T\|_2. \tag{3}$$

Neither inequality requires S and T to commute.

Under simultaneous GD, the first-order drift of D_ℓ cancels exactly, so its accumulated change is second order in the stepsize. Together with a spectral floor, Proposition 3.1 converts this control into stability of Θ_{sh}. Proofs and the endpoint-safe form used in the first-exit argument appear in Appendix B.

4 Gradient Transmission

The factor gradients transmit the prediction residual through the excluded products Q_ℓ and R_ℓ, so controlling the end-to-end map alone does not control layerwise optimization geometry. The spectral-balance region instead controls these products uniformly, yielding the following bounds.

Theorem 4.1 (Geometry-sensitive gradient transmission)**.** *Let $F \in \mathcal{B}(\rho_-, \rho_+, \theta)$. Then, for every $\ell = 1, \dots, L$,*

$$\sqrt{2}\,\alpha e^{-A_-}\sqrt{\mathcal{E}(F)} \le \|G_\ell\|_F \le \sqrt{2}\,\beta e^{A_+}\sqrt{\mathcal{E}(F)}. \tag{4}$$

Consequently,

$$\|\nabla_F \mathcal{L}(F)\|_F^2 := \sum_{\ell=1}^{L}\|G_\ell\|_F^2 \ge 2L\mu(\rho_-, \theta)\mathcal{E}(F), \tag{5}$$

where the depth-normalized PL coefficient is

$$\mu(\rho_-, \theta) = \alpha^2 e^{-2A_-} = \alpha^2 \exp\left(-\frac{2(L-1)}{L}\rho_- - \frac{2(L-1)}{L^2}\theta\right). \tag{6}$$

Thus, lower radial scale ρ_- and shape mismatch θ govern the gradient-transmission lower bound, whereas ρ_+ enters the upper bound through A_+ and will control finite-step stiffness. In particular, the three geometric coordinates play distinct roles rather than entering through a single condition number. The proof combines least-squares residual geometry with the excluded-product bounds of Section 3; see Appendix C.

5 Finite-Step Basin Invariance

The PL inequality of Section 4 is useful globally only if the GD trajectory remains in a region where the spectral bounds underlying it continue to hold. We therefore work with two nested spectral-balance regions $\mathcal{B}_0 = \mathcal{B}(\rho_-^{(0)}, \rho_+^{(0)}, \theta_0) \Subset \mathcal{B}_1 = \mathcal{B}(\rho_-^{(1)}, \rho_+^{(1)}, \theta_1)$, with strict margins

$$\rho_-^{(1)} > \rho_-^{(0)}, \qquad \rho_+^{(1)} > \rho_+^{(0)}, \qquad \theta_1 > \theta_0. \tag{7}$$

The inner basin specifies the admissible initialization, while the outer basin provides room for finite-step motion. The goal is to show that, for a sufficiently small depth-normalized step size, the trajectory never exhausts these margins.

Relative-step smoothness and descent. A finite GD update must be controlled not only at its endpoints but along the segment joining them. On $\mathcal{B}_1$, consider a perturbation $\Delta F = (\Delta F_1, \dots, \Delta F_L)$ satisfying

$$\sum_{\ell=1}^{L} \frac{\|\Delta F_\ell\|_2}{\sigma_{\min}(F_\ell)} \le \xi < 1.$$

Since each layerwise relative perturbation is at most $\xi < 1$, every factor remains invertible along the segment, with uniform spectral control. For later use, define

$$A_{\mathrm{rel}}^{(1)} := \frac{L-1}{L}\rho_+^{(1)} + \frac{\rho_-^{(1)}}{L} + \frac{2(L-1)}{L^2}\theta_1.$$

Lemma 5.1 (Relative-step basin smoothness)**.** *There is an explicit constant $C_{\mathrm{sm}}^{(1)}(\xi)$, given in Appendix D, such that the loss is $LC_{\mathrm{sm}}^{(1)}(\xi)$-smooth along every admissible segment in the product factor norm.*

For GD, $F_\ell^{(t+1)} = F_\ell^{(t)} - \eta G_\ell^{(t)}$ and $\eta = \frac{\eta_0}{L}$. The layerwise gradient bound from Theorem 4.1 gives

$$\sum_{\ell=1}^{L} \frac{\eta\|G_\ell^{(t)}\|_2}{\sigma_{\min}(F_\ell^{(t)})} \le \sqrt{2}\,\eta_0\beta e^{A_{\mathrm{rel}}^{(1)}}\sqrt{\mathcal{E}_t}. \tag{8}$$

Therefore, a suitable bound on η_0 makes every update segment admissible. If, in addition, $\eta_0 C_{\mathrm{sm}}^{(1)}(\xi) \le 1$, the descent estimate gives

$$\mathcal{L}_{t+1} \le \mathcal{L}_t - \frac{\eta}{2}\sum_{\ell=1}^{L}\|G_\ell^{(t)}\|_F^2. \tag{9}$$

Second-order imbalance drift. Define the adjacent-layer Gram imbalance

$$D_\ell^{(t)} = (F_{\ell+1}^{(t)})^\top F_{\ell+1}^{(t)} - F_\ell^{(t)}(F_\ell^{(t)})^\top.$$

The factor gradients satisfy

$$G_\ell^{(t)}(F_\ell^{(t)})^\top = (F_{\ell+1}^{(t)})^\top G_{\ell+1}^{(t)}.$$

Under simultaneous updates, this identity cancels the first-order variation of $D_\ell^{(t)}$ exactly, leaving

$$D_\ell^{(t+1)} - D_\ell^{(t)} = \eta^2\left[(G_{\ell+1}^{(t)})^\top G_{\ell+1}^{(t)} - G_\ell^{(t)}(G_\ell^{(t)})^\top\right]. \tag{10}$$

Thus, additive imbalance accumulates at second order in the step size rather than first order. Descent controls the resulting total squared-gradient motion, while Proposition 3.1 converts the accumulated imbalance into logarithmic shape control.

Controlling shape and radial motion. Define

$$M_0 = \exp\left(\frac{\rho_+^{(0)}}{L} + \frac{L-1}{L^2}\theta_0\right),$$

and

$$m_1 = \exp\left(-\frac{\rho_-^{(1)}}{L} - \frac{L-1}{L^2}\theta_1\right), \qquad m_{1,\xi} = (1-\xi)m_1.$$

Combining the exact drift identity with descent and the additive-to-logarithmic comparison gives

$$\Theta_{\mathrm{sh}}(F^{(t)}) \le \frac{M_0^2}{m_{1,\xi}^2}\theta_0 + \frac{2\eta_0}{m_{1,\xi}^2}\mathcal{E}_0. \tag{11}$$

In parallel, summability of the relative factor motion yields

$$\left|r_\pm(F^{(t)}) - r_\pm(F^{(0)})\right| \le C_{\mathrm{rad}}^{(1)}\sqrt{\mathcal{E}_0}. \tag{12}$$

The explicit expression for $C_{\mathrm{rad}}^{(1)}$ and the endpoint-safe derivations are given in Appendices E–F.

The two estimates have different origins. Shape stability exploits the second-order cancellation in (10), whereas radial stability follows from summable first-order relative motion. The nested basin provides sufficient margin for both effects simultaneously.

Forward invariance. We can now state the main finite-step result.

Theorem 5.2 (Forward invariance under depth-normalized GD)**.** *Let $F^{(0)} \in \mathcal{B}_0$ have invertible factors and let $\mathcal{E}_0 > 0$. Fix $0 < \xi < 1$, choose $\eta_0 > 0$, set $\eta = \eta_0/L$, and define*

$$\mu_1 := \alpha^2 e^{-2A_-^{(1)}}.$$

Suppose

$$0 < \eta_0 \le \min\left\{ [C_{\mathrm{sm}}^{(1)}(\xi)]^{-1}, \frac{\xi e^{-A_{\mathrm{rel}}^{(1)}}}{\sqrt{2}\,\beta\sqrt{\mathcal{E}_0}}, \frac{e^{2A_-^{(1)}}}{\alpha^2} \right\}, \tag{13}$$

and the basin margins satisfy

$$\rho_-^{(1)} - \rho_-^{(0)} > C_{\mathrm{rad}}^{(1)}\sqrt{\mathcal{E}_0}, \qquad \rho_+^{(1)} - \rho_+^{(0)} > C_{\mathrm{rad}}^{(1)}\sqrt{\mathcal{E}_0},$$

and

$$\theta_1 > \frac{M_0^2}{m_{1,\xi}^2}\theta_0 + \frac{2\eta_0}{m_{1,\xi}^2}\mathcal{E}_0.$$

Then

$$F^{(t)} \in \mathcal{B}_1 \qquad \text{for every } t \ge 0,$$

and (9) *holds at every iteration. In particular, every factor remains invertible for all $t \ge 0$. Moreover, the third stepsize condition gives*

$$0 < \eta_0\mu_1 \le 1.$$

The proof is by first exit. Before a hypothetical exit time $t_\star$, the relative-step, smoothness, and gradient-transmission bounds remain valid, so

$$\mathcal{E}_t \le (1 - \eta_0\mu_1)^t \mathcal{E}_0, \qquad 0 \le t \le t_\star. \tag{14}$$

This decay controls the accumulated second-order imbalance drift and radial motion, keeping all three spectral coordinates strictly within their outer margins at $t_\star$, a contradiction. Hence $\mathcal{B}_1$ is forward invariant and the estimates hold for all $t \ge 0$; see Appendix G for the endpoint-safe bootstrap.

6 Global Linear Convergence

Forward invariance turns the local geometric estimate of Section 4 into a global-in-time convergence guarantee. Indeed, Theorem 5.2 ensures that the trajectory remains in $\mathcal{B}_1$ at every iteration, so the same geometry-sensitive gradient-transmission bound applies uniformly along the entire trajectory.

Theorem 6.1 (Global linear convergence in the spectral-balance basin)**.** *Under the hypotheses of Theorem 5.2, gradient descent with $\eta = \eta_0/L$ satisfies, for every $t \ge 0$,*

$$\mathcal{E}(F^{(t)}) \le (1 - \eta_0\mu_1)^t \mathcal{E}(F^{(0)}) = \left(1 - \eta_0\alpha^2 e^{-2A_-^{(1)}}\right)^t \mathcal{E}(F^{(0)}), \tag{15}$$

where $\mu_1 = \alpha^2 e^{-2A_-^{(1)}}$. In particular, $\mathcal{L}(F^{(t)}) \to \mathcal{L}^\star$. Moreover, for any $0 < \varepsilon < \mathcal{E}(F^{(0)})$, it is sufficient to take

$$t \ge \frac{1}{\eta_0\mu_1} \log \frac{\mathcal{E}(F^{(0)})}{\varepsilon} = \frac{e^{2A_-^{(1)}}}{\eta_0\alpha^2} \log \frac{\mathcal{E}(F^{(0)})}{\varepsilon} \tag{16}$$

to ensure $\mathcal{E}(F^{(t)}) \le \varepsilon$.

The theorem gives global-in-time convergence *from* the spectral-balance basin, rather than convergence from arbitrary factorizations. Its role is therefore distinct from that of the forward-invariance theorem: the latter establishes that the geometry needed for gradient transmission persists under finite-step GD, whereas the present result converts that persistence into a uniform optimization rate along the entire trajectory.

The depth-normalized step makes one aspect of the depth dependence transparent. The gradient-transmission inequality contributes a factor L through the sum of the L layer gradients, while

$\eta = \eta_0/L$ cancels this explicit factor in the one-step contraction. In the contraction coefficient itself, depth remains through

$$A_-^{(1)} = \frac{L-1}{L}\rho_-^{(1)} + \frac{L-1}{L^2}\theta_1.$$

Additional depth dependence can enter through the admissible stepsize and the basin conditions of Theorem 5.2. Thus the theorem does not claim a depth-independent training complexity; rather, depth is separated from the explicit L-fold gradient aggregation and enters through the geometry and finite-step stability requirements.

The proof follows by combining the descent inequality with the uniform gradient-transmission bound maintained by forward invariance and iterating the resulting contraction. The complexity bound uses $1 - x \le e^{-x}$ for $x = \eta_0\mu_1$. Details are given in Appendix H.

7 Quantitative Obstructions

The preceding results give sufficient geometric conditions for global finite-step training. We now ask to what extent their dependence on the three spectral coordinates can be improved uniformly. To separate radial contraction, radial expansion, and inter-layer shape, we use scalar factorizations in which one adverse geometric direction can be varied while the others remain controlled. These examples are deliberately minimal: they test the scaling laws in the sufficient theory without introducing additional matrix-algebra effects.

For a fixed initialization F, call a normalized step η_0 one-step stable if $\mathcal{L}(F_{\eta_0}^+) \le \mathcal{L}(F)$, and let $\eta_{0,\mathrm{conn}}(F)$ denote the endpoint of the connected one-step-stable interval containing $\eta_0 = 0$.

Radial contraction. Consider the exactly balanced scalar factorization

$$f_\ell = e^{-\rho/L}, \qquad \ell = 1, \dots, L.$$

For this family,

$$\frac{\sum_{\ell=1}^{L} g_\ell^2}{2L\mathcal{E}} = e^{-2(L-1)\rho/L}. \tag{17}$$

Thus the lower-radial dependence in the gradient-transmission theorem is attained exactly. Uniformly over this class, the factor $e^{-2(L-1)\rho/L}$ cannot be replaced by one with a strictly better exponential dependence on ρ.

Radial expansion. The opposite radial direction affects finite-step stability rather than gradient transmission. For the balanced scalar factorization $f_\ell = e^{\rho/L}$ with $Y = 0$, the connected one-step stability threshold is

$$\eta_{0,\mathrm{conn}} = 2Le^{-2(L-1)\rho/L}. \tag{18}$$

Thus radial expansion produces an exponentially shrinking stable step-size scale. This reproduces the upper-radial exponential scale that enters the finite-step bounds.

Pure shape imbalance. Radial scale alone does not account for all finite-step obstructions. At depth two, consider the inverse pair

$$(f_1, f_2) = (a, a^{-1}), \qquad a \ge 1.$$

Its end-to-end product is fixed, and its radial coordinates are neutral: $f_2 f_1 = 1$ and $r_- = r_+ = 0$. Nevertheless, $\Theta_{\mathrm{sh}} = 4\log a$. Writing $C_\Theta = \cosh(\Theta_{\mathrm{sh}}/2)$, for $C_\Theta \ge \sqrt{2}$ the endpoint of the connected one-step-stable component containing zero is

$$\eta_{0,\mathrm{conn}} = 2C_\Theta - 2\sqrt{C_\Theta^2 - 2}. \tag{19}$$

Consequently,

$$\eta_{0,\mathrm{conn}} \sim 4e^{-\Theta_{\mathrm{sh}}/2} \qquad \text{as } \Theta_{\mathrm{sh}} \to \infty. \tag{20}$$

Thus, finite-step stability can deteriorate exponentially with inter-layer shape even when the represented predictor and both radial coordinates remain fixed.

These constructions give the following worst-case limitations.

Theorem 7.1 (Coordinate-wise obstructions to uniform finite-step guarantees)**.** *Over the corresponding geometric classes:*

1. *the lower-radial factor $e^{-2(L-1)\rho_-/L}$ in the gradient-transmission scaling is attained by exactly balanced scalar contractions;*
2. *the connected one-step stable normalized step-size scale can decay as $e^{-2(L-1)\rho_+/L}$ under exactly balanced scalar expansions; and*
3. *even with $r_- = r_+ = 0$ and a fixed end-to-end product, the connected one-step stable normalized step-size scale can decay as $e^{-\Theta_{\mathrm{sh}}/2}$ at depth two.*

Consequently, lower radial scale, upper radial scale, and inter-layer shape represent distinct worst-case limitations on uniform finite-step guarantees.

The obstruction theorem is existential, not a converse to Theorem 5.2: it identifies distinct worst-case scalings but does not make the spectral-balance basin necessary or equate one-step instability with failure of eventual convergence. Exact derivations and quantifiers appear in Appendix I.

A complementary near-identity regime. In particular, the pure-shape obstruction does not make Θ_{sh} an instance-wise instability score. The following result shows that, near the balanced identity, the same-predictor stability threshold is instead second-order robust.

Proposition 7.2 (Second-order robustness of the balanced-identity threshold)**.** *Fix the depth L. Consider the zero-target identity problem*

$$\mathcal{L}(F) = \frac{1}{2}\|F_L \cdots F_1\|_F^2$$

and suppose

$$F_L \cdots F_1 = I, \qquad F_\ell = I + E_\ell, \qquad \max_{\ell \in [L]} \|E_\ell\|_2 \le \varepsilon.$$

For GD with step size $\eta = \eta_0/L$, let $\eta_{0,\mathrm{conn}}(F)$ be the endpoint of the connected interval containing zero on which one GD step does not increase the loss. There exist $\varepsilon_L > 0$ and $C_L < \infty$, independent of the matrix dimension, such that whenever $\varepsilon \le \varepsilon_L$,

$$|\eta_{0,\mathrm{conn}}(F) - 2L| \le C_L \varepsilon^2.$$

Hence, for fixed depth,

$$\eta_{0,\mathrm{conn}}(F) = 2L + O_L(\varepsilon^2).$$

The mechanism is a first-order cancellation imposed by the exact product constraint. In this zero-target identity problem, $P(F) = I$ gives $G_\ell = F_\ell^{-\top}$, while $\prod_\ell (I + E_\ell) = I$ implies

$$\sum_\ell E_\ell = O_L(\varepsilon^2).$$

Consequently, writing $h = \eta_0/L$, the finite-step product near the balanced threshold obeys

$$P(F^+(h)) = (1-h)^L I + O_L(\varepsilon^2),$$

so the balanced-identity boundary $h = 2$ moves only at second order. A three-region argument closes the connected-component statement; the complete proof is in Appendix J.

8 Experiments

The experiments test the distinction between *uniform worst-case geometric control* and *instance-wise severity*. Appendix K contains the frozen protocols, numerical definitions, and full tables.

Exact laws. Scalar unit tests recover the contraction gradient-transmission scale $e^{-2(L-1)\rho/L}$ with maximum relative error 4.53×10^{-15}, the expansion threshold $2Le^{-2(L-1)\rho/L}$ with error 6.00×10^{-15}, and the inverse-pair threshold with error 1.78×10^{-12}. The latter has fitted asymptotic log-slope -0.500274 with $R^2 = 0.99999985$, matching the predicted $e^{-\Theta_{\mathrm{sh}}/2}$ scaling. The corresponding figure is in Appendix K.

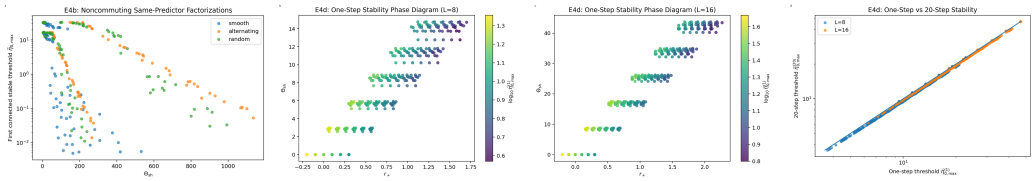

Figure 1: **Matrix-valued finite-step stability.** Same-predictor factorizations can have widely different connected stable steps (left); the predeclared theory-motivated score has substantial rank association with stability in a generic ensemble (middle); independently computed one-step and 20-step thresholds have Spearman correlation 0.99924 (right).

Matrix geometry. We numerically estimate the endpoint of the connected one-step-stable component, denoted $\widehat{\eta}^{(1)}_{0,\mathrm{conn}}$. Across 288 noncommuting telescoping factorizations satisfying $\|P(F) - I\|_F \le 1.29 \times 10^{-13}$, the threshold ranges from 4.86×10^{-3} to 32.79, showing that the represented predictor alone does not determine finite-step stability.

On a predeclared ensemble of 720 generic initializations spanning multiple dimensions, depths, and factorization geometries (Appendix K), the theory-motivated score

$$\Gamma(F) = \frac{2(L-1)}{L} \max\{r_+(F), 0\} + \frac{2(L-1)}{L^2} \Theta_{\mathrm{sh}}(F)$$

has Spearman correlation -0.765 at $L = 8$ and -0.742 at $L = 16$ with $\log \widehat{\eta}^{(1)}_{0,\mathrm{conn}}$. These are observational ordering statistics, not independent-coordinate effects or evidence that Γ is an optimal predictor: r_+ alone has marginal correlations -0.868 and -0.809 on the same ensemble.

Matched negative control and perturbative law. A prespecified same-predictor intervention accepts $732/840$ configurations (87.1%, below the frozen 90% operational gate) while matching the product and radial coordinates to numerical precision. Among 111 matched families, the median within-family threshold ratio is only 1.007 despite the prespecified variation in shape; across accepted configurations, the median

$$\widehat{\eta}^{(1)}_{0,\mathrm{conn}}/(2L) = 1.001.$$

We report the calibration shortfall without retuning the gate. The intervention therefore does not support a universal monotone shape–stability law.

Finally, an independent perturbative experiment directly checks the scaling predicted by Proposition 7.2. Across 576 configurations with $d \in \{4, 8\}$ and $L \in \{2, 4, 8, 16\}$, all product-constraint and connected-threshold checks pass. Fitting

$$\left| \widehat{\eta}^{(1)}_{0,\mathrm{conn}} - 2L \right| \propto \varepsilon^p$$

gives median $p = 1.94$ over the full frozen grid and $p = 1.98$ when restricted to the five smallest-ε levels per seed. This corroborates the predicted second-order perturbative law while allowing the expected higher-order deviations at finite ε.

9 Limitations and Conclusion

We establish a sufficient geometric regime for global finite-step training of deep linear residual networks. The analysis assumes invertible effective factors and worst-case constants, and the $2L$ perturbative result is fixed-depth rather than depth-uniform. Lower radial scale controls gradient transmission, upper radial scale controls finite-step stiffness, and logarithmic shape controls inter-layer mismatch; forward invariance converts this geometry into global-in-time linear convergence. Exact obstructions identify distinct worst-case limitations, while the near-identity result and matched experiments show that these scalings need not determine instance-wise stability. Thus, neither the end-to-end predictor nor aggregate shape alone determines finite-step stability: the realized factorization matters. Extending this framework to rank-deficient factors, nonlinear residual architectures, and instance-adaptive geometric guarantees remains open.

Reproducibility Statement

Anonymous reproduction code is available at https://anonymous.4open.science/r/spectral-balance-gd-ECC3/. The repository provides one-command reproduction scripts, pinned dependencies, frozen reference outputs, raw experiment tables, and automated verification for the reported experiments.

AI Use Statement

Generative AI tools were used as research-assistance tools for proof checking, code implementation and refactoring, grammar checking, and manuscript editing. All claims, mathematical proofs, citations, code, and numerical results were reviewed and verified by the authors, who take responsibility for the final submission and all AI-assisted artifacts.

References

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2019a. URL https://openreview.net/forum?id=SkMQg3C5K7.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32, pages 7411–7422, 2019b. URL https://papers.neurips.cc/paper/8960-implicit-regularization-in-deep-matrix-factorization.

Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 2021. doi: 10.1093/imaiai/iaaa039. URL https://doi.org/10.1093/imaiai/iaaa039.

Raphael Barboni, Gabriel Peyre, and Francois-Xavier Vialard. On global convergence of resnets: From finite to infinite width using linear parameterization. In *Advances in Neural Information Processing Systems*, volume 35, pages 16385–16397, 2022. doi: 10.52202/068431-1192. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/67ea314d1df751bbf99ab664ae3049a5-Abstract.html.

Peter L. Bartlett, David P. Helmbold, and Philip M. Long. Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 521–530. PMLR, 2018. URL https://proceedings.mlr.press/v80/bartlett18a.html.

Blake Bordelon and Cengiz Pehlevan. Deep linear network training dynamics from random initialization: Data, width, depth, and hyperparameter transfer. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 4968–4997. PMLR, 2025. URL https://proceedings.mlr.press/v267/bordelon25a.html.

Hengchao Chen, Xin Chen, Mohamad Elmasri, and Qiang Sun. Gradient descent in matrix factorization: Understanding large initialization. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, volume 244 of *Proceedings of Machine Learning Research*, pages 619–647. PMLR, 2024. URL https://proceedings.mlr.press/v244/chen24a.html.

Kun Chen, Dachao Lin, and Zhihua Zhang. On non-local convergence analysis of deep linear networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 3417–3443. PMLR, 2022. URL https://proceedings.mlr.press/v162/chen22p.html.

Kun Chen, Dachao Lin, and Zhihua Zhang. A non-local convergence analysis of gradient flow for deep linear networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

doi: 10.1109/TPAMI.2026.3704447. URL https://doi.org/10.1109/TPAMI.2026.3704447. Early access.

Clémentine C. J. Dominé, Nicolas Anguita, Alexandra M. Proca, Lukas Braun, Daniel Kunin, Pedro A. M. Mediano, and Andrew M. Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fde1a69a5b6e554b2f1f727197d2651d-Abstract-Conference.html.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, volume 31, pages 382–393, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/fe131d7f5a6b38b23cc967316c13dae2-Abstract.html.

Armin Eftekhari. Training linear neural networks: Non-local convergence and complexity results. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2836–2847. PMLR, 2020. URL https://proceedings.mlr.press/v119/eftekhari20a.html.

Avrajit Ghosh, Soo Min Kwon, Rongrong Wang, Saiprasad Ravishankar, and Qing Qu. Learning dynamics of deep matrix factorization beyond the edge of stability. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2d43f7a61b57f83619f82c971e4bddc0-Abstract-Conference.html.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f39ae9ff3a81f499230c4126e01f421b-Abstract.html.

Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, volume 30, pages 6151–6159, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/58191d2a914c6dae66371c9dcdc91b41-Abstract.html.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1611.04231.

Wei Hu, Lechao Xiao, and Jeffrey Pennington. Provable benefit of orthogonal initialization in optimizing deep linear networks. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2001.05992.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJflg30qKX.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, volume 29, pages 586–594, 2016. URL https://papers.nips.cc/paper/2016/hash/f2fc990265c712c49d51a18a32b39f0c-Abstract.html.

Thomas Laurent and James von Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2902–2907. PMLR, 2018. URL https://proceedings.mlr.press/v80/laurent18a.html.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=AHOs7Sm5H7R.

Sibylle Marcotte, Gabriel Peyré, and Rémi Gribonval. Intrinsic training dynamics of deep neural networks. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/184b2454410a88c280d29aa95fbac6e2-Abstract-Conference.html.

Hancheng Min, Salma Tarmoun, Rene Vidal, and Enrique Mallada. On the explicit role of initialization on the convergence and implicit bias of overparametrized linear networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7760–7768. PMLR, 2021. URL https://proceedings.mlr.press/v139/min21c.html.

Hancheng Min, Rene Vidal, and Enrique Mallada. On the convergence of gradient flow on multi-layer linear models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24850–24887. PMLR, 2023. URL https://proceedings.mlr.press/v202/min23d.html.

Gabin Maxime Nguegnang, Holger Rauhut, and Ulrich Terstiege. Convergence of gradient descent for learning linear neural networks. *Advances in Continuous and Discrete Models*, 2024(23), 2024. doi: 10.1186/s13662-023-03797-x. URL https://doi.org/10.1186/s13662-023-03797-x.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6120.

Ohad Shamir. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2691–2713. PMLR, 2019. URL https://proceedings.mlr.press/v99/shamir19a.html.

Salma Tarmoun, Guilherme Franca, Benjamin D. Haeffele, and Rene Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10153–10161. PMLR, 2021. URL https://proceedings.mlr.press/v139/tarmoun21a.html.

Difan Zou, Philip M. Long, and Quanquan Gu. On the global convergence of training deep linear resnets. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SEDpetA3xwZ.

A Polar Gauge and Spectral-Balance Geometry

This appendix develops the geometric facts used in Section 3. Throughout, the effective factors $F_1, \dots, F_L \in \mathbb{R}^{m\times m}$ are assumed invertible.

A.1 Recursive polar gauge

The representation

$$F_\ell = C_\ell S_\ell C_{\ell-1}^\top$$

underlying the spectral-balance geometry of Section 3 is a shared recursive polar gauge rather than a collection of unrelated polar decompositions. We first show that, for each fixed initial orthogonal frame, this gauge is well defined and unique.

Lemma A.1 (Recursive polar gauge)**.** *Fix any $C_0 \in O(m)$. For every invertible factorization $(F_1, \dots, F_L)$, there exist uniquely determined matrices*

$$C_\ell \in O(m), \qquad S_\ell \succ 0, \qquad \ell \in [L],$$

such that

$$F_\ell = C_\ell S_\ell C_{\ell-1}^\top. \tag{21}$$

They are obtained recursively from the right polar decomposition

$$F_\ell C_{\ell-1} = C_\ell S_\ell. \tag{22}$$

Equivalently,

$$S_\ell = \left(C_{\ell-1}^\top F_\ell^\top F_\ell C_{\ell-1}\right)^{1/2}, \qquad C_\ell = F_\ell C_{\ell-1} S_\ell^{-1}. \tag{23}$$

Proof. Assume that $C_{\ell-1} \in O(m)$ has already been constructed. Since F_ℓ is invertible, so is $F_\ell C_{\ell-1}$. Its right polar decomposition is unique and has the form

$$F_\ell C_{\ell-1} = C_\ell S_\ell,$$

where C_ℓ is orthogonal and $S_\ell \succ 0$. The positive factor is

$$S_\ell = \left[(F_\ell C_{\ell-1})^\top (F_\ell C_{\ell-1})\right]^{1/2} = \left(C_{\ell-1}^\top F_\ell^\top F_\ell C_{\ell-1}\right)^{1/2}.$$

Multiplying the polar decomposition by S_ℓ^{-1} gives

$$C_\ell = F_\ell C_{\ell-1} S_\ell^{-1}.$$

Moreover,

$$C_\ell^\top C_\ell = S_\ell^{-1} C_{\ell-1}^\top F_\ell^\top F_\ell C_{\ell-1} S_\ell^{-1} = S_\ell^{-1} S_\ell^2 S_\ell^{-1} = I.$$

Thus, the recursion is well defined. For the fixed initial frame C_0, uniqueness of the right polar decomposition at each layer gives uniqueness of all subsequent C_ℓ and S_ℓ. □

If $F_\ell = U_\ell H_\ell$ and $H_\ell = (F_\ell^\top F_\ell)^{1/2}$, is the ordinary right polar decomposition of the individual factor; then orthogonal equivariance of the positive square root gives

$$S_\ell = C_{\ell-1}^\top H_\ell C_{\ell-1}, \qquad C_\ell = U_\ell C_{\ell-1}. \tag{24}$$

This representation is also convenient for evaluating the geometry numerically.

A.2 Gauge independence of the shape functional

Although the recursive gauge depends on the initial orthogonal frame C_0, the resulting shape functional does not.

Lemma A.2 (Gauge independence)**.** *Let*

$$(C_\ell, S_\ell)_{\ell=1}^{L} \quad \textit{and} \quad (C'_\ell, S'_\ell)_{\ell=1}^{L}$$

be recursive polar gauges generated from two initial frames $C_0, C'_0 \in O(m)$. Then there exists a fixed $R \in O(m)$ such that

$$C'_\ell = C_\ell R, \qquad S'_\ell = R^\top S_\ell R \qquad \textit{for every } \ell \in [L]. \tag{25}$$

Consequently, if $Z_\ell = \log S_\ell$ and $Z'_\ell = \log S'_\ell$, then

$$Z'_\ell = R^\top Z_\ell R, \tag{26}$$

and

$$L \sum_{\ell=1}^{L-1} \|Z'_{\ell+1} - Z'_\ell\|_2 = L \sum_{\ell=1}^{L-1} \|Z_{\ell+1} - Z_\ell\|_2. \tag{27}$$

Proof. Write

$$C'_0 = C_0 R, \qquad R = C_0^\top C'_0 \in O(m).$$

Suppose inductively that $C'_{\ell-1} = C_{\ell-1}R$. From

$$F_\ell C_{\ell-1} = C_\ell S_\ell$$

we obtain

$$F_\ell C'_{\ell-1} = F_\ell C_{\ell-1} R = C_\ell S_\ell R = (C_\ell R)(R^\top S_\ell R).$$

The first factor is orthogonal and the second is positive definite. Uniqueness of the right polar decomposition therefore gives

$$C'_\ell = C_\ell R, \qquad S'_\ell = R^\top S_\ell R.$$

Induction proves (25). Functional calculus under orthogonal conjugation gives

$$\log(R^\top S_\ell R) = R^\top (\log S_\ell) R,$$

which proves (26). Since the operator norm is invariant under orthogonal conjugation,

$$\|Z'_{\ell+1} - Z'_\ell\|_2 = \|Z_{\ell+1} - Z_\ell\|_2$$

for every ℓ, proving (27). □

Hence,

$$\Theta_{\mathrm{sh}}(F) = L \sum_{\ell=1}^{L-1} \|Z_{\ell+1} - Z_\ell\|_2 \tag{28}$$

is well defined as a functional of the factorization, independently of the arbitrary initial frame C_0. Although Θ_{sh} is defined through adjacent differences, the centered-path estimate below shows that it controls every Z_ℓ relative to the mean logarithmic shape.

A.3 Centered path estimate

Define

$$\bar{Z} := \frac{1}{L} \sum_{j=1}^{L} Z_j.$$

Lemma A.3 (Centered path estimate)**.** *For every $\ell \in [L]$,*

$$\|Z_\ell - \bar{Z}\|_2 \le \frac{L-1}{L^2} \Theta_{\mathrm{sh}}(F). \tag{29}$$

Proof. By the triangle inequality,

$$\|Z_\ell - \bar{Z}\|_2 = \left\| \frac{1}{L} \sum_{j=1}^{L} (Z_\ell - Z_j) \right\|_2 \leq \frac{1}{L} \sum_{j=1}^{L} \|Z_\ell - Z_j\|_2.$$

Set

$$d_q := \|Z_{q+1} - Z_q\|_2, \qquad q = 1, \ldots, L-1.$$

For every j,

$$\|Z_\ell - Z_j\|_2 \leq \sum_{q=\min\{j,\ell\}}^{\max\{j,\ell\}-1} d_q \leq \sum_{q=1}^{L-1} d_q.$$

The term $j = \ell$ vanishes, so at most $L-1$ nonzero terms occur. Therefore,

$$\|Z_\ell - \bar{Z}\|_2 \leq \frac{L-1}{L} \sum_{q=1}^{L-1} d_q = \frac{L-1}{L^2} \Theta_{\mathrm{sh}}(F).$$

□

A.4 Layerwise and excluded-product spectral bounds

For each layer, define

$$a_\ell := \log \sigma_{\min}(F_\ell) = \lambda_{\min}(Z_\ell), \qquad b_\ell := \log \|F_\ell\|_2 = \lambda_{\max}(Z_\ell).$$

These identities hold because F_ℓ and S_ℓ have the same singular values. The extreme-eigenvalue maps are 1-Lipschitz with respect to the operator norm, so

$$|a_i - a_j| \leq \|Z_i - Z_j\|_2, \qquad |b_i - b_j| \leq \|Z_i - Z_j\|_2. \tag{30}$$

Thus, the matrix-valued shape control also bounds the variation of the layerwise logarithmic extreme singular values.

Let

$$\bar{a} := \frac{1}{L} \sum_{\ell=1}^{L} a_\ell = \frac{r_-(F)}{L}, \qquad \bar{b} := \frac{1}{L} \sum_{\ell=1}^{L} b_\ell = \frac{r_+(F)}{L}.$$

Applying the same path argument as in Lemma A.3 gives

$$|a_\ell - \bar{a}| \leq \frac{L-1}{L^2} \Theta_{\mathrm{sh}}(F), \qquad |b_\ell - \bar{b}| \leq \frac{L-1}{L^2} \Theta_{\mathrm{sh}}(F). \tag{31}$$

Lemma A.4 (Single-factor spectral control)**.** *If*

$$F \in \mathcal{B}(\rho_-, \rho_+, \theta),$$

then, for every $\ell \in [L]$,

$$\sigma_{\min}(F_\ell) \geq \exp\left[-\frac{\rho_-}{L} - \frac{L-1}{L^2}\theta \right], \tag{32}$$

and

$$\|F_\ell\|_2 \leq \exp\left[\frac{\rho_+}{L} + \frac{L-1}{L^2}\theta \right]. \tag{33}$$

Proof. Since $r_-(F) \geq -\rho_-$,

$$\bar{a} = \frac{r_-(F)}{L} \geq -\frac{\rho_-}{L}.$$

Combining this with (31) gives

$$a_\ell \geq -\frac{\rho_-}{L} - \frac{L-1}{L^2}\theta,$$

and exponentiation proves (32).

Similarly,

$$\bar{b} = \frac{r_+(F)}{L} \le \frac{\rho_+}{L},$$

so

$$b_\ell \le \frac{\rho_+}{L} + \frac{L-1}{L^2}\theta.$$

Exponentiation proves (33). □

Recall

$$Q_\ell = F_L \cdots F_{\ell+1}, \qquad R_\ell = F_{\ell-1} \cdots F_1.$$

Lemma A.5 (One-factor excluded-product control)**.** *If*

$$F \in \mathcal{B}(\rho_-, \rho_+, \theta),$$

then, for every $\ell \in [L]$,

$$\sigma_{\min}(Q_\ell)\sigma_{\min}(R_\ell) \ge e^{-A_-}, \tag{34}$$

and

$$\|Q_\ell\|_2 \|R_\ell\|_2 \le e^{A_+}, \tag{35}$$

where

$$A_- = \frac{L-1}{L}\rho_- + \frac{L-1}{L^2}\theta, \qquad A_+ = \frac{L-1}{L}\rho_+ + \frac{L-1}{L^2}\theta. \tag{36}$$

Proof. By multiplicativity of the smallest-singular-value lower bound,

$$\log[\sigma_{\min}(Q_\ell)\sigma_{\min}(R_\ell)] \ge \sum_{j \ne \ell} a_j = r_-(F) - a_\ell.$$

Equation (31) gives

$$a_\ell \le \frac{r_-(F)}{L} + \frac{L-1}{L^2}\theta.$$

Hence,

$$\begin{aligned} r_-(F) - a_\ell &\ge \frac{L-1}{L} r_-(F) - \frac{L-1}{L^2}\theta \\ &\ge -\frac{L-1}{L}\rho_- - \frac{L-1}{L^2}\theta \\ &= -A_-. \end{aligned}$$

Exponentiating proves (34).

For the upper bound,

$$\log[\|Q_\ell\|_2 \|R_\ell\|_2] \le \sum_{j \ne \ell} b_j = r_+(F) - b_\ell.$$

Again by (31),

$$b_\ell \ge \frac{r_+(F)}{L} - \frac{L-1}{L^2}\theta.$$

Therefore,

$$\begin{aligned} r_+(F) - b_\ell &\le \frac{L-1}{L} r_+(F) + \frac{L-1}{L^2}\theta \\ &\le \frac{L-1}{L}\rho_+ + \frac{L-1}{L^2}\theta \\ &= A_+. \end{aligned}$$

Exponentiating proves (35). □

The second-order product expansion used in the basin-smoothness argument requires the analogous estimate after excluding two factors.

Lemma A.6 (Two-factor excluded-product control). *Assume $L \geq 2$ and let $i, j \in [L]$ with $i \neq j$. If*

$$F \in \mathcal{B}(\rho_-, \rho_+, \theta),$$

then

$$\prod_{q \notin \{i,j\}} \|F_q\|_2 \leq e^{A_{++}}, \qquad A_{++} = \frac{L-2}{L}\rho_+ + \frac{2(L-1)}{L^2}\theta, \tag{37}$$

where the empty product is interpreted as 1 when $L = 2$.

Proof. By (31),

$$b_i + b_j \geq \frac{2}{L} r_+(F) - \frac{2(L-1)}{L^2}\theta.$$

Consequently,

$$\begin{aligned}\sum_{q \notin \{i,j\}} b_q &= r_+(F) - b_i - b_j \\ &\leq \frac{L-2}{L} r_+(F) + \frac{2(L-1)}{L^2}\theta \\ &\leq \frac{L-2}{L}\rho_+ + \frac{2(L-1)}{L^2}\theta \\ &= A_{++}.\end{aligned}$$

Exponentiating proves the claim. □

A.5 Exact bridge to additive balancedness

For

$$D_\ell = F_{\ell+1}^\top F_{\ell+1} - F_\ell F_\ell^\top,$$

the recursive polar gauge gives an exact correspondence between the additive Gram defect and the adjacent squared-polar defect.

Lemma A.7 (Polar representation of the additive defect). *For every $\ell = 1, \ldots, L-1$,*

$$S_{\ell+1}^2 - S_\ell^2 = C_\ell^\top D_\ell C_\ell. \tag{38}$$

In particular,

$$\|S_{\ell+1}^2 - S_\ell^2\|_2 = \|D_\ell\|_2. \tag{39}$$

Proof. From

$$F_{\ell+1} = C_{\ell+1} S_{\ell+1} C_\ell^\top,$$

we obtain

$$F_{\ell+1}^\top F_{\ell+1} = C_\ell S_{\ell+1}^2 C_\ell^\top.$$

Likewise,

$$F_\ell F_\ell^\top = C_\ell S_\ell^2 C_\ell^\top.$$

Subtracting gives

$$D_\ell = C_\ell (S_{\ell+1}^2 - S_\ell^2) C_\ell^\top,$$

which is equivalent to (38). Orthogonal invariance of the operator norm then gives (39). □

Thus, the additive Gram defect and the adjacent squared-polar defect are not merely comparable: they are orthogonally conjugate and therefore have exactly the same operator norm.

B Additive and Logarithmic Balancedness

This appendix establishes the noncommutative conversion between adjacent additive Gram defects and the logarithmic shape variables used in the finite-step invariance argument. The two directions require different spectral information: an upper spectral bound converts logarithmic mismatch into an additive defect, whereas a lower spectral floor converts an additive defect back into logarithmic mismatch.

Proposition B.1 (Additive versus logarithmic balance)**.** *Let $S, T \succ 0$.*

If

$$S, T \succeq \underline{s} I$$

for some $\underline{s} > 0$, then

$$\|\log S - \log T\|_2 \le \frac{1}{2\underline{s}^{\,2}} \|S^2 - T^2\|_2. \tag{40}$$

If

$$S, T \preceq \overline{s} I$$

for some $\overline{s} > 0$, then

$$\|S^2 - T^2\|_2 \le 2\overline{s}^{\,2} \|\log S - \log T\|_2. \tag{41}$$

Both statements hold without assuming that S and T commute.

B.1 Additive defect controls logarithmic mismatch

Proof of the first implication. Set

$$H := S - T, \qquad C := S^2 - T^2.$$

Then,

$$C = SH + HT. \tag{42}$$

Indeed,

$$SH + HT = S(S - T) + (S - T)T = S^2 - T^2.$$

This identity is exact and does not require S and T to commute.

Since $S, T \succeq \underline{s} I$, the Sylvester equation

$$SH + HT = C$$

has the integral representation

$$H = \int_0^\infty e^{-tS} C e^{-tT}\, dt. \tag{43}$$

To verify this, define

$$\Psi(t) := e^{-tS} C e^{-tT}.$$

Then,

$$-\Psi'(t) = S e^{-tS} C e^{-tT} + e^{-tS} C e^{-tT} T.$$

Since $S, T \succeq \underline{s} I$, $\Psi(t) \to 0$ as $t \to \infty$. Therefore,

$$\int_0^\infty -\Psi'(t)\, dt = \Psi(0) = C,$$

so the matrix on the right-hand side of (43) solves the Sylvester equation. The solution is unique because the spectra of S and $-T$ are disjoint.

Moreover,

$$\|e^{-tS}\|_2 \le e^{-\underline{s}t}, \qquad \|e^{-tT}\|_2 \le e^{-\underline{s}t},$$

and therefore,

$$\begin{aligned} \|S - T\|_2 &\le \int_0^\infty \|e^{-tS}\|_2 \|C\|_2 \|e^{-tT}\|_2\, dt \\ &\le \|C\|_2 \int_0^\infty e^{-2\underline{s}t}\, dt \\ &= \frac{1}{2\underline{s}} \|S^2 - T^2\|_2. \end{aligned} \tag{44}$$

We next use the standard integral representation of the matrix logarithm: for every positive-definite matrix R,

$$\log R = \int_0^\infty \left(\frac{1}{1+t} I - (R+tI)^{-1} \right) dt.$$

Subtracting the representations for S and T gives

$$\log S - \log T = \int_0^\infty \left[(T+tI)^{-1} - (S+tI)^{-1} \right] dt. \tag{45}$$

The resolvent identity yields

$$(T+tI)^{-1} - (S+tI)^{-1} = (S+tI)^{-1}(S-T)(T+tI)^{-1}. \tag{46}$$

Since $S, T \succeq \underline{s} I$,

$$\|(S+tI)^{-1}\|_2, \|(T+tI)^{-1}\|_2 \leq \frac{1}{\underline{s}+t}.$$

Consequently,

$$\begin{aligned} \|\log S - \log T\|_2 &\leq \|S-T\|_2 \int_0^\infty \frac{dt}{(\underline{s}+t)^2} \\ &= \frac{1}{\underline{s}} \|S-T\|_2. \end{aligned} \tag{47}$$

Combining (44) and (47) proves

$$\|\log S - \log T\|_2 \leq \frac{1}{2\underline{s}^2} \|S^2 - T^2\|_2.$$

□

B.2 Logarithmic mismatch controls additive defect

Proof of the second implication. Set

$$U := \log S, \qquad V := \log T,$$

so that $S = e^U$ and $T = e^V$. Since S and T are positive definite, U and V are symmetric.

The Duhamel identity

$$e^U - e^V = \int_0^1 e^{(1-t)U} (U-V) e^{tV} \, dt \tag{48}$$

holds without any commutativity assumption. To verify it, define

$$\Phi(t) := e^{(1-t)U} e^{tV}.$$

Differentiating the two exponential factors gives

$$\begin{aligned} \Phi'(t) &= -U e^{(1-t)U} e^{tV} + e^{(1-t)U} V e^{tV} \\ &= -e^{(1-t)U} (U-V) e^{tV}, \end{aligned}$$

where only the identities

$$U e^{(1-t)U} = e^{(1-t)U} U, \qquad V e^{tV} = e^{tV} V$$

are used; no commutativity between U and V is required. Integrating over $t \in [0,1]$ yields

$$e^V - e^U = -\int_0^1 e^{(1-t)U} (U-V) e^{tV} \, dt,$$

which proves (48).

The assumption $S, T \preceq \overline{s} I$ implies

$$\lambda_{\max}(U) \leq \log \overline{s}, \qquad \lambda_{\max}(V) \leq \log \overline{s}.$$

Hence, for every $t \in [0,1]$,

$$\|e^{(1-t)U}\|_2 \le \overline{s}^{\,1-t}, \qquad \|e^{tV}\|_2 \le \overline{s}^{\,t}.$$

Applying (48),

$$\begin{aligned}\|S - T\|_2 &\le \int_0^1 \overline{s}^{\,1-t}\overline{s}^{\,t}\|U - V\|_2\, dt \\ &= \overline{s}\|\log S - \log T\|_2. \end{aligned} \tag{49}$$

Finally,

$$S^2 - T^2 = S(S-T) + (S-T)T, \tag{50}$$

which again requires no commutativity. Therefore

$$\begin{aligned}\|S^2 - T^2\|_2 &\le (\|S\|_2 + \|T\|_2)\|S - T\|_2 \\ &\le 2\overline{s}^{\,2}\|\log S - \log T\|_2. \end{aligned} \tag{51}$$

This proves (41). □

B.3 Application to adjacent layers

We now apply Proposition B.1 to the recursive polar factors. Recall from Lemma A.7 that

$$\|S_{\ell+1}^2 - S_\ell^2\|_2 = \|D_\ell\|_2, \qquad D_\ell = F_{\ell+1}^\top F_{\ell+1} - F_\ell F_\ell^\top. \tag{52}$$

Corollary B.2 (Adjacent additive-to-log conversion)**.** *Suppose*

$$S_\ell, S_{\ell+1} \succeq \underline{s} I$$

for some $\underline{s} > 0$. Then,

$$\|Z_{\ell+1} - Z_\ell\|_2 \le \frac{1}{2\underline{s}^{\,2}}\|D_\ell\|_2, \tag{53}$$

where $Z_j = \log S_j$. Consequently, if the same lower spectral floor holds for every adjacent pair,

$$\Theta_{\mathrm{sh}}(F) \le \frac{L}{2\underline{s}^{\,2}} \sum_{\ell=1}^{L-1} \|D_\ell\|_2. \tag{54}$$

Proof. By (52),

$$\|S_{\ell+1}^2 - S_\ell^2\|_2 = \|D_\ell\|_2.$$

Applying (40) with $S = S_{\ell+1}$ and $T = S_\ell$ gives

$$\|Z_{\ell+1} - Z_\ell\|_2 \le \frac{1}{2\underline{s}^{\,2}}\|D_\ell\|_2.$$

Summing over ℓ and using

$$\Theta_{\mathrm{sh}}(F) = L \sum_{\ell=1}^{L-1} \|Z_{\ell+1} - Z_\ell\|_2$$

proves (54). □

Corollary B.3 (Adjacent log-to-additive conversion)**.** *Suppose*

$$S_\ell, S_{\ell+1} \preceq \overline{s} I$$

for some $\overline{s} > 0$. Then,

$$\|D_\ell\|_2 \le 2\overline{s}^{\,2}\|Z_{\ell+1} - Z_\ell\|_2. \tag{55}$$

Consequently, if the same upper spectral bound holds for every adjacent pair,

$$\sum_{\ell=1}^{L-1} \|D_\ell\|_2 \le \frac{2\overline{s}^{\,2}}{L}\Theta_{\mathrm{sh}}(F). \tag{56}$$

Proof. Equation (52) and (41) give

$$\|D_\ell\|_2 \le 2\overline{s}^{\,2}\|Z_{\ell+1} - Z_\ell\|_2.$$

Summing over ℓ and using the definition of Θ_{sh} proves (56). □

The two conversions play different roles in the finite-step argument. At initialization, an upper spectral bound converts the prescribed logarithmic shape into a bound on the additive Gram defects. The exact discrete imbalance identity then controls the subsequent change of these defects. At a candidate endpoint, a lower spectral floor converts the resulting additive defects back into logarithmic shape. In the first-exit proof, this lower floor is supplied by the spectral margin and relative-step admissibility before the hypothetical exit; it is not inferred from the desired endpoint shape bound.

C Proof of the Geometry-Sensitive Gradient Transmission Theorem

This appendix proves Theorem 4.1. The argument has three components: an exact decomposition of the least-squares excess, a prediction-space gradient bound, and transmission of that bound through the excluded layer products controlled by the spectral-balance geometry.

C.1 Least-squares excess decomposition

Recall

$$M(F) := BP(F)A, \qquad E(F) := M(F)X - Y.$$

Let

$$\Phi \in \arg\min_{\Theta \in \mathbb{R}^{k\times d}} \frac{1}{2}\|\Theta X - Y\|_F^2$$

be any least-squares minimizer, and define

$$R^\star := \Phi X - Y.$$

Lemma C.1 (Orthogonal decomposition of the excess loss)**.** *The least-squares residual satisfies*

$$R^\star X^\top = 0. \tag{57}$$

If

$$Z(F) := (M(F) - \Phi)X, \tag{58}$$

then

$$E(F) = Z(F) + R^\star, \qquad \langle Z(F), R^\star\rangle_F = 0. \tag{59}$$

Consequently,

$$\mathcal{E}(F) = \mathcal{L}(F) - \mathcal{L}^\star = \frac{1}{2}\|Z(F)\|_F^2, \tag{60}$$

and

$$E(F)X^\top = Z(F)X^\top. \tag{61}$$

Proof. Consider

$$f(\Theta) = \frac{1}{2}\|\Theta X - Y\|_F^2.$$

For any perturbation H,

$$\frac{d}{dt} f(\Phi + tH)\Big|_{t=0} = \langle \Phi X - Y, HX\rangle_F = \langle (\Phi X - Y)X^\top, H\rangle_F.$$

First-order optimality of the unconstrained least-squares problem therefore gives

$$R^\star X^\top = 0.$$

By definition,

$$E(F) = M(F)X - Y = (M(F) - \Phi)X + (\Phi X - Y) = Z(F) + R^\star.$$

Moreover,

$$\langle Z(F), R^\star\rangle_F = \operatorname{tr}\big((M(F)-\Phi)X(R^\star)^\top\big) = \operatorname{tr}\big((M(F)-\Phi)[R^\star X^\top]^\top\big) = 0.$$

Hence,

$$\|E(F)\|_F^2 = \|Z(F)\|_F^2 + \|R^\star\|_F^2.$$

Since

$$\mathcal{L}^\star = \frac{1}{2}\|R^\star\|_F^2,$$

we obtain (60). Finally,

$$E(F)X^\top = Z(F)X^\top + R^\star X^\top = Z(F)X^\top.$$

□

C.2 Prediction-space gradient bound

We first record two elementary matrix inequalities.

Lemma C.2 (Row-space singular-value bound)**.** *Let $X \in \mathbb{R}^{d\times n}$ have rank $r \geq 1$, and suppose that every row of $Z \in \mathbb{R}^{k\times n}$ belongs to* $\operatorname{row}(X)$*. Then*

$$\sigma_r(X)\|Z\|_F \leq \|ZX^\top\|_F \leq \|X\|_2\|Z\|_F. \tag{62}$$

Proof. Let

$$X = U_r\Sigma_r V_r^\top$$

be a compact singular-value decomposition. Since

$$\operatorname{row}(X) = \operatorname{span}(V_r^\top),$$

there exists $C \in \mathbb{R}^{k\times r}$ such that

$$Z = CV_r^\top.$$

Therefore,

$$\|Z\|_F = \|C\|_F$$

and

$$ZX^\top = C\Sigma_r U_r^\top, \qquad \|ZX^\top\|_F = \|C\Sigma_r\|_F.$$

The singular values of Σ_r then give

$$\sigma_r(X)\|C\|_F \leq \|C\Sigma_r\|_F \leq \sigma_1(X)\|C\|_F,$$

which proves the claim. □

Lemma C.3 (Two-sided Frobenius multiplication)**.** *For compatible matrices U, H, V,*

$$\|UHV\|_F \leq \|U\|_2\|V\|_2\|H\|_F. \tag{63}$$

If U has full column rank and V has full row rank, then

$$\|UHV\|_F \geq \sigma_{\min}(U)\sigma_{\min}(V)\|H\|_F. \tag{64}$$

Proof. The upper bound follows from

$$\|UHV\|_F \leq \|U\|_2\|HV\|_F \leq \|U\|_2\|V\|_2\|H\|_F.$$

For the lower bound, full column rank of U gives, for every compatible matrix K,

$$\|UK\|_F^2 = \sum_j \|Uk_j\|_2^2 \geq \sigma_{\min}(U)^2\|K\|_F^2.$$

Similarly, since V has full row rank, $V^\top$ has full column rank, and hence

$$\|HV\|_F = \|V^\top H^\top\|_F \geq \sigma_{\min}(V)\|H^\top\|_F = \sigma_{\min}(V)\|H\|_F.$$

Applying the first inequality with $K = HV$ and then the second gives

$$\|UHV\|_F \geq \sigma_{\min}(U)\|HV\|_F \geq \sigma_{\min}(U)\sigma_{\min}(V)\|H\|_F,$$

which proves (64). □

Lemma C.4 (Prediction-space gradient bound)**.** *Under the assumptions of Section 3,*

$$\sqrt{2}\,\alpha\sqrt{\mathcal{E}(F)} \le \|B^\top E(F) X^\top A^\top\|_F \le \sqrt{2}\,\beta\sqrt{\mathcal{E}(F)}, \tag{65}$$

where

$$\alpha = \sigma_{\min}(B)\sigma_{\min}(A)\sigma_r(X), \qquad \beta = \|B\|_2\|AX\|_2.$$

Proof. Every row of

$$Z(F) = (M(F) - \Phi)X$$

belongs to $\mathrm{row}(X)$. Moreover, Lemma C.1 gives

$$E(F)X^\top = Z(F)X^\top.$$

Because B has full row rank, $B^\top$ has full column rank, while the full column rank of A implies that $A^\top$ has full row rank. Hence Lemmas C.3 and C.2 yield

$$\begin{aligned}\|B^\top E(F) X^\top A^\top\|_F &= \|B^\top Z(F) X^\top A^\top\|_F \\ &\ge \sigma_{\min}(B)\sigma_{\min}(A)\|Z(F)X^\top\|_F \\ &\ge \sigma_{\min}(B)\sigma_{\min}(A)\sigma_r(X)\|Z(F)\|_F \\ &= \sqrt{2}\,\alpha\sqrt{\mathcal{E}(F)}.\end{aligned}$$

For the upper bound,

$$Z(F)X^\top A^\top = Z(F)(AX)^\top,$$

and therefore,

$$\begin{aligned}\|B^\top E(F) X^\top A^\top\|_F &= \|B^\top Z(F)(AX)^\top\|_F \\ &\le \|B\|_2\|AX\|_2\|Z(F)\|_F \\ &= \sqrt{2}\,\beta\sqrt{\mathcal{E}(F)}.\end{aligned}$$

□

C.3 Exact layer gradients

Lemma C.5 (Exact layer-gradient formula)**.** *Let*

$$Q_\ell = F_L \cdots F_{\ell+1}, \qquad R_\ell = F_{\ell-1}\cdots F_1,$$

with empty products equal to the identity. Then

$$G_\ell := \nabla_{F_\ell}\mathcal{L}(F) = Q_\ell^\top B^\top E(F) X^\top A^\top R_\ell^\top. \tag{66}$$

Proof. For a perturbation H of the ℓ-th factor,

$$DP(F)[H] = Q_\ell H R_\ell.$$

Hence,

$$\begin{aligned}D\mathcal{L}(F)[H] &= \langle E(F), BQ_\ell H R_\ell AX\rangle_F \\ &= \mathrm{tr}\big(E(F)^\top BQ_\ell H R_\ell AX\big) \\ &= \mathrm{tr}\big(R_\ell AX E(F)^\top BQ_\ell H\big) \\ &= \mathrm{tr}\Big(\big[Q_\ell^\top B^\top E(F) X^\top A^\top R_\ell^\top\big]^\top H\Big).\end{aligned}$$

Therefore,

$$\nabla_{F_\ell}\mathcal{L}(F) = Q_\ell^\top B^\top E(F) X^\top A^\top R_\ell^\top.$$

□

C.4 Transmission through excluded products

Lemma C.6 (Layerwise gradient transmission)**.** *Suppose*

$$F \in \mathcal{B}(\rho_-, \rho_+, \theta),$$

and define

$$A_- = \frac{L-1}{L}\rho_- + \frac{L-1}{L^2}\theta, \qquad A_+ = \frac{L-1}{L}\rho_+ + \frac{L-1}{L^2}\theta.$$

Then, for every $\ell \in [L]$,

$$\sqrt{2}\,\alpha e^{-A_-}\sqrt{\mathcal{E}(F)} \le \|G_\ell\|_F \le \sqrt{2}\,\beta e^{A_+}\sqrt{\mathcal{E}(F)}. \tag{67}$$

Proof. By Lemma C.5,

$$G_\ell = Q_\ell^\top \left[B^\top E(F) X^\top A^\top\right] R_\ell^\top.$$

Since the factors are invertible, Q_ℓ and R_ℓ are invertible (with the identity allowed for empty products). Therefore Lemma C.3 gives

$$\|G_\ell\|_F \ge \sigma_{\min}(Q_\ell)\sigma_{\min}(R_\ell)\|B^\top E(F) X^\top A^\top\|_F$$

and

$$\|G_\ell\|_F \le \|Q_\ell\|_2 \|R_\ell\|_2 \|B^\top E(F) X^\top A^\top\|_F.$$

By Lemma A.5,

$$\sigma_{\min}(Q_\ell)\sigma_{\min}(R_\ell) \ge e^{-A_-}, \qquad \|Q_\ell\|_2\|R_\ell\|_2 \le e^{A_+}.$$

Combining these estimates with Lemma C.4 proves (67). □

C.5 Proof of Theorem 4.1

Proof of Theorem 4.1. Lemma C.6 gives, for every $\ell \in [L]$,

$$\sqrt{2}\,\alpha e^{-A_-}\sqrt{\mathcal{E}(F)} \le \|G_\ell\|_F \le \sqrt{2}\,\beta e^{A_+}\sqrt{\mathcal{E}(F)}.$$

Squaring the lower bound and summing over the L layers yields

$$\|\nabla_F \mathcal{L}(F)\|_F^2 := \sum_{\ell=1}^{L} \|G_\ell\|_F^2 \ge 2L\alpha^2 e^{-2A_-}\mathcal{E}(F). \tag{68}$$

Define the depth-normalized PL coefficient

$$\mu(\rho_-, \theta) := \alpha^2 e^{-2A_-} = \alpha^2 \exp\left[-\frac{2(L-1)}{L}\rho_- - \frac{2(L-1)}{L^2}\theta\right]. \tag{69}$$

Then, (68) can be written as

$$\|\nabla_F \mathcal{L}(F)\|_F^2 \ge 2L\,\mu(\rho_-, \theta)\,\mathcal{E}(F). \tag{70}$$

Equivalently, under the standard convention

$$\frac{1}{2}\|\nabla_F \mathcal{L}(F)\|_F^2 \ge \mu_{\mathrm{PL}}\mathcal{E}(F),$$

the corresponding full-gradient PL coefficient is

$$\mu_{\mathrm{PL}} = L\,\mu(\rho_-, \theta).$$

This proves the theorem. □

Corollary C.7 (No nonoptimal stationary point in a finite basin)**.** *If*

$$F \in \mathcal{B}(\rho_-, \rho_+, \theta)$$

and

$$G_\ell = 0 \qquad \textit{for every } \ell \in [L],$$

then

$$\mathcal{E}(F) = 0, \qquad \mathcal{L}(F) = \mathcal{L}^\star.$$

Proof. Equation (68) gives

$$0 = \|\nabla_F \mathcal{L}(F)\|_F^2 \geq 2L\alpha^2 e^{-2A_-}\mathcal{E}(F).$$

Since $\alpha > 0$ and $A_- < \infty$, the prefactor is strictly positive. Therefore

$$\mathcal{E}(F) = 0,$$

and hence,

$$\mathcal{L}(F) = \mathcal{L}^\star.$$

□

D Relative-Step Smoothness and One-Step Descent

This appendix establishes the smoothness and one-step descent estimates used in the proof of Theorem 5.2. The key point is that the perturbation is measured relative to the current spectral scale of each factor. This yields segment bounds that depend on the total relative motion rather than on independent worst-case bounds for every prefix and suffix product. The required excluded-product estimates are provided by Appendix A.

D.1 Relative-step smoothness

Let

$$\Delta = (\Delta_1, \dots, \Delta_L)$$

be a simultaneous perturbation, and consider the segment

$$F_\ell(s) = F_\ell + s\Delta_\ell, \qquad s \in [0,1]. \tag{71}$$

Define

$$\delta_\ell := \frac{\|\Delta_\ell\|_2}{\sigma_{\min}(F_\ell)}. \tag{72}$$

Throughout this section we impose the aggregate relative-step condition

$$\sum_{\ell=1}^{L} \delta_\ell \leq \xi, \qquad 0 < \xi < 1. \tag{73}$$

In particular,

$$\delta_\ell \leq \xi < 1 \qquad \text{for every } \ell,$$

and Weyl's inequality gives

$$\begin{aligned} \sigma_{\min}(F_\ell(s)) &\geq \sigma_{\min}(F_\ell) - s\|\Delta_\ell\|_2 \\ &= (1 - s\delta_\ell)\sigma_{\min}(F_\ell) \\ &\geq (1 - \delta_\ell)\sigma_{\min}(F_\ell) > 0. \end{aligned} \tag{74}$$

Hence, every factor remains invertible along the entire segment.

Let

$$P(s) = F_L(s) \cdots F_1(s). \tag{75}$$

Lemma D.1 (Relative-step basin smoothness)**.** *Suppose*

$$F \in \mathcal{B}(\rho_-, \rho_+, \theta), \qquad \mathcal{L}(F) \leq \mathcal{L}_0,$$

and let Δ *satisfy* (73)*. Define*

$$A_+ = \frac{L-1}{L}\rho_+ + \frac{L-1}{L^2}\theta \tag{76}$$

and, for $L \geq 2$*,*

$$A_{++} = \frac{L-2}{L}\rho_+ + \frac{2(L-1)}{L^2}\theta. \tag{77}$$

Also define

$$\mathscr{R}_\xi := \sqrt{2\mathcal{L}_0} + \beta_F e^{\rho_+}(e^\xi - 1). \tag{78}$$

Then,

$$\mathcal{L}(F+\Delta) \le \mathcal{L}(F) + \sum_{\ell=1}^{L} \langle G_\ell, \Delta_\ell \rangle_F + \frac{L}{2} C_{\mathrm{sm}}^{\mathrm{rel}} \sum_{\ell=1}^{L} \|\Delta_\ell\|_F^2, \tag{79}$$

where, for $L \ge 2$*,*

$$C_{\mathrm{sm}}^{\mathrm{rel}} = \beta^2 e^{2(A_+ + \xi)} + \beta \mathscr{R}_\xi e^{A_{++} + \xi}. \tag{80}$$

For $L = 1$*, the second-derivative product term vanishes, and one may take*

$$C_{\mathrm{sm}}^{\mathrm{rel}} = \beta^2.$$

Proof. For each ℓ, define

$$Q_\ell(s) = F_L(s) \cdots F_{\ell+1}(s), \qquad R_\ell(s) = F_{\ell-1}(s) \cdots F_1(s).$$

Differentiating the product gives

$$P'(s) = \sum_{\ell=1}^{L} Q_\ell(s) \Delta_\ell R_\ell(s). \tag{81}$$

Since

$$\|\Delta_j\|_2 \le \delta_j \sigma_{\min}(F_j) \le \delta_j \|F_j\|_2,$$

we have

$$\|F_j(s)\|_2 \le \|F_j\|_2 + s\|\Delta_j\|_2 \le \|F_j\|_2 (1 + s\delta_j).$$

For the product excluding layer ℓ,

$$\prod_{j \ne \ell} (1 + s\delta_j) \le \exp\left(s \sum_{j \ne \ell} \delta_j \right) \le e^{\xi}.$$

Together with the one-factor excluded-product bound from Lemma A.5, this yields

$$\|Q_\ell(s)\|_2 \|R_\ell(s)\|_2 \le e^{A_+ + \xi}. \tag{82}$$

Therefore,

$$\|P'(s)\|_F \le e^{A_+ + \xi} \sum_{\ell=1}^{L} \|\Delta_\ell\|_F,$$

and Cauchy–Schwarz gives

$$\|P'(s)\|_F^2 \le L e^{2(A_+ + \xi)} \sum_{\ell=1}^{L} \|\Delta_\ell\|_F^2. \tag{83}$$

For $L \ge 2$, differentiating once more gives

$$P''(s) = 2 \sum_{1 \le i < j \le L} Q_j(s) \Delta_j M_{ji}(s) \Delta_i R_i(s), \tag{84}$$

where

$$M_{ji}(s) = F_{j-1}(s) \cdots F_{i+1}(s).$$

Each summand contains every factor except F_i and F_j. Hence,

$$\prod_{q \notin \{i,j\}} (1 + s\delta_q) \le \exp\left(s \sum_{q \notin \{i,j\}} \delta_q \right) \le e^{\xi}.$$

The two-factor excluded-product bound therefore gives

$$\|Q_j(s)\|_2\|M_{ji}(s)\|_2\|R_i(s)\|_2 \le e^{A_{++}+\xi}.$$

Consequently,

$$\begin{aligned}\|P''(s)\|_F &\le 2e^{A_{++}+\xi}\sum_{i<j}\|\Delta_i\|_F\|\Delta_j\|_F \\ &\le (L-1)e^{A_{++}+\xi}\sum_{\ell=1}^{L}\|\Delta_\ell\|_F^2 \\ &\le Le^{A_{++}+\xi}\sum_{\ell=1}^{L}\|\Delta_\ell\|_F^2. \end{aligned} \tag{85}$$

Here, we use

$$2\sum_{i<j}a_ia_j \le (L-1)\sum_{i=1}^{L}a_i^2.$$

For $L = 1$, $P''(s) = 0$. We next bound the residual along the segment. For each summand in (81),

$$\begin{aligned}\|Q_\ell(s)\Delta_\ell R_\ell(s)\|_2 &\le \left(\prod_{j\neq\ell}\|F_j\|_2(1+s\delta_j)\right)\|\Delta_\ell\|_2 \\ &\le e^{r_+(F)}\delta_\ell\prod_{j\neq\ell}(1+s\delta_j) \\ &\le e^{r_+(F)}e^{s\sum_j\delta_j}\delta_\ell.\end{aligned}$$

Summing over ℓ and using $r_+(F) \le \rho_+$ gives

$$\|P'(s)\|_2 \le e^{\rho_+}e^{s\sum_j\delta_j}\sum_{\ell=1}^{L}\delta_\ell \le e^{\rho_+}e^{s\xi}\xi. \tag{86}$$

Therefore, for every $s \in [0, 1]$,

$$\begin{aligned}\|P(s) - P(0)\|_2 &\le \int_0^s \|P'(u)\|_2\,du \\ &\le e^{\rho_+}\int_0^s \xi e^{u\xi}\,du \\ &= e^{\rho_+}(e^{s\xi}-1) \\ &\le e^{\rho_+}(e^{\xi}-1).\end{aligned} \tag{87}$$

Let

$$E(s) = BP(s)AX - Y.$$

Since

$$\|E(0)\|_F = \sqrt{2\mathcal{L}(F)} \le \sqrt{2\mathcal{L}_0},$$

we obtain

$$\begin{aligned}\|E(s)\|_F &\le \|E(0)\|_F + \|B(P(s)-P(0))AX\|_F \\ &\le \sqrt{2\mathcal{L}_0} + \beta_F e^{\rho_+}(e^{\xi}-1) \\ &= \mathscr{R}_\xi.\end{aligned} \tag{88}$$

Define

$$\phi(s) = \frac{1}{2}\|E(s)\|_F^2.$$

Then,

$$\phi''(s) = \|BP'(s)AX\|_F^2 + \langle E(s), BP''(s)AX\rangle_F. \tag{89}$$

Using

$$\|BHAX\|_F \le \|B\|_2\|AX\|_2\|H\|_F = \beta\|H\|_F,$$

together with (83), (85), and (88), gives, for $L \ge 2$,

$$\phi''(s) \le L\left[\beta^2 e^{2(A_+ + \xi)} + \beta\mathscr{R}_\xi e^{A_{++}+\xi}\right]\sum_{\ell=1}^{L}\|\Delta_\ell\|_F^2. \tag{90}$$

For $L = 1$, the $P''(s)$ term vanishes and

$$\phi''(s) \le \beta^2\|\Delta_1\|_F^2.$$

Finally, Taylor's formula with integral remainder,

$$\phi(1) = \phi(0) + \phi'(0) + \int_0^1 (1-s)\phi''(s)\,ds,$$

together with

$$\phi'(0) = \sum_{\ell=1}^{L}\langle G_\ell, \Delta_\ell\rangle_F,$$

yields (79). □

For $L \ge 2$, the two curvature contributions in $C_{\mathrm{sm}}^{\mathrm{rel}}$ have the same leading shape exponent:

$$2\frac{L-1}{L^2}\theta$$

in the first term, and

$$\frac{2(L-1)}{L^2}\theta$$

through A_{++} in the second. At depth two, this exponent is $\theta/2$, the same exponential scale that appears in the pure-shape stability obstruction of Section 7. This agreement concerns the exponential scale; it does not assert that the sufficient smoothness constant is sharp.

D.2 One-step descent and relative spectral safety

We now apply the relative-step estimate to a GD update

$$\Delta_\ell = -\eta G_\ell, \qquad \eta = \frac{\eta_0}{L}.$$

Let

$$\mathcal{B}_1 = \mathcal{B}(\rho_-^{(1)}, \rho_+^{(1)}, \theta_1)$$

be the outer basin. Define

$$A_+^{(1)} = \frac{L-1}{L}\rho_+^{(1)} + \frac{L-1}{L^2}\theta_1$$

and

$$A_{\mathrm{rel}}^{(1)} := A_+^{(1)} + \frac{\rho_-^{(1)}}{L} + \frac{L-1}{L^2}\theta_1. \tag{91}$$

Equivalently,

$$A_{\mathrm{rel}}^{(1)} = \frac{L-1}{L}\rho_+^{(1)} + \frac{\rho_-^{(1)}}{L} + \frac{2(L-1)}{L^2}\theta_1. \tag{92}$$

Inside $\mathcal{B}_1$, the layerwise gradient-transmission bound from Theorem 4.1 gives

$$\|G_\ell\|_F \le \sqrt{2}\,\beta e^{A_+^{(1)}}\sqrt{\mathcal{E}}. \tag{93}$$

The single-factor spectral estimate from Lemma A.4 gives

$$\sigma_{\min}(F_\ell)^{-1} \le \exp\left[\frac{\rho_-^{(1)}}{L} + \frac{L-1}{L^2}\theta_1\right]. \tag{94}$$

Therefore,

$$\sum_{\ell=1}^{L} \frac{\eta\|G_\ell\|_2}{\sigma_{\min}(F_\ell)} \le \sum_{\ell=1}^{L} \frac{\eta\|G_\ell\|_F}{\sigma_{\min}(F_\ell)} \le \sqrt{2}\,\eta_0\beta e^{A^{(1)}_{\mathrm{rel}}}\sqrt{\mathcal{E}}. \tag{95}$$

Thus, whenever

$$F \in \mathcal{B}_1 \qquad \text{and} \qquad \mathcal{E}(F) \le \mathcal{E}_0,$$

the aggregate relative-step condition (73) is guaranteed by

$$\eta_0 \le \frac{\xi e^{-A^{(1)}_{\mathrm{rel}}}}{\sqrt{2}\,\beta\sqrt{\mathcal{E}_0}}. \tag{96}$$

To formulate the bootstrap without circularity, let $t_\star$ denote a hypothetical first exit time from $\mathcal{B}_1$. Then,

$$F^{(t)} \in \mathcal{B}_1 \qquad \text{for every } t < t_\star.$$

The estimates below are asserted only for $t < t_\star$; the proof of Theorem 5.2 will subsequently rule out the exit.

Define

$$\mathscr{R}^{(1)}_\xi = \sqrt{2\mathcal{L}_0} + \beta_F e^{\rho^{(1)}_+}(e^\xi - 1) \tag{97}$$

and, for $L \ge 2$,

$$C^{(1)}_{\mathrm{sm}}(\xi) = \beta^2 e^{2(A^{(1)}_+ + \xi)} + \beta\mathscr{R}^{(1)}_\xi e^{A^{(1)}_{++}+\xi}, \tag{98}$$

where

$$A^{(1)}_{++} = \frac{L-2}{L}\rho^{(1)}_+ + \frac{2(L-1)}{L^2}\theta_1.$$

For $L = 1$, set

$$C^{(1)}_{\mathrm{sm}}(\xi) := \beta^2.$$

Assume

$$\eta_0 C^{(1)}_{\mathrm{sm}}(\xi) \le 1. \tag{99}$$

We now prove inductively that, for every $t < t_\star$,

$$\mathcal{L}(F^{(t)}) \le \mathcal{L}_0, \qquad \mathcal{E}_t \le \mathcal{E}_0,$$

and that the corresponding GD segment satisfies the relative-step condition. The claim holds at $t = 0$. If it holds at time $t < t_\star$, then (95) and (96) imply

$$\sum_{\ell=1}^{L} \frac{\eta\|G^{(t)}_\ell\|_2}{\sigma_{\min}(F^{(t)}_\ell)} \le \xi.$$

Hence, Lemma D.1, applied with $\Delta_\ell = -\eta G^{(t)}_\ell$, gives

$$\mathcal{L}(F^{(t+1)}) \le \mathcal{L}(F^{(t)}) - \eta\left(1 - \frac{\eta L}{2}C^{(1)}_{\mathrm{sm}}(\xi)\right)\sum_{\ell=1}^{L}\|G^{(t)}_\ell\|_F^2 \le \mathcal{L}(F^{(t)}) - \frac{\eta}{2}\sum_{\ell=1}^{L}\|G^{(t)}_\ell\|_F^2. \tag{100}$$

Thus the loss, and therefore the excess loss, is nonincreasing before the hypothetical first exit. This closes the induction.

Let

$$A^{(1)}_- = \frac{L-1}{L}\rho^{(1)}_- + \frac{L-1}{L^2}\theta_1,$$

and define the depth-normalized PL coefficient

$$\mu_1 := \alpha^2 e^{-2A^{(1)}_-}. \tag{101}$$

For every $t < t_\star$, Theorem 4.1 gives

$$\sum_{\ell=1}^{L} \|G_\ell^{(t)}\|_F^2 \geq 2L\mu_1 \mathcal{E}_t.$$

Substituting this into (100) and using $\eta = \eta_0/L$ yields

$$\mathcal{E}_{t+1} \leq (1 - \eta_0\mu_1)\mathcal{E}_t, \qquad t < t_\star. \tag{102}$$

To interpret the right-hand side as a nonnegative contraction factor, assume

$$0 < \eta_0\mu_1 \leq 1. \tag{103}$$

Equivalently,

$$\eta_0 \leq \frac{e^{2A_-^{(1)}}}{\alpha^2} = \frac{1}{\mu_1}. \tag{104}$$

Therefore,

$$\mathcal{E}_t \leq (1 - \eta_0\mu_1)^t \mathcal{E}_0, \qquad t < t_\star. \tag{105}$$

These estimates are deliberately restricted to the pre-exit interval. Once Theorem 5.2 rules out a finite first exit, the same descent and contraction bounds hold for all $t \geq 0$.

E Discrete Imbalance Dynamics and Shape Stability

This appendix establishes the shape-control component of Theorem 5.2. The argument has two steps. First, the adjacent Gram imbalance evolves only at second order under simultaneous gradient descent because its first-order drift cancels exactly. Second, the additive-to-logarithmic comparison from Appendix B converts control of this discrete defect into control of the logarithmic shape variable.

For $L = 1$, there are no adjacent interfaces and $\Theta_{\mathrm{sh}} \equiv 0$, so the shape-stability statement is immediate. We therefore assume $L \geq 2$ throughout this appendix.

E.1 Exact discrete imbalance dynamics

Recall the adjacent Gram imbalance

$$D_\ell = F_{\ell+1}^\top F_{\ell+1} - F_\ell F_\ell^\top, \qquad \ell = 1, \dots, L-1. \tag{106}$$

The layer gradients satisfy an exact compatibility identity.

Lemma E.1 (Adjacent gradient identity)**.** *For every $\ell = 1, \dots, L-1$,*

$$G_\ell F_\ell^\top = F_{\ell+1}^\top G_{\ell+1}. \tag{107}$$

Proof. Recall

$$Q_\ell = F_L \cdots F_{\ell+1}, \qquad R_\ell = F_{\ell-1} \cdots F_1.$$

Then,

$$Q_\ell = Q_{\ell+1}F_{\ell+1}, \qquad R_{\ell+1} = F_\ell R_\ell.$$

Using the exact layer-gradient formula,

$$G_\ell = Q_\ell^\top B^\top E X^\top A^\top R_\ell^\top,$$

we obtain

$$\begin{aligned} G_\ell F_\ell^\top &= F_{\ell+1}^\top Q_{\ell+1}^\top B^\top E X^\top A^\top R_\ell^\top F_\ell^\top \\ &= F_{\ell+1}^\top Q_{\ell+1}^\top B^\top E X^\top A^\top R_{\ell+1}^\top \\ &= F_{\ell+1}^\top G_{\ell+1}. \end{aligned}$$

□

Consider one simultaneous GD update

$$F_\ell^+ = F_\ell - \eta G_\ell.$$

Expanding the adjacent defect gives

$$\begin{aligned} D_\ell^+ - D_\ell = &- \eta \left(G_{\ell+1}^\top F_{\ell+1} + F_{\ell+1}^\top G_{\ell+1} \right) \\ &+ \eta \left(G_\ell F_\ell^\top + F_\ell G_\ell^\top \right) \\ &+ \eta^2 \left(G_{\ell+1}^\top G_{\ell+1} - G_\ell G_\ell^\top \right). \end{aligned} \tag{108}$$

Equation (107), together with its transpose, cancels the first-order terms exactly. Therefore, along the GD trajectory,

$$D_\ell^{(t+1)} - D_\ell^{(t)} = \eta^2 \left[(G_{\ell+1}^{(t)})^\top G_{\ell+1}^{(t)} - G_\ell^{(t)} (G_\ell^{(t)})^\top \right]. \tag{109}$$

Thus, the discrete change of adjacent Gram imbalance is second order in the step size.

Taking operator norms and summing over interfaces gives

$$\begin{aligned} \sum_{\ell=1}^{L-1} \left\| D_\ell^{(t+1)} - D_\ell^{(t)} \right\|_2 &\le \eta^2 \sum_{\ell=1}^{L-1} \left(\|G_{\ell+1}^{(t)}\|_2^2 + \|G_\ell^{(t)}\|_2^2 \right) \\ &\le 2\eta^2 \sum_{\ell=1}^{L} \|G_\ell^{(t)}\|_F^2. \end{aligned} \tag{110}$$

Let $t_\star$ denote a hypothetical first exit time from the outer basin $\mathcal{B}_1$. For every $t < t_\star$, Appendix D gives

$$\mathcal{L}(F^{(t+1)}) \le \mathcal{L}(F^{(t)}) - \frac{\eta}{2} \sum_{\ell=1}^{L} \|G_\ell^{(t)}\|_F^2.$$

Since $\mathcal{L}^\star$ is constant, the same inequality holds for the excess loss:

$$\mathcal{E}_{t+1} \le \mathcal{E}_t - \frac{\eta}{2} \sum_{\ell=1}^{L} \|G_\ell^{(t)}\|_F^2.$$

Consequently, for every integer T satisfying

$$1 \le T \le t_\star,$$

telescoping from $t = 0$ to $T - 1$ yields

$$\begin{aligned} \frac{\eta}{2} \sum_{t=0}^{T-1} \sum_{\ell=1}^{L} \|G_\ell^{(t)}\|_F^2 &\le \mathcal{E}_0 - \mathcal{E}_T \\ &\le \mathcal{E}_0. \end{aligned} \tag{111}$$

Hence, by the triangle inequality in time,

$$\begin{aligned} \sum_{\ell=1}^{L-1} \|D_\ell^{(T)} - D_\ell^{(0)}\|_2 &\le \sum_{t=0}^{T-1} \sum_{\ell=1}^{L-1} \|D_\ell^{(t+1)} - D_\ell^{(t)}\|_2 \\ &\le 2\eta^2 \sum_{t=0}^{T-1} \sum_{\ell=1}^{L} \|G_\ell^{(t)}\|_F^2 \\ &\le 4\eta \mathcal{E}_0 \\ &= \frac{4\eta_0}{L} \mathcal{E}_0. \end{aligned} \tag{112}$$

E.2 From additive imbalance to shape stability

The recursive polar gauge identifies the adjacent Gram defect with the difference of the squared polar factors. By Lemma A.7,

$$S_{\ell+1}^2 - S_\ell^2 = C_\ell^\top D_\ell C_\ell, \tag{113}$$

and therefore,

$$\|S_{\ell+1}^2 - S_\ell^2\|_2 = \|D_\ell\|_2. \tag{114}$$

Suppose that, at a given endpoint,

$$\sigma_{\min}(F_\ell) \geq m \qquad \text{for every } \ell.$$

Since F_ℓ and S_ℓ have the same singular values,

$$S_\ell \succeq mI.$$

Corollary B.2 then gives

$$\|Z_{\ell+1} - Z_\ell\|_2 \leq \frac{1}{2m^2}\|D_\ell\|_2. \tag{115}$$

Summing over adjacent layers yields

$$\Theta_{\mathrm{sh}}(F) \leq \frac{L}{2m^2}\sum_{\ell=1}^{L-1}\|D_\ell\|_2. \tag{116}$$

We now specialize this estimate to the first-exit argument. Define the single-factor lower spectral scale of the outer basin by

$$m_1 := \exp\left[-\frac{\rho_-^{(1)}}{L} - \frac{L-1}{L^2}\theta_1\right]. \tag{117}$$

For every pre-exit iterate $t < t_\star$,

$$F^{(t)} \in \mathcal{B}_1,$$

and Lemma A.4 therefore gives

$$\sigma_{\min}(F_\ell^{(t)}) \geq m_1 \qquad \text{for every } \ell. \tag{118}$$

At the candidate first-exit endpoint $t_\star$, membership in $\mathcal{B}_1$ has not yet been established and therefore cannot be used to infer a spectral floor. Instead, the relative-step condition on the segment from $F^{(t_\star-1)}$ to $F^{(t_\star)}$ gives

$$\sigma_{\min}(F_\ell^{(t_\star)}) \geq \left(1 - \delta_\ell^{(t_\star-1)}\right)\sigma_{\min}(F_\ell^{(t_\star-1)}).$$

Since

$$\sum_{\ell=1}^{L}\delta_\ell^{(t_\star-1)} \leq \xi,$$

we have

$$\delta_\ell^{(t_\star-1)} \leq \xi \qquad \text{for every } \ell.$$

Combining this with (118) yields

$$\sigma_{\min}(F_\ell^{(t_\star)}) \geq m_{1,\xi}, \qquad m_{1,\xi} := (1-\xi)m_1. \tag{119}$$

This lower bound uses only pre-exit basin membership and relative-step safety; it does not assume that the candidate endpoint already satisfies the desired shape bound.

More generally, the same endpoint floor holds for every $1 \leq T \leq t_\star$: for $T < t_\star$ it follows directly from (118), since $m_1 \geq m_{1,\xi}$, while for $T = t_\star$ it follows from (119). Therefore, (116) gives

$$\Theta_{\mathrm{sh}}(F^{(T)}) \leq \frac{L}{2m_{1,\xi}^2}\sum_{\ell=1}^{L-1}\|D_\ell^{(T)}\|_2, \qquad 1 \leq T \leq t_\star. \tag{120}$$

It remains to control the initial additive defects. Define

$$M_0 := \exp\left[\frac{\rho_+^{(0)}}{L} + \frac{L-1}{L^2}\theta_0\right]. \tag{121}$$

Since

$$F^{(0)} \in \mathcal{B}_0,$$

Lemma A.4 gives

$$\|F_\ell^{(0)}\|_2 \le M_0.$$

Because $F_\ell^{(0)}$ and $S_\ell^{(0)}$ have the same singular values,

$$S_\ell^{(0)} \preceq M_0 I.$$

Corollary B.3 therefore yields

$$\|D_\ell^{(0)}\|_2 \le 2M_0^2 \|Z_{\ell+1}^{(0)} - Z_\ell^{(0)}\|_2.$$

Summing over ℓ and using

$$\Theta_{\mathrm{sh}}(F^{(0)}) = L\sum_{\ell=1}^{L-1} \|Z_{\ell+1}^{(0)} - Z_\ell^{(0)}\|_2 \le \theta_0$$

gives

$$\sum_{\ell=1}^{L-1} \|D_\ell^{(0)}\|_2 \le \frac{2M_0^2}{L}\theta_0. \tag{122}$$

For every $1 \le T \le t_\star$,

$$\sum_{\ell=1}^{L-1} \|D_\ell^{(T)}\|_2 \le \sum_{\ell=1}^{L-1} \|D_\ell^{(0)}\|_2 + \sum_{\ell=1}^{L-1} \|D_\ell^{(T)} - D_\ell^{(0)}\|_2.$$

Combining (122), (112), and (120) yields

$$\Theta_{\mathrm{sh}}(F^{(T)}) \le \frac{M_0^2}{m_{1,\xi}^2}\theta_0 + \frac{2\eta_0}{m_{1,\xi}^2}\mathcal{E}_0 \qquad \text{for every } 1 \le T \le t_\star. \tag{123}$$

In particular, the strict shape-margin condition

$$\frac{M_0^2}{m_{1,\xi}^2}\theta_0 + \frac{2\eta_0}{m_{1,\xi}^2}\mathcal{E}_0 < \theta_1 \tag{124}$$

implies

$$\Theta_{\mathrm{sh}}(F^{(t_\star)}) < \theta_1.$$

Hence, the shape coordinate cannot be responsible for a hypothetical first exit from the outer basin.

The prefactor

$$\frac{M_0^2}{m_{1,\xi}^2}$$

arises from two distinct noncommutative conversions: the upper spectral bound at initialization converts logarithmic shape into additive Gram imbalance, while the lower spectral floor at the candidate endpoint converts the resulting additive defect back into logarithmic shape. No sharpness claim is made for this prefactor; the obstruction results concern the principal exponential radial and shape scales.

F Radial Stability

This appendix establishes the radial-control component of Theorem 5.2. The argument combines the relative-step safety from Appendix D with the geometric decay of the excess loss before a hypothetical first exit. The resulting cumulative motion of both radial coordinates is finite and remains strictly within the prescribed outer-basin margins.

F.1 One-step radial motion

Recall

$$r_-(F) = \sum_{\ell=1}^{L} \log \sigma_{\min}(F_\ell), \qquad r_+(F) = \sum_{\ell=1}^{L} \log \|F_\ell\|_2.$$

Consider a GD step

$$F_\ell^+ = F_\ell - \eta G_\ell,$$

and define

$$\delta_\ell := \frac{\eta \|G_\ell\|_2}{\sigma_{\min}(F_\ell)}. \tag{125}$$

Before a hypothetical first exit, the relative-step condition from Appendix D guarantees

$$\sum_{\ell=1}^{L} \delta_\ell \le \xi, \qquad 0 < \xi < 1. \tag{126}$$

In particular,

$$0 \le \delta_\ell \le \xi < 1 \qquad \text{for every } \ell.$$

For the lower radial coordinate, Weyl's inequality gives

$$\sigma_{\min}(F_\ell^+) \ge \sigma_{\min}(F_\ell) - \eta \|G_\ell\|_2 = (1 - \delta_\ell)\sigma_{\min}(F_\ell),$$

while

$$\sigma_{\min}(F_\ell^+) \le \sigma_{\min}(F_\ell) + \eta \|G_\ell\|_2 = (1 + \delta_\ell)\sigma_{\min}(F_\ell).$$

Therefore,

$$1 - \delta_\ell \le \frac{\sigma_{\min}(F_\ell^+)}{\sigma_{\min}(F_\ell)} \le 1 + \delta_\ell.$$

Since $\delta_\ell \le \xi < 1$,

$$-\log(1 - \delta_\ell) \le \frac{\delta_\ell}{1 - \xi}, \qquad \log(1 + \delta_\ell) \le \delta_\ell \le \frac{\delta_\ell}{1 - \xi}.$$

Hence,

$$\left|\log \sigma_{\min}(F_\ell^+) - \log \sigma_{\min}(F_\ell)\right| \le \frac{\delta_\ell}{1 - \xi}. \tag{127}$$

For the upper radial coordinate, the perturbation inequality for the operator norm gives

$$\left|\|F_\ell^+\|_2 - \|F_\ell\|_2\right| \le \eta \|G_\ell\|_2.$$

Since

$$\sigma_{\min}(F_\ell) \le \|F_\ell\|_2,$$

we have

$$\frac{\eta \|G_\ell\|_2}{\|F_\ell\|_2} \le \delta_\ell.$$

Consequently,

$$(1 - \delta_\ell)\|F_\ell\|_2 \le \|F_\ell^+\|_2 \le (1 + \delta_\ell)\|F_\ell\|_2.$$

The same logarithmic estimates therefore imply

$$\left|\log \|F_\ell^+\|_2 - \log \|F_\ell\|_2\right| \le \frac{\delta_\ell}{1 - \xi}. \tag{128}$$

Summing either (127) or (128) over layers gives

$$|r_\pm(F^+) - r_\pm(F)| \le \frac{1}{1 - \xi} \sum_{\ell=1}^{L} \delta_\ell. \tag{129}$$

Let $t_\star$ denote a hypothetical first exit time from the outer basin $\mathcal{B}_1$. For every $t < t_\star$,

$$F^{(t)} \in \mathcal{B}_1.$$

The GD relative-step estimate from Appendix D therefore gives

$$\sum_{\ell=1}^{L} \delta_\ell^{(t)} \le \sqrt{2}\,\eta_0 \beta e^{A_{\mathrm{rel}}^{(1)}} \sqrt{\mathcal{E}_t}.$$

Combining this estimate with (129) yields

$$|r_\pm(F^{(t+1)}) - r_\pm(F^{(t)})| \le \frac{\sqrt{2}\,\eta_0 \beta e^{A_{\mathrm{rel}}^{(1)}}}{1-\xi} \sqrt{\mathcal{E}_t}, \qquad t < t_\star. \tag{130}$$

F.2 Summable radial motion

For every $t < t_\star$, the one-step descent and gradient-transmission estimates give

$$\mathcal{E}_{t+1} \le (1 - \eta_0 \mu_1)\mathcal{E}_t, \qquad \mu_1 = \alpha^2 e^{-2A_-^{(1)}}. \tag{131}$$

The contraction-safety condition

$$0 < \eta_0 \mu_1 \le 1$$

therefore implies, inductively,

$$\mathcal{E}_t \le (1 - \eta_0 \mu_1)^t \mathcal{E}_0, \qquad 0 \le t \le t_\star. \tag{132}$$

Here, the estimate at $t = t_\star$ follows by applying the one-step contraction at time $t_\star - 1$; no membership of $F^{(t_\star)}$ in the outer basin is used.

For $x \in (0, 1]$,

$$1 - \sqrt{1-x} = \frac{x}{1+\sqrt{1-x}} \ge \frac{x}{2}.$$

Hence,

$$\sum_{t=0}^{\infty} (1 - \eta_0 \mu_1)^{t/2} = \frac{1}{1 - \sqrt{1 - \eta_0 \mu_1}} \le \frac{2}{\eta_0 \mu_1}. \tag{133}$$

Let T be any integer satisfying

$$1 \le T \le t_\star.$$

Summing (130) over $t = 0, \dots, T-1$ and using (132) gives

$$\begin{aligned} |r_\pm(F^{(T)}) - r_\pm(F^{(0)})| &\le \frac{\sqrt{2}\,\eta_0 \beta e^{A_{\mathrm{rel}}^{(1)}}}{1-\xi} \sqrt{\mathcal{E}_0} \sum_{t=0}^{T-1} (1 - \eta_0 \mu_1)^{t/2} \\ &\le \frac{2\sqrt{2}\,\beta}{(1-\xi)\mu_1} e^{A_{\mathrm{rel}}^{(1)}} \sqrt{\mathcal{E}_0} \\ &= C_{\mathrm{rad}}^{(1)} \sqrt{\mathcal{E}_0}, \end{aligned} \tag{134}$$

where

$$C_{\mathrm{rad}}^{(1)} := \frac{2\sqrt{2}\,\beta}{(1-\xi)\alpha^2} \exp\left(A_{\mathrm{rel}}^{(1)} + 2A_-^{(1)}\right). \tag{135}$$

In the last equality, we use

$$\mu_1^{-1} = \alpha^{-2} e^{2A_-^{(1)}}.$$

Thus,

$$\sup_{1 \le T \le t_\star} |r_\pm(F^{(T)}) - r_\pm(F^{(0)})| \le C_{\mathrm{rad}}^{(1)} \sqrt{\mathcal{E}_0}. \tag{136}$$

A notable feature of (134) is that the explicit stepsize η_0 cancels. Reducing the step size decreases each individual radial displacement, while the corresponding linear-convergence timescale grows proportionally to its reciprocal. This observation concerns the resulting upper bound; it does not assert that the actual trajectory length is stepsize-independent.

F.3 Radial margin condition

Assume the strict radial margin conditions

$$\rho_-^{(1)} - \rho_-^{(0)} > C_{\mathrm{rad}}^{(1)}\sqrt{\mathcal{E}_0}, \tag{137}$$

and

$$\rho_+^{(1)} - \rho_+^{(0)} > C_{\mathrm{rad}}^{(1)}\sqrt{\mathcal{E}_0}. \tag{138}$$

These are equivalently summarized by

$$C_{\mathrm{rad}}^{(1)}\sqrt{\mathcal{E}_0} < \min\left\{\rho_-^{(1)} - \rho_-^{(0)},\, \rho_+^{(1)} - \rho_+^{(0)}\right\}.$$

At a hypothetical first-exit endpoint $t_\star$,

$$r_-(F^{(0)}) \geq -\rho_-^{(0)}.$$

Therefore,

$$\begin{aligned} r_-(F^{(t_\star)}) &\geq r_-(F^{(0)}) - \left|r_-(F^{(t_\star)}) - r_-(F^{(0)})\right| \\ &\geq -\rho_-^{(0)} - C_{\mathrm{rad}}^{(1)}\sqrt{\mathcal{E}_0} \\ &> -\rho_-^{(1)}. \end{aligned} \tag{139}$$

Similarly,

$$r_+(F^{(0)}) \leq \rho_+^{(0)}$$

implies

$$\begin{aligned} r_+(F^{(t_\star)}) &\leq r_+(F^{(0)}) + \left|r_+(F^{(t_\star)}) - r_+(F^{(0)})\right| \\ &\leq \rho_+^{(0)} + C_{\mathrm{rad}}^{(1)}\sqrt{\mathcal{E}_0} \\ &< \rho_+^{(1)}. \end{aligned} \tag{140}$$

Hence, neither radial coordinate can reach its corresponding outer-basin boundary at a hypothetical first exit. The use of absolute radial motion in (134) is conservative, motion in the opposite direction is harmless, but provides a uniform sufficient condition for the first-exit argument.

G Proof of the Forward-Invariance Theorem

This appendix combines the relative-step smoothness, shape-stability, and radial-stability estimates from Appendices D–F to prove Theorem 5.2. The remaining step is to close these estimates simultaneously through a nested first-exit argument.

G.1 Nested first-exit argument

Proof of Theorem 5.2. Write

$$\mathcal{B}_0 = \mathcal{B}\big(\rho_-^{(0)}, \rho_+^{(0)}, \theta_0\big), \qquad \mathcal{B}_1 = \mathcal{B}\big(\rho_-^{(1)}, \rho_+^{(1)}, \theta_1\big),$$

with the strict nesting conditions

$$\rho_-^{(1)} > \rho_-^{(0)}, \qquad \rho_+^{(1)} > \rho_+^{(0)}, \qquad \theta_1 > \theta_0.$$

Let

$$F^{(0)} \in \mathcal{B}_0$$

have invertible factors, and set

$$\mathcal{E}_0 = \mathcal{E}(F^{(0)}).$$

If $\mathcal{E}_0 = 0$, then Lemma C.4 gives

$$B^\top E(F^{(0)}) X^\top A^\top = 0.$$

The exact layer-gradient formula therefore implies

$$G_\ell^{(0)} = 0 \qquad \text{for every } \ell.$$

Hence the GD trajectory is stationary and remains in $\mathcal{B}_1$ for all time. We therefore assume throughout the remainder of the proof that

$$\mathcal{E}_0 > 0.$$

Fix $0 < \xi < 1$ and choose

$$\eta = \frac{\eta_0}{L}, \qquad \eta_0 > 0,$$

such that

$$\eta_0 \le \min\left\{ \frac{1}{C_{\mathrm{sm}}^{(1)}(\xi)},\ \frac{\xi e^{-A_{\mathrm{rel}}^{(1)}}}{\sqrt{2}\,\beta\sqrt{\mathcal{E}_0}},\ \frac{e^{2A_-^{(1)}}}{\alpha^2} \right\}. \tag{141}$$

Assume also the strict radial-margin conditions

$$\rho_-^{(1)} - \rho_-^{(0)} > C_{\mathrm{rad}}^{(1)}\sqrt{\mathcal{E}_0}, \tag{142}$$

and

$$\rho_+^{(1)} - \rho_+^{(0)} > C_{\mathrm{rad}}^{(1)}\sqrt{\mathcal{E}_0}, \tag{143}$$

together with the strict shape-margin condition

$$\frac{M_0^2}{m_{1,\xi}^2}\theta_0 + \frac{2\eta_0}{m_{1,\xi}^2}\mathcal{E}_0 < \theta_1. \tag{144}$$

Suppose, toward a contradiction, that the trajectory leaves $\mathcal{B}_1$. Let

$$t_\star := \min\left\{ t \ge 1 : F^{(t)} \notin \mathcal{B}_1 \right\}.$$

Then,

$$F^{(t)} \in \mathcal{B}_1 \qquad \text{for every } 0 \le t < t_\star. \tag{145}$$

Pre-exit descent and contraction. We first verify the estimates used along the pre-exit trajectory. We claim inductively that

$$\mathcal{E}_t \le \mathcal{E}_0, \qquad \mathcal{L}(F^{(t)}) \le \mathcal{L}(F^{(0)}), \qquad 0 \le t \le t_\star. \tag{146}$$

The claim is immediate at $t = 0$. Suppose it holds at some $t < t_\star$. By (145),

$$F^{(t)} \in \mathcal{B}_1.$$

The layerwise gradient-transmission estimate, the single-factor spectral bound, and the second condition in (141) therefore give

$$\sum_{\ell=1}^{L} \frac{\eta \|G_\ell^{(t)}\|_2}{\sigma_{\min}(F_\ell^{(t)})} \le \xi. \tag{147}$$

Hence, the complete GD segment from $F^{(t)}$ to $F^{(t+1)}$ satisfies the relative-step assumptions of Appendix D.

The first condition in (141) gives

$$\eta_0 C_{\mathrm{sm}}^{(1)}(\xi) \le 1.$$

The one-step smoothness estimate therefore yields

$$\mathcal{L}(F^{(t+1)}) \le \mathcal{L}(F^{(t)}) - \frac{\eta}{2}\sum_{\ell=1}^{L} \|G_\ell^{(t)}\|_F^2, \qquad t < t_\star. \tag{148}$$

In particular, both the loss and the excess loss are nonincreasing, which closes the induction in (146), including the endpoint $t = t_\star$.

Define the depth-normalized PL coefficient of the outer basin by

$$\mu_1 := \alpha^2 e^{-2A_-^{(1)}}. \tag{149}$$

For every $t < t_\star$, the gradient-transmission theorem gives

$$\sum_{\ell=1}^{L} \|G_\ell^{(t)}\|_F^2 \geq 2L\mu_1 \mathcal{E}_t. \tag{150}$$

Substituting this into (148) and using $\eta = \eta_0/L$ yields

$$\mathcal{E}_{t+1} \leq (1 - \eta_0\mu_1)\mathcal{E}_t, \qquad t < t_\star. \tag{151}$$

The third condition in (141) is exactly

$$\eta_0 \leq \frac{1}{\mu_1}.$$

Since $\eta_0 > 0$ and $\mu_1 > 0$,

$$0 < \eta_0\mu_1 \leq 1.$$

Iterating (151) therefore gives

$$\mathcal{E}_t \leq (1 - \eta_0\mu_1)^t \mathcal{E}_0, \qquad 0 \leq t \leq t_\star. \tag{152}$$

The value at $t = t_\star$ uses only the contraction step based at $F^{(t_\star-1)} \in \mathcal{B}_1$; no basin membership of the candidate endpoint is assumed.

Endpoint spectral safety. The relative-step condition (147) on the final pre-exit segment $F^{(t_\star-1)} \to F^{(t_\star)}$ guarantees that every factor at the candidate endpoint remains invertible. Indeed, since

$$F^{(t_\star-1)} \in \mathcal{B}_1,$$

Lemma A.4 gives

$$\sigma_{\min}(F_\ell^{(t_\star-1)}) \geq m_1.$$

Moreover, if

$$\delta_\ell^{(t_\star-1)} := \frac{\eta\|G_\ell^{(t_\star-1)}\|_2}{\sigma_{\min}(F_\ell^{(t_\star-1)})},$$

then

$$\sum_{\ell=1}^{L} \delta_\ell^{(t_\star-1)} \leq \xi,$$

and hence,

$$\delta_\ell^{(t_\star-1)} \leq \xi \qquad \text{for every } \ell.$$

Weyl's inequality therefore yields

$$\begin{aligned}\sigma_{\min}(F_\ell^{(t_\star)}) &\geq \left(1 - \delta_\ell^{(t_\star-1)}\right) \sigma_{\min}(F_\ell^{(t_\star-1)}) \\ &\geq (1-\xi)m_1 = m_{1,\xi} > 0.\end{aligned} \tag{153}$$

Thus the recursive polar gauge, logarithmic factors, radial coordinates, and shape coordinate are all well defined at the candidate exit endpoint.

Radial coordinates. Appendix F gives

$$|r_\pm(F^{(t_\star)}) - r_\pm(F^{(0)})| \leq C_{\mathrm{rad}}^{(1)}\sqrt{\mathcal{E}_0}. \tag{154}$$

Since

$$r_-(F^{(0)}) \geq -\rho_-^{(0)},$$

the lower radial margin (142) gives

$$\begin{aligned}r_-(F^{(t_\star)}) &\geq -\rho_-^{(0)} - C_{\mathrm{rad}}^{(1)}\sqrt{\mathcal{E}_0} \\ &> -\rho_-^{(1)}.\end{aligned} \tag{155}$$

Similarly,

$$r_+(F^{(0)}) \le \rho_+^{(0)}$$

and (143) imply

$$\begin{aligned} r_+(F^{(t_\star)}) &\le \rho_+^{(0)} + C_{\mathrm{rad}}^{(1)}\sqrt{\mathcal{E}_0} \\ &< \rho_+^{(1)}. \end{aligned} \tag{156}$$

Shape coordinate. Appendix E gives

$$\Theta_{\mathrm{sh}}(F^{(t_\star)}) \le \frac{M_0^2}{m_{1,\xi}^2}\theta_0 + \frac{2\eta_0}{m_{1,\xi}^2}\mathcal{E}_0. \tag{157}$$

By the strict shape margin (144),

$$\Theta_{\mathrm{sh}}(F^{(t_\star)}) < \theta_1. \tag{158}$$

We have therefore shown simultaneously that

$$r_-(F^{(t_\star)}) > -\rho_-^{(1)}, \qquad r_+(F^{(t_\star)}) < \rho_+^{(1)}, \qquad \Theta_{\mathrm{sh}}(F^{(t_\star)}) < \theta_1,$$

while all factors of $F^{(t_\star)}$ remain invertible. Hence,

$$F^{(t_\star)} \in \operatorname{int}\mathcal{B}_1 \subseteq \mathcal{B}_1,$$

contradicting the definition of $t_\star$ as the first iterate outside $\mathcal{B}_1$. Therefore, no finite first-exit time exists, and

$$F^{(t)} \in \mathcal{B}_1 \qquad \text{for every } t \ge 0. \tag{159}$$

In particular, every factor remains invertible for all time. Since the relative-step smoothness and gradient-transmission estimates are now valid at every iteration, the preceding pre-exit estimates extend globally:

$$\mathcal{L}(F^{(t+1)}) \le \mathcal{L}(F^{(t)}) - \frac{\eta}{2}\sum_{\ell=1}^{L}\|G_\ell^{(t)}\|_F^2, \qquad t \ge 0, \tag{160}$$

and

$$\mathcal{E}_{t+1} \le (1-\eta_0\mu_1)\mathcal{E}_t, \qquad t \ge 0. \tag{161}$$

This completes the proof. □

H Proof of the Global Linear-Convergence Theorem

This appendix proves Theorem 6.1. Once forward invariance has been established, the argument is direct: the spectral-balance geometry remains valid along the entire trajectory, so the gradient-transmission inequality and the one-step descent estimate can be applied uniformly at every iteration.

Throughout, we retain the hypotheses of Theorem 5.2. In particular,

$$\eta = \frac{\eta_0}{L}, \qquad \eta_0 > 0,$$

and we define the depth-normalized PL coefficient

$$\mu_1 := \alpha^2 e^{-2A_-^{(1)}}, \qquad A_-^{(1)} = \frac{L-1}{L}\rho_-^{(1)} + \frac{L-1}{L^2}\theta_1. \tag{162}$$

We also write

$$\mathcal{E}_t = \mathcal{L}(F^{(t)}) - \mathcal{L}^\star.$$

H.1 Uniform contraction along the invariant trajectory

By Theorem 5.2,

$$F^{(t)} \in \mathcal{B}_1 = \mathcal{B}\big(\rho_-^{(1)}, \rho_+^{(1)}, \theta_1\big) \qquad \text{for every } t \geq 0.$$

Hence, heorem 4.1 applies at every iterate and gives the uniform gradient-transmission bound

$$\sum_{\ell=1}^{L} \|G_\ell^{(t)}\|_F^2 \geq 2L\mu_1 \mathcal{E}_t. \tag{163}$$

Equivalently, in terms of the full product-space gradient,

$$\|\nabla_F \mathcal{L}(F^{(t)})\|_F^2 := \sum_{\ell=1}^{L} \|G_\ell^{(t)}\|_F^2 \geq 2L\mu_1 \mathcal{E}_t.$$

Forward invariance also makes the relative-step descent estimate valid at every iteration:

$$\mathcal{L}(F^{(t+1)}) \leq \mathcal{L}(F^{(t)}) - \frac{\eta}{2} \sum_{\ell=1}^{L} \|G_\ell^{(t)}\|_F^2. \tag{164}$$

Subtracting $\mathcal{L}^\star$ and applying (163) yields

$$\begin{aligned} \mathcal{E}_{t+1} &\leq \mathcal{E}_t - \frac{\eta}{2}\,(2L\mu_1 \mathcal{E}_t) \\ &= (1 - \eta L \mu_1)\mathcal{E}_t \\ &= (1 - \eta_0 \mu_1)\mathcal{E}_t. \end{aligned} \tag{165}$$

The stepsize condition in Theorem 5.2 includes

$$\eta_0 \leq \frac{e^{2A_-^{(1)}}}{\alpha^2} = \frac{1}{\mu_1}.$$

Since $\eta_0 > 0$ and $\mu_1 > 0$,

$$0 < \eta_0 \mu_1 \leq 1. \tag{166}$$

Therefore (165) is a valid contraction inequality at every iteration.

H.2 Proof of Theorem 6.1

Proof of Theorem 6.1. Iterating (165) gives

$$\mathcal{E}_t \leq (1 - \eta_0 \mu_1)^t \mathcal{E}_0 \qquad \text{for every } t \geq 0. \tag{167}$$

Substituting

$$\mu_1 = \alpha^2 e^{-2A_-^{(1)}}$$

gives

$$\mathcal{E}_t \leq \left(1 - \eta_0 \alpha^2 e^{-2A_-^{(1)}}\right)^t \mathcal{E}_0, \tag{168}$$

which is the claimed global-in-time linear rate within the invariant outer basin.

Since

$$0 \leq 1 - \eta_0 \mu_1 < 1,$$

it follows that

$$\mathcal{E}_t \longrightarrow 0.$$

Hence,

$$\mathcal{L}(F^{(t)}) \longrightarrow \mathcal{L}^\star. \tag{169}$$

For the iteration complexity, use

$$1 - x \leq e^{-x} \qquad \text{for } x \in [0, 1].$$

Equation (167) therefore implies

$$\mathcal{E}_t \le e^{-\eta_0 \mu_1 t} \mathcal{E}_0. \tag{170}$$

For any

$$0 < \varepsilon < \mathcal{E}_0,$$

it is sufficient that

$$e^{-\eta_0 \mu_1 t} \mathcal{E}_0 \le \varepsilon,$$

or equivalently,

$$t \ge \frac{1}{\eta_0 \mu_1} \log \frac{\mathcal{E}_0}{\varepsilon}.$$

Substituting (162) gives the sufficient iteration bound

$$t \ge \frac{e^{2A_-^{(1)}}}{\eta_0 \alpha^2} \log \frac{\mathcal{E}_0}{\varepsilon}. \tag{171}$$

□

H.3 Interpretation of the guarantee

Recall that

$$\mathcal{L}^\star = \min_{\Theta \in \mathbb{R}^{k \times d}} \frac{1}{2} \|\Theta X - Y\|_F^2.$$

Thus, the theorem establishes convergence of the training loss to the unconstrained least-squares value induced by the data:

$$\lim_{t \to \infty} \mathcal{L}(F^{(t)}) = \min_{\Theta \in \mathbb{R}^{k \times d}} \frac{1}{2} \|\Theta X - Y\|_F^2. \tag{172}$$

No interpolation assumption is required: the optimal least-squares residual may be nonzero. Likewise, X need not have full row rank; the rate depends on its smallest nonzero singular value through

$$\alpha = \sigma_{\min}(B) \sigma_{\min}(A) \sigma_r(X).$$

The conclusion is an objective-value guarantee. It does not assert convergence to a unique parameter-space factorization, nor is such uniqueness expected in general: distinct collections of factors can represent the same end-to-end prediction map.

Finally, the depth-normalized stepsize explains why the contraction coefficient itself contains no additional explicit multiplicative factor of L. The full-gradient transmission inequality contributes a factor L,

$$\|\nabla_F \mathcal{L}(F)\|_F^2 \ge 2L\mu_1 \mathcal{E}(F),$$

while

$$\eta = \frac{\eta_0}{L}$$

cancels this factor in (165). Depth nevertheless remains in the contraction coefficient through

$$A_-^{(1)} = \frac{L-1}{L} \rho_-^{(1)} + \frac{L-1}{L^2} \theta_1,$$

and additional depth dependence may enter through the admissible stepsize and basin conditions required for forward invariance. Thus the theorem does not assert a depth-independent convergence guarantee.

I Proof of the Coordinate-Wise Obstruction Theorem

This appendix proves the three obstruction mechanisms summarized in Theorem 7.1. The constructions are scalar subclasses of the model and are therefore valid special cases of the matrix-valued setting. They establish worst-case limitations on the geometric dependences appearing in the sufficient theory. The radial-collapse construction concerns the gradient-transmission scale, whereas the radial-expansion and pure-shape constructions concern uniform one-step loss-nonincreasing stability

under finite-step GD. None of the results below is a converse to Theorem 5.2, and failure of one-step loss nonincrease does not imply failure of eventual convergence.

Throughout, take

$$A = B = X = 1.$$

For scalar factors $f_1, \dots, f_L$, write

$$p := \prod_{\ell=1}^{L} f_\ell,$$

and, for target $y \in \{0, 1\}$, define

$$\mathcal{L}_y(f_1, \dots, f_L) = \frac{1}{2}(p - y)^2. \tag{173}$$

The scalar layer gradient is

$$g_\ell = \frac{\partial \mathcal{L}_y}{\partial f_\ell} = (p - y) \prod_{j \neq \ell} f_j. \tag{174}$$

When all factors are positive,

$$r_- = r_+ = \sum_{\ell=1}^{L} \log f_\ell = \log p. \tag{175}$$

At depth $L = 2$,

$$\Theta_{\mathrm{sh}} = 2|\log f_2 - \log f_1|. \tag{176}$$

For a fixed initialization F, call a normalized stepsize $\eta_0 \geq 0$ one-step stable if

$$\mathcal{L}(F^+_{\eta_0}) \leq \mathcal{L}(F), \tag{177}$$

and let $\eta_{0,\mathrm{conn}}(F)$ denote the right endpoint of the connected one-step-stable interval containing $\eta_0 = 0$.

I.1 Radial collapse and exact gradient-transmission degeneration

Proposition I.1 (Exact radial-collapse obstruction)**.** *For every $L \geq 2$ and $\rho > 0$, there exists an exactly shape-balanced scalar factorization satisfying*

$$r_- = -\rho, \qquad \Theta_{\mathrm{sh}} = 0,$$

for which

$$\frac{\sum_{\ell=1}^{L} g_\ell^2}{2L\mathcal{L}_1} = \exp\left[-\frac{2(L-1)}{L}\rho\right]. \tag{178}$$

Hence, the lower-radial exponential dependence in the gradient-transmission inequality of Theorem 4.1 is attained exactly.

Proof. Take

$$f_1 = \cdots = f_L = s, \qquad s = e^{-\rho/L} \in (0, 1), \tag{179}$$

and set $y = 1$. Then,

$$p = s^L = e^{-\rho}, \qquad r_- = r_+ = -\rho, \qquad \Theta_{\mathrm{sh}} = 0.$$

By (174),

$$g_\ell = (s^L - 1)s^{L-1}$$

for every ℓ, so

$$\sum_{\ell=1}^{L} g_\ell^2 = L(1 - s^L)^2 s^{2L-2}.$$

Since

$$2L\mathcal{L}_1 = L(1 - s^L)^2,$$

we obtain

$$\frac{\sum_\ell g_\ell^2}{2L\mathcal{L}_1} = s^{2L-2} = \exp\left[-\frac{2(L-1)}{L}\rho\right].$$

□

Remark I.2 (Singular collapse boundary). At the limiting singular factorization

$$f_1 = \cdots = f_L = 0, \qquad L \geq 2,$$

with $y = 1$, every layer gradient vanishes because every excluded product contains a zero factor, while

$$\mathcal{L}_1 = \frac{1}{2}.$$

Thus, the singular boundary contains a non-global stationary point. The preceding proposition is more informative for the present purpose because it shows the exact exponential degeneration of the gradient-transmission ratio while all factors remain invertible.

I.2 Radial expansion and exact one-step stability

Proposition I.3 (Exact radial-expansion one-step threshold)**.** *For every $L \geq 2$ and $\rho > 0$, there exists an exactly shape-balanced scalar factorization satisfying*

$$r_+ = \rho, \qquad \Theta_{\mathrm{sh}} = 0,$$

such that, under the depth-normalized step

$$\eta = \frac{\eta_0}{L},$$

the connected one-step-stable interval containing zero is

$$0 \leq \eta_0 \leq 2L \exp\left[-\frac{2(L-1)}{L}\rho\right]. \tag{180}$$

Equivalently,

$$\eta_{0,\mathrm{conn}} = 2L \exp\left[-\frac{2(L-1)}{L}\rho\right]. \tag{181}$$

Proof. Set $y = 0$ and take

$$f_1 = \cdots = f_L = s, \qquad s = e^{\rho/L} > 1.$$

Then,

$$p = s^L, \qquad r_- = r_+ = \rho, \qquad \Theta_{\mathrm{sh}} = 0.$$

The scalar loss is

$$\mathcal{L}_0(f) = \frac{1}{2}s^{2L},$$

and

$$g_\ell = s^{2L-1}.$$

One simultaneous GD step gives

$$f_\ell^+ = s\left(1 - \frac{\eta_0}{L}s^{2L-2}\right).$$

Hence,

$$p^+ = s^L\left(1 - \frac{\eta_0}{L}s^{2L-2}\right)^L,$$

and therefore,

$$\frac{\mathcal{L}_0(f^+)}{\mathcal{L}_0(f)} = \left|1 - \frac{\eta_0}{L}s^{2L-2}\right|^{2L}. \tag{182}$$

Loss nonincrease is equivalent to

$$\left|1 - \frac{\eta_0}{L}s^{2L-2}\right| \leq 1.$$

For $\eta_0 \geq 0$, this holds exactly when

$$0 \leq \eta_0 \leq 2Ls^{-2L+2},$$

which becomes

$$0 \leq \eta_0 \leq 2L \exp\left[-\frac{2(L-1)}{L}\rho\right].$$

Because this stable set is itself a connected interval containing zero, its right endpoint is precisely (181). □

Corollary I.4 (Upper-radial obstruction to uniform one-step descent)**.** *Any one-step loss-nonincreasing GD guarantee that is uniform over exactly balanced scalar factorizations with $r_+ = \rho$ must allow the admissible normalized stepsize scale to deteriorate at least at exponential order*

$$\exp\left[-\frac{2(L-1)}{L}\rho\right].$$

Proof. Proposition I.3 provides, for every $\rho > 0$, an instance whose exact connected one-step-stable normalized stepsize threshold is

$$2Le^{-2(L-1)\rho/L}.$$

A uniform one-step loss-nonincreasing guarantee with an exponentially larger admissible stepsize scale would fail on this family. □

I.3 Pure shape imbalance at neutral radial scale

We next keep the end-to-end product and both radial coordinates fixed while allowing only inter-layer shape to increase.

Lemma I.5 (Exact inverse-pair one-step map)**.** *Let $L = 2$, $y = 0$, and*

$$f_1 = a, \qquad f_2 = a^{-1}, \qquad a \geq 1.$$

Then,

$$P(F) = 1, \qquad r_- = r_+ = 0, \qquad \Theta_{\mathrm{sh}} = 4\log a.$$

Under the depth-normalized step

$$\eta = \frac{\eta_0}{2},$$

the one-step loss ratio is

$$\frac{\mathcal{L}_0(f^+)}{\mathcal{L}_0(f)} = \left[1 - \eta_0 \cosh\left(\frac{\Theta_{\mathrm{sh}}}{2}\right) + \frac{\eta_0^2}{4}\right]^2. \tag{183}$$

Proof. Since $f_1 f_2 = 1$,

$$r_- = r_+ = 0.$$

At depth two,

$$\Theta_{\mathrm{sh}} = 2|\log a^{-1} - \log a| = 4\log a.$$

Moreover,

$$a^2 + a^{-2} = 2\cosh\left(\frac{\Theta_{\mathrm{sh}}}{2}\right).$$

Because $p = 1$ and $y = 0$,

$$\mathcal{L}_0(f_1, f_2) = \frac{1}{2},$$

while

$$g_1 = a^{-1}, \qquad g_2 = a.$$

Thus,

$$f_1^+ = a - \frac{\eta_0}{2a}, \qquad f_2^+ = a^{-1} - \frac{\eta_0 a}{2}.$$

Multiplication gives

$$p^+ = 1 - \frac{\eta_0}{2}(a^2 + a^{-2}) + \frac{\eta_0^2}{4},$$

and therefore

$$p^+ = 1 - \eta_0 \cosh\left(\frac{\Theta_{\mathrm{sh}}}{2}\right) + \frac{\eta_0^2}{4}.$$

Since

$$\frac{\mathcal{L}_0(f^+)}{\mathcal{L}_0(f)} = (p^+)^2,$$

the result follows. □

Lemma I.6 (Connected pure-shape stability threshold)**.** *Let*

$$C_\Theta := \cosh\left(\frac{\Theta_{\mathrm{sh}}}{2}\right).$$

If

$$C_\Theta \geq \sqrt{2},$$

then the connected component containing $\eta_0 = 0$ of the one-step-stable set has right endpoint

$$\eta_{0,\mathrm{conn}}(\Theta_{\mathrm{sh}}) = 2C_\Theta - 2\sqrt{C_\Theta^2 - 2}. \tag{184}$$

Proof. By Lemma I.5, one-step loss nonincrease is equivalent to

$$|q(\eta_0)| \leq 1, \qquad q(\eta_0) = 1 - C_\Theta \eta_0 + \frac{\eta_0^2}{4}.$$

Now

$$q(0) = 1, \qquad q'(\eta_0) = -C_\Theta + \frac{\eta_0}{2},$$

so q decreases immediately to the right of zero and attains its minimum at

$$\eta_0 = 2C_\Theta.$$

At that point,

$$q_{\min} = 1 - C_\Theta^2.$$

When $C_\Theta \geq \sqrt{2}$, we have $q_{\min} \leq -1$, so the first boundary encountered by the stable component containing zero occurs at

$$q(\eta_0) = -1.$$

Solving

$$1 - C_\Theta \eta_0 + \frac{\eta_0^2}{4} = -1$$

gives

$$\eta_0 = 2C_\Theta \pm 2\sqrt{C_\Theta^2 - 2}.$$

The first connected boundary is the smaller positive root, which proves (184). □

Lemma I.7 (Asymptotic pure-shape threshold)**.** *As $\Theta_{\mathrm{sh}} \to \infty$,*

$$\eta_{0,\mathrm{conn}}(\Theta_{\mathrm{sh}}) \sim 4e^{-\Theta_{\mathrm{sh}}/2}. \tag{185}$$

Proof. Rationalizing the exact expression gives

$$\eta_{0,\mathrm{conn}} = \frac{4}{C_\Theta + \sqrt{C_\Theta^2 - 2}}.$$

As $\Theta_{\mathrm{sh}} \to \infty$,

$$C_\Theta = \cosh\left(\frac{\Theta_{\mathrm{sh}}}{2}\right) \sim \frac{1}{2} e^{\Theta_{\mathrm{sh}}/2},$$

and

$$\sqrt{C_\Theta^2 - 2} \sim C_\Theta.$$

Hence,

$$\eta_{0,\mathrm{conn}} \sim \frac{2}{C_\Theta} \sim 4e^{-\Theta_{\mathrm{sh}}/2}.$$

□

Proposition I.8 (Pure-shape obstruction to uniform one-step stability)**.** *For every fixed $\eta_0 > 0$, there exists a depth-two scalar factorization with*

$$P(F) = 1, \qquad r_- = r_+ = 0,$$

and sufficiently large Θ_{sh} for which the first GD step increases the loss. Along the inverse-pair family,

$$\eta_{0,\mathrm{conn}}(\Theta_{\mathrm{sh}}) \sim 4e^{-\Theta_{\mathrm{sh}}/2}. \tag{186}$$

Proof. The asymptotic threshold follows from Lemma I.7. For the fixed-stepsize statement, fix $\eta_0 > 0$. From (183),

$$q(\eta_0) = 1 - \eta_0 \cosh\left(\frac{\Theta_{\mathrm{sh}}}{2}\right) + \frac{\eta_0^2}{4}.$$

As $\Theta_{\mathrm{sh}} \to \infty$,

$$q(\eta_0) \sim -\frac{\eta_0}{2} e^{\Theta_{\mathrm{sh}}/2},$$

and hence,

$$|q(\eta_0)| \to \infty.$$

Therefore, for sufficiently large Θ_{sh},

$$\mathcal{L}_0(f^+) > \mathcal{L}_0(f).$$

More quantitatively,

$$\frac{\mathcal{L}_0(f^+)}{\mathcal{L}_0(f)} = q(\eta_0)^2 \sim \frac{\eta_0^2}{4} e^{\Theta_{\mathrm{sh}}},$$

so the one-step loss amplification is unbounded along this radially neutral family. □

I.4 An independent local calculation at the minimizer manifold

The same depth-two exponential scale also appears in an independent local linear-stability calculation around the manifold of equivalent global minimizers.

Proposition I.9 (Normal linear stability of the inverse-pair minimizers)**.** *Let $L = 2$, $y = 1$, and*

$$(f_1, f_2) = (a, a^{-1}), \qquad a \geq 1.$$

Then, this point is a global minimizer, and the Hessian is

$$H_a = \begin{pmatrix} a^{-2} & 1 \\ 1 & a^2 \end{pmatrix}. \tag{187}$$

Its eigenvalues are

$$0, \qquad a^2 + a^{-2}.$$

For GD with raw step $\eta = \eta_0/2$, the normal linearized direction is strictly contractive exactly when

$$0 < \eta_0 < \frac{4}{a^2 + a^{-2}} = 2 \operatorname{sech}\left(\frac{\Theta_{\mathrm{sh}}}{2}\right). \tag{188}$$

The right-hand side satisfies

$$2 \operatorname{sech}\left(\frac{\Theta_{\mathrm{sh}}}{2}\right) \sim 4e^{-\Theta_{\mathrm{sh}}/2}.$$

Proof. With $y = 1$,

$$\mathcal{L}_1(f_1, f_2) = \frac{1}{2}(f_1 f_2 - 1)^2.$$

At (a, a^{-1}), the residual vanishes, so the point is a global minimizer. The gradient is

$$\nabla \mathcal{L}_1 = (f_1 f_2 - 1) \begin{pmatrix} f_2 \\ f_1 \end{pmatrix}.$$

At a zero-residual point, the Hessian is the rank-one matrix

$$\begin{pmatrix} f_2 \\ f_1 \end{pmatrix} \begin{pmatrix} f_2 & f_1 \end{pmatrix},$$

which gives (187). Its determinant is zero and its trace is $a^2 + a^{-2}$, yielding the stated eigenvalues.

The zero eigenvalue is tangent to the minimizer manifold $f_1 f_2 = 1$. In the normal direction, the linearized GD multiplier is

$$1 - \frac{\eta_0}{2}(a^2 + a^{-2}).$$

Strict contractivity is therefore equivalent to

$$\left|1 - \frac{\eta_0}{2}(a^2 + a^{-2})\right| < 1,$$

or

$$0 < \eta_0 < \frac{4}{a^2 + a^{-2}}.$$

Since

$$a^2 + a^{-2} = 2\cosh\left(\frac{\Theta_{\mathrm{sh}}}{2}\right),$$

the threshold equals

$$2\ \mathrm{sech}\left(\frac{\Theta_{\mathrm{sh}}}{2}\right) \sim 4e^{-\Theta_{\mathrm{sh}}/2}.$$

□

I.5 Proof of Theorem 7.1

Proof of Theorem 7.1. The three claims follow from independent constructions.

For radial contraction, Proposition I.1 gives, for every $L \geq 2$ and $\rho_- > 0$, an exactly shape-balanced instance satisfying

$$r_- = -\rho_-, \qquad \Theta_{\mathrm{sh}} = 0,$$

for which

$$\frac{\sum_\ell g_\ell^2}{2L\mathcal{L}_1} = \exp\left[-\frac{2(L-1)}{L}\rho_-\right].$$

Thus, the lower-radial exponent in the depth-normalized gradient-transmission bound is attained exactly.

For radial expansion, Proposition I.3 gives, for every $L \geq 2$ and $\rho_+ > 0$, an exactly shape-balanced instance satisfying

$$r_+ = \rho_+, \qquad \Theta_{\mathrm{sh}} = 0,$$

whose exact connected one-step-stable normalized stepsize threshold is

$$\eta_{0,\mathrm{conn}} = 2L\exp\left[-\frac{2(L-1)}{L}\rho_+\right].$$

Hence, any uniform one-step loss-nonincreasing guarantee over this class must deteriorate with the same upper-radial exponential order.

For pure shape, Proposition I.8 gives a depth-two family with

$$P(F) = 1, \qquad r_- = r_+ = 0,$$

and arbitrarily large Θ_{sh} for which

$$\eta_{0,\mathrm{conn}} \sim 4e^{-\Theta_{\mathrm{sh}}/2}.$$

Thus, no fixed positive normalized stepsize can guarantee one-step loss nonincrease uniformly over this radially neutral class as $\Theta_{\mathrm{sh}} \to \infty$.

At depth two, the shape exponent appearing in the sufficient finite-step scale is

$$\frac{2(L-1)}{L^2} = \frac{1}{2}.$$

The inverse-pair construction therefore matches the $e^{-\Theta_{\mathrm{sh}}/2}$ exponential order appearing in the sufficient finite-step control at depth two. This is a matching worst-case exponential order, not a claim that the sufficient theorem is necessary or that every prefactor and auxiliary basin margin in Theorem 5.2 is sharp. These constructions establish the three claims of Theorem 7.1. □

I.6 Quantifier-level interpretation

The obstruction statements are existential worst-case results and should be read at the level of uniform guarantees.

For radial collapse, for every (L, ρ) there exists an exactly balanced invertible instance whose depth-normalized gradient-transmission ratio equals

$$e^{-2(L-1)\rho/L}.$$

Thus, a gradient-transmission lower bound that is uniform over this class cannot have a strictly better lower-radial exponential dependence.

For radial expansion, for every (L, ρ) there exists an exactly balanced instance whose connected one-step-stable normalized stepsize threshold equals

$$2Le^{-2(L-1)\rho/L}.$$

Thus a uniform one-step loss-nonincreasing guarantee cannot admit an exponentially larger normalized stepsize scale.

For pure shape, for every fixed $\eta_0 > 0$ there exists a radially neutral depth-two factorization with sufficiently large Θ_{sh} for which the first GD step increases the loss. Equivalently, the connected one-step stability threshold along this family decays as

$$4e^{-\Theta_{\mathrm{sh}}/2}.$$

None of these statements implies that every factorization outside the sufficient basin is unstable, that every factorization with the same geometric coordinates realizes the worst case, or that failure of one-step loss nonincrease precludes eventual convergence. In particular, the obstructions concern uniform worst-case control rather than an instance-wise ordering of stability severity. This distinction is consistent with the commuting and matched same-predictor controls reported in Appendix K.

J Second-Order Robustness of the Balanced-Identity Threshold

This appendix proves Proposition 7.2. The result concerns the zero-target identity problem

$$\mathcal{L}(F) = \frac{1}{2}\|P(F)\|_F^2, \qquad P(F) = F_L \cdots F_1,$$

at an exact same-predictor factorization satisfying

$$P(F) = I.$$

Throughout this appendix, L is fixed. Constants denoted C_L, c_L, ε_L may depend on L but not on the matrix dimension.

J.1 Exact gradient identity and product cancellation

Let

$$Q_\ell = F_L \cdots F_{\ell+1}, \qquad R_\ell = F_{\ell-1} \cdots F_1.$$

At $P(F) = I$, the zero-target gradient formula gives

$$G_\ell = Q_\ell^\top P(F) R_\ell^\top = Q_\ell^\top R_\ell^\top.$$

Since

$$Q_\ell F_\ell R_\ell = I,$$

we have

$$F_\ell R_\ell = Q_\ell^{-1},$$

and hence,

$$G_\ell F_\ell^\top = Q_\ell^\top R_\ell^\top F_\ell^\top = Q_\ell^\top (F_\ell R_\ell)^\top = I.$$

Therefore,

$$G_\ell = F_\ell^{-\top}. \tag{189}$$

Writing

$$h = \frac{\eta_0}{L},$$

one GD step is exactly

$$F_\ell^+(h) = F_\ell - hF_\ell^{-\top}. \tag{190}$$

Now, write

$$F_\ell = I + E_\ell$$

and assume

$$\max_{1\le\ell\le L} \|E_\ell\|_2 \le \varepsilon < 1.$$

The exact product constraint gives

$$(I + E_L)\cdots(I + E_1) = I.$$

Expanding by degree and isolating the linear part,

$$\sum_{\ell=1}^{L} E_\ell = -\sum_{k=2}^{L} \sum_{i_k>\cdots>i_1} E_{i_k}\cdots E_{i_1}.$$

Consequently,

$$\begin{aligned}\left\|\sum_{\ell=1}^{L} E_\ell\right\|_2 &\le \sum_{k=2}^{L} \binom{L}{k}\varepsilon^k \\ &= (1+\varepsilon)^L - 1 - L\varepsilon \\ &\le C_L\varepsilon^2 \end{aligned} \tag{191}$$

for all sufficiently small ε. The same estimate holds for $\sum_\ell E_\ell^\top$.

Thus, the exact same-predictor constraint removes the entire first-order aggregate perturbation.

J.2 A DIMENSION-FREE C^1 EXPANSION NEAR THE BALANCED BOUNDARY

The Neumann expansion gives

$$(I + E_\ell)^{-\top} = I - E_\ell^\top + R_\ell,$$

where

$$\|R_\ell\|_2 \le \frac{\varepsilon^2}{1-\varepsilon}. \tag{192}$$

Combining this with (190),

$$F_\ell^+(h) = (1-h)I + E_\ell + hE_\ell^\top - hR_\ell. \tag{193}$$

Fix a compact interval J containing 2 but bounded away from 1; for example,

$$J = [3/2, 5/2].$$

Put

$$a = 1 - h, \qquad K_\ell(h) = E_\ell + hE_\ell^\top - hR_\ell.$$

Uniformly for $h \in J$,

$$|a| \ge \frac{1}{2}, \qquad \|K_\ell(h)\|_2 \le C\varepsilon.$$

Moreover,

$$K_\ell'(h) = E_\ell^\top - R_\ell,$$

so

$$\|K_\ell'(h)\|_2 \le C\varepsilon.$$

The first-order aggregate satisfies

$$\sum_{\ell=1}^{L} K_\ell(h) = \sum_{\ell=1}^{L} E_\ell + h\sum_{\ell=1}^{L} E_\ell^\top - h\sum_{\ell=1}^{L} R_\ell.$$

Using (191) and (192),

$$\left\| \sum_{\ell=1}^{L} K_\ell(h) \right\|_2 \le C_L \varepsilon^2 \qquad (h \in J). \tag{194}$$

Differentiating the same expression gives

$$\sum_{\ell=1}^{L} K'_\ell(h) = \sum_{\ell=1}^{L} E_\ell^\top - \sum_{\ell=1}^{L} R_\ell,$$

and therefore,

$$\left\| \sum_{\ell=1}^{L} K'_\ell(h) \right\|_2 \le C_L \varepsilon^2. \tag{195}$$

Expanding the finite product,

$$P(F^+(h)) = a^L I + a^{L-1} \sum_{\ell=1}^{L} K_\ell(h) + \sum_{k=2}^{L} a^{L-k} \sum_{i_k > \cdots > i_1} K_{i_k}(h) \cdots K_{i_1}(h).$$

Every term in the final sum contains at least two K factors. Hence, (194) implies, uniformly on J,

$$\left\| P(F^+(h)) - (1-h)^L I \right\|_2 \le C_L \varepsilon^2. \tag{196}$$

Differentiating the finite-product expansion, every differentiated higher-order term still contains either at least two $O(\varepsilon)$ factors or one $O(\varepsilon)$ factor together with one $O(\varepsilon)$ derivative. Together with (195), this gives

$$\left\| \frac{d}{dh} P(F^+(h)) + L(1-h)^{L-1} I \right\|_2 \le C_L \varepsilon^2 \tag{197}$$

uniformly on J. Define the normalized one-step loss ratio

$$\Psi_F(h) := \frac{\mathcal{L}(F^+(h))}{\mathcal{L}(F)} = \frac{1}{d} \|P(F^+(h))\|_F^2,$$

where

$$\mathcal{L}(F) = \frac{d}{2}$$

because $P(F) = I$.

Write

$$P(F^+(h)) = (1-h)^L I + R(h).$$

Then,

$$\Psi_F(h) = |1-h|^{2L} + \frac{2(1-h)^L}{d} \operatorname{tr} R(h) + \frac{1}{d} \|R(h)\|_F^2.$$

Using

$$\frac{|\operatorname{tr} R|}{d} \le \|R\|_2, \qquad \frac{\|R\|_F^2}{d} \le \|R\|_2^2,$$

we obtain

$$\left| \frac{2(1-h)^L}{d} \operatorname{tr} R(h) + \frac{1}{d} \|R(h)\|_F^2 \right| \le 2|1-h|^L \|R(h)\|_2 + \|R(h)\|_2^2.$$

The same argument after differentiation, together with (197), yields the dimension-free estimates

$$\Psi_F(h) = |1-h|^{2L} + O_L(\varepsilon^2), \tag{198}$$

$$\Psi'_F(h) = \frac{d}{dh} |1-h|^{2L} + O_L(\varepsilon^2), \tag{199}$$

uniformly on J.

J.3 IDENTIFICATION OF THE CONNECTED THRESHOLD

At the balanced identity,

$$\Psi_I(h) = |1-h|^{2L},$$

so the connected one-step loss-nonincreasing interval containing zero is exactly $[0, 2]$. At the right endpoint,

$$\Psi_I(2) = 1, \qquad \Psi_I'(2) = 2L > 0.$$

By the uniform C^1 estimates (198)– (199), there exist a fixed neighborhood U of 2 and $\varepsilon_L > 0$ such that, whenever $\varepsilon \le \varepsilon_L$, the function

$$h \longmapsto \Psi_F(h) - 1$$

has a unique transverse zero

$$h_\star \in U.$$

Moreover, since the unperturbed derivative at the crossing equals $2L$, the root displacement is second order:

$$|h_\star - 2| \le C_L \varepsilon^2. \tag{200}$$

It remains to verify that $h_\star$ is the first boundary of the one-step-stable component issuing from zero.

At $h = 0$,

$$\Psi_F(0) = 1.$$

Since

$$F_\ell^+(h) = F_\ell - hG_\ell,$$

differentiating the loss along this direction gives

$$\frac{d}{dh}\mathcal{L}(F^+(h))\Big|_{h=0} = -\sum_{\ell=1}^{L} \|G_\ell\|_F^2.$$

Using (189) and $\mathcal{L}(F) = d/2$,

$$\Psi_F'(0) = -\frac{2}{d}\sum_{\ell=1}^{L} \|F_\ell^{-\top}\|_F^2. \tag{201}$$

Because

$$\|F_\ell\|_2 \le 1 + \varepsilon,$$

every singular value of $F_\ell^{-\top}$ is at least $(1+\varepsilon)^{-1}$, and therefore

$$\|F_\ell^{-\top}\|_F^2 \ge \frac{d}{(1+\varepsilon)^2}.$$

Consequently,

$$\Psi_F'(0) \le -\frac{2L}{(1+\varepsilon)^2} < 0. \tag{202}$$

Furthermore, for fixed L and sufficiently small ε, the finite-product representation of $P(F^+(h))$ and its first two derivatives gives a dimension-free bound

$$|\Psi_F''(h)| \le C_L \tag{203}$$

on a fixed neighborhood $[0, h_0]$ of the origin. Combining (202) and (203), after decreasing h_0 and ε_L if necessary, yields

$$\Psi_F(h) < 1 \qquad \text{for every } 0 < h \le h_0. \tag{204}$$

Next, choose a fixed $\delta > 0$ small enough that

$$h_0 < \delta < 2 - \delta$$

and that the neighborhood of the crossing near 2 overlaps the right end of the remaining interval. On every compact subinterval strictly inside $(0, 2)$,

$$|1-h|^{2L} < 1.$$

In particular,

$$\gamma_{L,\delta} := 1 - \max_{h\in[\delta,\,2-\delta]} |1-h|^{2L} > 0.$$

The finite-product map $(F,h) \mapsto \Psi_F(h)$ is continuous uniformly over this compact interval for fixed L and sufficiently small ε. Hence, after decreasing ε_L if necessary,

$$\Psi_F(h) < 1 \qquad \text{for every } h \in [\delta, 2-\delta]. \tag{205}$$

Finally, the C^1 expansion near $h = 2$ gives a unique transverse crossing $h_\star$ satisfying (200), with

$$\Psi_F(h) < 1$$

immediately to its left and

$$\Psi_F(h) > 1$$

immediately to its right. Choosing the fixed neighborhoods above with overlap, and then choosing ε_L sufficiently small, shows that there is no earlier crossing between 0 and $h_\star$. Therefore $h_\star$ is exactly the right endpoint of the connected one-step-stable interval containing zero.

Since

$$\eta_0 = Lh$$

and L is fixed,

$$\eta_{0,\mathrm{conn}}(F) = Lh_\star.$$

Thus,

$$|\eta_{0,\mathrm{conn}}(F) - 2L| = L|h_\star - 2| \le C_L\varepsilon^2,$$

after absorbing the fixed factor L into C_L. This proves Proposition 7.2.

K Experimental Details

This appendix describes the constructions, numerical protocols, and diagnostics underlying Section 8. Unless stated otherwise, all computations use double precision.

Throughout the matrix experiments we set

$$A = B = X = I.$$

The radial-contraction experiment uses the scalar target $Y = 1$ to realize the corresponding obstruction construction. All other experiments use $Y = 0$. Thus, except for radial contraction,

$$\mathcal{L}(F) = \frac{1}{2}\|P(F)\|_F^2, \qquad P(F) = F_L \cdots F_1.$$

Gradient descent uses the depth-normalized update

$$F_\ell^+ = F_\ell - \frac{\eta_0}{L}\nabla_{F_\ell}\mathcal{L}(F). \tag{206}$$

All reported stability thresholds therefore refer to the normalized stepsize η_0, rather than to the raw layerwise stepsize η_0/L.

Whenever a one-step stability threshold is reported, it refers to the endpoint of the connected loss-nonincreasing component containing $\eta_0 = 0$. We denote its numerical estimate by $\widehat{\eta}^{(1)}_{0,\mathrm{conn}}$. Analogously, $\widehat{\eta}^{(20)}_{0,\mathrm{conn}}$ denotes the connected threshold for the 20-step criterion defined below.

K.1 Scalar obstruction laws

The first three experiments correspond to the scalar constructions in Section 7. Because closed-form expressions are available, these experiments primarily verify the threshold-search implementation and illustrate the predicted scaling laws.

Radial contraction. For

$$L \in \{2, 4, 8, 16, 32\},$$

we take

$$f_1 = \cdots = f_L = s, \qquad s = e^{-\rho/L},$$

with $\rho \in [0.25, 8]$ evaluated on a fixed grid. With scalar target $Y = 1$,

$$\mathcal{L} = \frac{1}{2}(s^L - 1)^2.$$

We evaluate the depth-normalized gradient-transmission coefficient

$$\widehat{\mu}_{\mathrm{tr}} := \frac{\sum_{\ell=1}^{L} g_\ell^2}{2L\mathcal{L}}. \tag{207}$$

Direct substitution gives the exact law

$$\widehat{\mu}_{\mathrm{tr}} = e^{-2(L-1)\rho/L}. \tag{208}$$

Across the complete grid, the maximum relative error with respect to (208) is 4.53×10^{-15}. The maximum absolute error of the fitted logarithmic slope is 8.88×10^{-16}, and the minimum coefficient of determination is $R^2 = 1$.

Radial expansion. We next take

$$f_1 = \cdots = f_L = s, \qquad s = e^{\rho/L},$$

with zero target. One normalized GD step gives the exact loss ratio

$$\frac{\mathcal{L}(F^+)}{\mathcal{L}(F)} = \left|1 - \frac{\eta_0}{L} s^{2L-2}\right|^{2L}. \tag{209}$$

The connected loss-nonincreasing interval containing $\eta_0 = 0$ has endpoint

$$\eta_{0,\mathrm{conn}} = 2Le^{-2(L-1)\rho/L}. \tag{210}$$

Numerical bisection recovers this threshold with maximum relative error 6.00×10^{-15}. The maximum absolute error in the fitted logarithmic slope is 6.66×10^{-16}, and the minimum R^2 is 1.

Pure shape imbalance. For the depth-two inverse pair, the one-step loss ratio is

$$\frac{\mathcal{L}(F^+)}{\mathcal{L}(F)} = \left(1 - \eta_0 \cosh(\Theta_{\mathrm{sh}}/2) + \frac{\eta_0^2}{4}\right)^2. \tag{211}$$

Writing

$$C_\Theta = \cosh(\Theta_{\mathrm{sh}}/2),$$

when

$$C_\Theta \geq \sqrt{2},$$

the endpoint of the connected loss-nonincreasing component containing $\eta_0 = 0$ is

$$\eta_{0,\mathrm{conn}} = 2C_\Theta - 2\sqrt{C_\Theta^2 - 2}. \tag{212}$$

In particular,

$$\eta_{0,\mathrm{conn}} \sim 4e^{-\Theta_{\mathrm{sh}}/2} \qquad \text{as } \Theta_{\mathrm{sh}} \to \infty. \tag{213}$$

For $\Theta_{\mathrm{sh}} \in [2, 12]$, the condition $C_\Theta \geq \sqrt{2}$ is satisfied. The maximum relative error between the numerical threshold and the exact finite-Θ_{sh} expression is 1.78×10^{-12}. Fitting the logarithmic threshold over $\Theta_{\mathrm{sh}} \geq 6$ gives slope -0.500274 with $R^2 = 0.99999985$.

K.2 Commuting factorizations of a fixed predictor

We next examine whether shape imbalance alone necessarily realizes the worst-case stability obstruction.

Let Q be fixed orthogonal and let $v \in \{-1, +1\}^d$ contain equal numbers of positive and negative entries. For coefficients $c_1, \dots, c_L$ satisfying

$$\sum_{\ell=1}^{L} c_\ell = 0, \qquad \sum_{\ell=1}^{L} |c_\ell| = 1,$$

define

$$F_\ell = Q \operatorname{diag}(e^{a c_\ell v})\, Q^\top. \tag{214}$$

The factors commute, and therefore

$$P(F) = F_L \cdots F_1 = I.$$

Moreover,

$$r_+ = a, \qquad r_- = -a,$$

for every coefficient pattern. We use smooth, block, alternating, and spike patterns to vary the distribution of inter-layer shape while keeping these radial coordinates fixed.

The experimental grid uses

$$d = 16, \qquad L \in \{8, 16\}, \qquad a \in \{0.25, 0.5, 0.75, 1, 1.25, 1.5\}.$$

The maximum numerical product error satisfies

$$\|P(F) - I\|_F < 4.5 \times 10^{-14}.$$

Across this commuting family, the measured connected one-step stability thresholds vary comparatively modestly. This provides a useful control: large aggregate shape does not force every factorization to realize the worst-case pure-shape obstruction.

K.3 Noncommuting factorizations of the same predictor

To separate end-to-end prediction from internal noncommutative geometry, set

$$H_0 = H_L = I$$

and define

$$F_\ell = H_\ell H_{\ell-1}^{-1}. \tag{215}$$

Then,

$$P(F) = F_L \cdots F_1 = I$$

by telescoping.

The intermediate matrices H_ℓ are positive definite with controlled spectra and rotating eigenbases. We use smooth, alternating, and independently generated rotation families with

$$d = 8, \qquad L \in \{8, 16\},$$

spectral amplitudes

$$a \in \{0.5, 1, 1.5, 2\},$$

rotation amplitudes

$$b \in \{0, 0.5, 1, 1.5\},$$

and three fixed random seeds, yielding 288 configurations in total. The maximum numerical product error is below

$$1.3 \times 10^{-13}.$$

Despite representing essentially the same end-to-end predictor, these factorizations exhibit a wide range of measured connected stable normalized stepsizes:

$$4.86 \times 10^{-3} \lesssim \widehat{\eta}^{(1)}_{0,\mathrm{conn}} \lesssim 32.79. \tag{216}$$

The observed associations between stability and the geometric variables are descriptive rather than causal, because the coordinates are not randomized independently. The reported correlations are

$$\operatorname{corr}\left(\log \widehat{\eta}^{(1)}_{0,\mathrm{conn}}, r_+\right) = -0.652,$$

$$\operatorname{corr}\left(\log \widehat{\eta}^{(1)}_{0,\mathrm{conn}}, \Theta_{\mathrm{sh}}\right) = -0.480,$$

and the Pearson correlation between $\log \widehat{\eta}^{(1)}_{0,\mathrm{conn}}$ and the maximum adjacent commutator norm is -0.526.

K.4 Fixed-stepsize diagnostic

We generate matrix-valued initializations from

$$Z_\ell = \frac{\rho}{L} I + \frac{\alpha_{\exp}}{\sqrt{L}} G_\ell, \qquad F_\ell = e^{Z_\ell}, \tag{217}$$

where each G_ℓ is a fixed random symmetric traceless matrix normalized to operator norm one. The grid is

$$L \in \{8, 16\},$$
$$\rho \in \{-0.2, -0.1, 0, 0.1, 0.2, 0.3\},$$
$$\alpha_{\exp} \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\},$$

with ten fixed seeds, yielding 720 initializations.

As a prespecified diagnostic, we run 100 GD iterations with $\eta_0 = 1$. All 720 runs remain stable under this fixed stepsize. We retain this negative result rather than selecting a larger stepsize after observing the outcome. The experiment therefore motivates measuring the stability threshold itself on exactly the same initialization set.

K.5 Generic matrix stability thresholds

We reuse all 720 initializations from Section K.4, without filtering or resampling.

For a given initialization, define the numerical one-step threshold as

$$\widehat{\eta}^{(1)}_{0,\mathrm{conn}} := \sup\left\{b > 0 : \mathcal{L}(F^+_{\eta_0}) \le \mathcal{L}(F) \text{ for every } \eta_0 \in [0, b]\right\}. \tag{218}$$

The numerical search begins at 10^{-8}, doubles the candidate stepsize until loss nonincrease fails, and uses the hard cap 2^{16}. Once the threshold is bracketed, we perform 90 bisection iterations. Numerical loss nonincrease is declared when

$$\frac{\mathcal{L}(F^+)}{\mathcal{L}(F)} \le 1 + 10^{-12}. \tag{219}$$

All 720 thresholds are successfully bracketed. Their observed range is

$$3.714 \lesssim \widehat{\eta}^{(1)}_{0,\mathrm{conn}} \lesssim 46.560. \tag{220}$$

To summarize the geometry appearing in the finite-step bounds, we use the theory-motivated score

$$\Gamma(F) = \frac{2(L-1)}{L} \max\{r_+(F), 0\} + \frac{2(L-1)}{L^2} \Theta_{\mathrm{sh}}(F). \tag{221}$$

The coefficients in Γ are fixed by the theoretical exponents; they are not fitted to the experimental thresholds. The score is used only as a compact observational summary of geometric stability ordering. It is neither the full sufficient condition of Theorem 5.2 nor claimed to be an optimal empirical predictor.

Within depth, the Spearman correlation between Γ and $\log \widehat{\eta}^{(1)}_{0,\mathrm{conn}}$ is

$$-0.765 \qquad (L = 8),$$

and

$$-0.742 \qquad (L = 16).$$

For comparison, the corresponding Spearman correlations with $(r_+, \Theta_{\mathrm{sh}})$ are

$$(-0.868, -0.643) \qquad (L = 8)$$

and

$$(-0.809, -0.618) \qquad (L = 16).$$

In particular, r_+ alone is more strongly rank-correlated with the one-step threshold than Γ on this grid. We therefore interpret Γ as a theory-motivated geometric summary rather than as evidence that combining the two coordinates improves empirical prediction.

We also compute an independently defined 20-step stability threshold. For each initialization, let $\widehat{\eta}^{(20)}_{0,\mathrm{conn}}$ denote the endpoint of the connected interval containing zero on which

$$\max_{0 \le t \le 20} \mathcal{L}_t \le 1.05\, \mathcal{L}_0, \qquad \mathcal{L}_{20} < \mathcal{L}_0. \tag{222}$$

The observed range is

$$3.587 \le \widehat{\eta}^{(20)}_{0,\mathrm{conn}} \le 46.560.$$

The one-step and 20-step thresholds induce nearly identical rankings on this initialization set:

$$\mathrm{Spearman}\left(\widehat{\eta}^{(1)}_{0,\mathrm{conn}}, \widehat{\eta}^{(20)}_{0,\mathrm{conn}}\right) = 0.99924. \tag{223}$$

This empirical agreement does not make the one-step threshold a global convergence certificate.

K.6 Computation of the spectral geometry

For each factor, let

$$F_\ell = U_\ell H_\ell, \qquad H_\ell = (F_\ell^\top F_\ell)^{1/2},$$

be its ordinary right polar decomposition. Starting from $C_0 = I$, define recursively

$$C_\ell = U_\ell C_{\ell-1}, \qquad S_\ell = C_{\ell-1}^\top H_\ell C_{\ell-1}, \qquad Z_\ell = \log S_\ell.$$

By Appendix A, this is exactly the recursive polar gauge used in the theory. We therefore evaluate

$$\Theta_{\mathrm{sh}}(F) = L \sum_{\ell=1}^{L-1} \|Z_{\ell+1} - Z_\ell\|_2 \tag{224}$$

directly, rather than through a surrogate statistic. The radial coordinates are

$$r_-(F) = \sum_{\ell=1}^{L} \log \sigma_{\min}(F_\ell), \qquad r_+(F) = \sum_{\ell=1}^{L} \log \|F_\ell\|_2.$$

All positive-definite square roots and matrix logarithms are evaluated using symmetric eigendecomposition.

K.7 Reproducibility and interpretation

The three scalar experiments are deterministic. The noncommuting same-predictor experiment uses the complete predeclared grid of spectral amplitudes, rotation amplitudes, rotation families, and random seeds. The generic matrix experiment uses all 720 initializations from the fixed-step diagnostic without filtering or resampling.

All reported one-step stability thresholds refer to the connected loss-nonincreasing component containing zero. The commuting family and the fixed-stepsize diagnostic are retained as controls rather than retuned after their outcomes were observed. The sufficient spectral-balance basin is not interpreted as a necessary empirical boundary. Source code, configuration files, exact random seeds, and reproduction scripts are available in the anonymous repository: `https://anonymous.4open.science/r/spectral-balance-gd-ECC3/`.

K.8 Matched shape intervention at fixed predictor and radial geometry

The matched negative control was specified before the full run. We use

$$H_0 = H_L = I, \qquad F_\ell = H_\ell H_{\ell-1}^{-1},$$

so that

$$P(F) = I$$

exactly in exact arithmetic. The intermediate positive-definite matrices have rotated eigenspaces, and a two-parameter spectral calibration matches prescribed targets for (r_-, r_+). The grid uses

$$d \in \{8, 16\}, \qquad L \in \{8, 16\},$$

three radial regimes, seven intervention levels, and ten seeds, for 840 configurations. A configuration is accepted only when the frozen product and radial matching tolerances are met; no threshold, seed, or tolerance is retuned after observing the results.

The frozen calibration accepts $732/840$ configurations, an overall rate of 87.14%, below the prespecified 90% operational gate. Acceptance rates are 95.2%, 96.2%, 76.7%, and 80.5% for $(d, L) = (8, 8), (8, 16), (16, 8), (16, 16)$, respectively. Among accepted configurations, the maximum normalized product error is 1.78×10^{-15} and the maximum radial mismatch is 2.13×10^{-11}. We retain this calibration shortfall as a reported limitation rather than changing the gate post hoc.

There are 111 matched families satisfying the frozen shape-span criterion. Only 18.9% have a negative fitted within-family slope of $\log \widehat{\eta}^{(1)}_{0,\mathrm{conn}}$ against Θ_{sh}; the mean slope is

$$7.24 \times 10^{-4},$$

with bootstrap 95% interval

$$[4.74, 9.79] \times 10^{-4}.$$

We do not interpret the positive sign as a universal beneficial effect of shape. The more stable finding is the small threshold variation within the accepted matched families: the median within-family max/min threshold ratio is approximately 1.007, the 90th percentile is approximately 1.071, the 95th percentile is approximately 1.088, and the maximum is approximately 1.132. Thus, within this matched construction, the prespecified variation in shape does not produce a monotone worst-case-like decrease of the connected one-step threshold.

A mechanism diagnostic over all 732 accepted configurations gives

$$\mathrm{median}\, \frac{\widehat{\eta}^{(1)}_{0,\mathrm{conn}}}{2L} \approx 1.001.$$

Within a typical matched family, the median max/min ratios are approximately 1.0011 for total gradient energy, 1.0021 for the second directional derivative of the loss, and 1.0010 for the local quadratic threshold estimate. In contrast, the corresponding median ratios are approximately 1.80 for the maximum excluded-product transmission envelope and 2.85 for the squared end-to-end Jacobian operator norm. Thus the uniform operator-norm envelopes can vary substantially while quantities associated with the realized GD direction remain nearly unchanged.

The local quadratic estimate should not, however, be interpreted as an accurate predictor of the absolute connected threshold: across accepted configurations,

$$\mathrm{median}\, \frac{\widehat{\eta}^{(1)}_{0,\mathrm{conn}}}{\widehat{\eta}_{0,\mathrm{quad}}} \approx 29.86,$$

with an observed range of approximately 13.95 to 31.52. The diagnostic is therefore informative about within-family variation, not about the absolute threshold scale.

This behavior is consistent with the local mechanism captured by Proposition 7.2: a worst-case geometric envelope may vary substantially even when a particular near-balanced same-predictor realization retains a threshold close to the balanced identity value.

For a longer-horizon diagnostic, 101 representative accepted configurations are run for up to 500 steps at 0.25, 0.50, 0.80, and 1.05 times their measured initial one-step connected threshold. At 0.25 and

0.50 times the threshold, all 101 runs reach step 500; the median final loss ratios are approximately 1.84×10^{-9} and 4.07×10^{-25}, respectively. At 0.80 times the threshold, $96/101$ runs reach step 500, with median final loss ratio approximately 5.74×10^{-7} among the completed runs, while five runs violate the prespecified continuation criterion before step 10. At 1.05 times the threshold, all configurations have a recorded first step, but none continues to step 10 under the prespecified continuation criterion.

Thus the initial one-step threshold is informative on this diagnostic but is not itself a global trajectory certificate. In particular, being below the initial one-step threshold does not guarantee that later iterates remain inside a stable geometric region. This distinction is consistent with the role of forward basin invariance in the main theorem.

K.9 Second-order balanced-identity scaling check

We independently test Proposition 7.2 using exact telescoping factorizations satisfying

$$P(F) = I.$$

The frozen grid uses

$$d \in \{4, 8\}, \qquad L \in \{2, 4, 8, 16\},$$

nine perturbation amplitudes and eight seeds, for 576 configurations. For each configuration we measure

$$\varepsilon = \max_{\ell} \|F_\ell - I\|_2$$

and the numerical endpoint $\widehat{\eta}^{(1)}_{0,\mathrm{conn}}$ of the connected one-step loss-nonincreasing interval. All 576 product and connectedness checks pass; the maximum product error is 1.13×10^{-15}.

For each (d, L) pair, a log–log fit of

$$\left|\widehat{\eta}^{(1)}_{0,\mathrm{conn}} - 2L\right|$$

against ε gives fitted exponents between 1.929 and 1.984, with median 1.941 and R^2 between 0.9930 and 0.9974. The normalized connected threshold satisfies

$$\operatorname{median} \frac{\widehat{\eta}^{(1)}_{0,\mathrm{conn}}}{2L} \approx 1.00143,$$

with observed range approximately

$$[1.00003, 1.04959].$$

As a prespecified asymptotic diagnostic, we repeat the fit using the five smallest perturbation levels per seed. The median fitted exponent becomes 1.979, with the fitted exponents approximately between 1.953 and 1.985. This is consistent with the second-order law predicted by Proposition 7.2. We report both the full-range and small-ε fits to expose finite-ε higher-order effects rather than selecting only the more favorable asymptotic range. These numerical fits corroborate the theorem but are not used as evidence for its proof.

L Comparison with Closely Related Prior Work

Table 1 summarizes the closely related results discussed in Section 2 along the dimensions most relevant to our analysis. The comparison distinguishes the model and initialization regimes, optimization guarantees, finite-step assumptions, and geometric quantities controlled by each result. It is intended to clarify the scope and technical relationship of the results rather than to rank the strength of the respective guarantees.

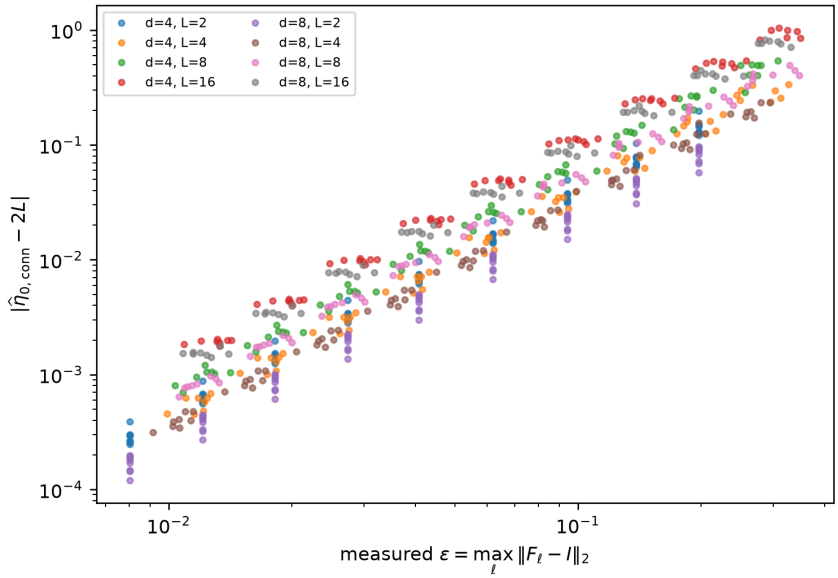

Figure 2: **Perturbative check of the balanced-identity threshold.** The deviation $|\widehat{\eta}^{(1)}_{0,\mathrm{conn}} - 2L|$ approaches the predicted second-order scaling as the factors approach identity.

Table 1: Theorem-level comparison with representative closely related work. "Finite-step" denotes positive-step gradient descent rather than a gradient-flow-only result. Entries summarize the aspects most relevant to our comparison and are not intended as exhaustive descriptions of the cited papers.

Work	Finite-step	Initialization / geometry	Relevant result	Distinction in this work
Du–Hu–Lee (Du et al., 2018)	Partial	Layer-norm balancing in homogeneous models; asymmetric matrix factorization	GF invariance/automatic balancing; positive-step GD results for matrix factorization	Foundational balancing mechanism; our analysis instead uses recursive-polar radial/shape geometry and an explicit invariant spectral basin.
Gidel–Bach–Lacoste-Julien (Gidel et al., 2019)	✓	Vanishing initialization; two-layer linear network	Discrete dynamics and sequential reduced-rank learning	Closely related discrete-dynamics setting, but different optimization question and no nested spectral-balance basin.
Arora et al. (Arora et al., 2019a)	✓	Approximately balanced deep-linear initialization with a rank-margin condition	Linear-rate GD convergence to a global optimum under their assumptions	Classical balanced-GD predecessor; our sufficient region separates lower radial scale, upper radial scale, and logarithmic inter-layer shape.
Hu et al. (Hu et al., 2020)	✓	Orthogonal and Gaussian initialization in deep linear networks	Provable convergence advantage of orthogonal initialization, including improved width–depth dependence	Initialization-specific convergence analysis rather than a deterministic invariant basin over general nonzero factorizations.
Zou et al. (Zou et al., 2020)	✓	Zero initialization of hidden residual weights	Global GD/SGD convergence for deep linear ResNets under their width and transformation assumptions	Residual-network predecessor centered at zero residual initialization; our basin allows a broader class of nonzero invertible effective factorizations.
Chen–Lin–Zhang (Chen et al., 2026)	–	Arbitrary balanced starts; deep linear networks with at least one one-neuron layer	Nonlocal GF convergence with explicit stage-dependent rates	Strong recent nonlocal result, but continuous time and under a specific architectural condition.
Min–Vidal–Mallada (Min et al., 2023)	–	Imbalance matrices and least singular value of the end-to-end product	Exponential GF convergence under gradient dominance and suitable initialization	Closest continuous-time geometric comparison; our result controls positive-step dynamics through separate radial and logarithmic-shape coordinates.
Nguegnang–Rauhut–Terstiege (Nguegnang et al., 2024)	✓	Approximate balancedness and stepsize conditions	Positive-step GD convergence to critical points, with stronger global-minimum conclusions in specific depth/rank regimes	Closest discrete balancedness comparison; our proof converts additive imbalance drift into logarithmic shape control inside a nested spectral basin.
Chen et al. (Chen et al., 2024)	✓	Large initialization in matrix factorization	GD trajectory analysis and incremental-learning phenomena	Relevant to large-scale finite-step behavior, but not to deep residual factorization geometry or a three-coordinate invariant basin.
Dominé et al. (Dominé et al., 2025)	–	λ-balanced deep-linear initialization	Exact GF dynamics across lazy and rich regimes	Exact initialization-dependent continuous-time dynamics; our focus is positive-step basin invariance and stability.
Marcotte–Peyré–Gribonval (Marcotte et al., 2026)	–	Balanced and relaxed-balanced structures	Conditions for intrinsic lower-dimensional GF dynamics	Broadens the class of balanced structures supporting intrinsic flow dynamics; our result concerns finite-step invariant geometry and stability.
Ghosh et al. (Ghosh et al., 2025)	✓	Deep matrix factorization near and beyond the edge of stability	Large-step dynamics, oscillations, and breakdown of gradient-flow conservation laws	Studies edge/beyond-edge behavior rather than a sufficient globally invariant finite-step basin.
This work	✓	$(r_-, r_+, \Theta_{\mathrm{sh}})$ **spectral-balance geometry**	**Geometry-sensitive PL bound, finite-step basin invariance, and global linear convergence from the basin**	**Also provides coordinate-isolating radial and pure-shape one-step obstruction families.**