A PRISM HIERARCHY OF LEARNING REGIMES IN LARGE LINEAR AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Theoretical studies of machine learning models commonly consider different limiting regimes in which the learning dynamics of gradient descent becomes theoretically tractable. It is, however, desirable to have a systematically obtained picture of qualitatively different extreme learning regimes for a particular type of models. In this paper we propose such a picture for large weight-tied linear autoencoders characterized by input and latent dimensions, initialization magnitude, and training set size. This model is nonlinear in the weights and its gradient flow does not have a general theoretical solution. We show that at the level of the formal loss-expansion hierarchy, its extreme regimes are naturally associated with faces of a triangular prism. In particular, there are five basic extreme regimes associated with the 2-faces of the prism: (1) large-data, (2) small-data, (3) mean-field, (4) narrow-latent, and (5) free. For regimes (1,2,3,4), we derive and rigorously characterize the limiting train and population dynamics under gradient flow, rigorously prove convergence to these limits, and get good agreement with experimental results. The resulting limits reveal spectral critical slowing, initialization-driven trapping away from the PCA optimum, and regimes in which population loss converges exponentially while training loss relaxes only algebraically.

1 INTRODUCTION

A common route to tractable learning dynamics is to take an asymptotic limit in which model width, data size, or initialization scale in a particular way. Different limits can simplify the same training problem differently, including lazy-training linearizations (Chizat et al., 2019; Jacot et al., 2018; Lee et al., 2018), mean-field limits (Sirignano and Spiliopoulos, 2020; Rotskoff and Vanden-Eijnden, 2018), and solvable linear-network dynamics (Saxe et al., 2013; 2019; Tu et al., 2024). Such analyses usually select a regime first and then study its dynamics. For a model with several scalable parameters, this leaves a broader question: *What are its distinct extreme learning regimes, how are they related, and what dynamics arise in each?* Answering this directly is difficult because even a few hyperparameters admit infinitely many joint large-system scalings.

We study this question for the weight-tied linear autoencoder (Baldi and Hornik, 1989; Baldi, 2012). Although linear in its input, the model is nonlinear in the weights and its gradient flow is not generally explicitly solvable. Moreover, with a finite training set, train and population performance can evolve differently. We let the input dimension p, latent width n, training-set size m, and initialization variance σ^2 scale jointly, and analyze both train and population learning trajectories.

We identify five basic extreme regimes – *large-data*, *small-data*, *mean-field*, *narrow-latent*, and *free* – together with intermediate phases combining them. In four of the five basic regimes we derive limiting train and population dynamics and rigorously prove convergence on compact time intervals. The resulting effective systems are qualitatively different: they range from spectral and decoupled covariance dynamics to a nonlinear alignment system and an infinite hierarchy of leakage moments. They also reveal spectral critical slowing, initialization-dependent trapping away from the PCA optimum, and parametrically different relaxation rates of train and population losses.

To organize the possible scalings systematically, we build on the coefficient-expansion viewpoint of Yarotsky et al. (2026), in which regimes are distinguished by dominant terms in formal time expansions. We make the associated phase structure explicit through leading-support signatures across expansion orders and extend the classification jointly to train and population losses. For

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.

the autoencoder, the resulting leading supports form triangular prisms, whose normal fan gives a dual triangular-bipyramid phase diagram. Thus the formal classification identifies the basic extreme regimes before their trajectories are solved, while the limiting analyses show that these regimes correspond to genuinely different learning dynamics.

Our contributions.

1. We derive and rigorously justify limiting train and population dynamics in four basic extreme regimes: *large-data*, *small-data*, *mean-field*, and *narrow-latent* (Section 4). The limits range from explicit spectral flows to a nonlinear alignment system and an infinite moment hierarchy. Numerical experiments on CIFAR10 agree well with the corresponding finite-dimensional predictions (Figs. 2 and 3).
2. The limiting solutions expose several distinct mechanisms: spectral critical slowing when relevant dimensions become comparable, initialization-dependent alignment or trapping away from the PCA optimum, and strong separation between train and population relaxation rates. In particular, in narrow-latent alignment the population residual can decay exponentially while the training residual decays only algebraically (Sections 4 and 5).
3. Building on the dominant-term expansion viewpoint of Yarotsky et al. (2026), we formalize the phase structure through leading-support signatures and extend the classification jointly to train and population losses. For the weight-tied linear autoencoder, this yields a triangular-prism hierarchy, equivalently a dual triangular-bipyramid phase diagram with five basic extreme regimes and additional phases combining them (Section 3 and Fig. 1).

2 Problem statement

We consider a shallow linear weight-tied autoencoder with *input dimension* p and *latent dimension* n: $f_{\boldsymbol{U}}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x} \in \mathbb{R}^p$ is the input, $\boldsymbol{A} = \boldsymbol{U}^\top \boldsymbol{U}$ is the model linear map, and $\boldsymbol{U} \in \mathbb{R}^{n\times p}$ is the weight matrix we train. Consider a training set $\boldsymbol{X} \in \mathbb{R}^{p\times m}$ of size m and define the *train* square loss for this model:

$$\widehat{L}(\boldsymbol{U}) = \frac{\left\|\boldsymbol{X} - \boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{X}\right\|_F^2}{2pm} = \frac{\operatorname{Tr}\left[(\boldsymbol{I}-\boldsymbol{A})\boldsymbol{S}(\boldsymbol{I}-\boldsymbol{A})\right]}{2p}, \qquad \boldsymbol{S} = \frac{1}{m}\boldsymbol{X}\boldsymbol{X}^\top \in \mathbb{R}^{p\times p}. \tag{1}$$

In order to define the *population* loss, we assume data isotropy:

$$\boldsymbol{\Sigma} := \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{I}; \qquad L(\boldsymbol{U}) = \mathbb{E}_{\boldsymbol{x}}\left[\frac{\left\|\boldsymbol{x} - \boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{x}\right\|^2}{2p}\right] = \frac{\|\boldsymbol{I}-\boldsymbol{A}\|_F^2}{2p}. \tag{2}$$

We train our model with gradient flow with learning rate η:

$$\frac{d\boldsymbol{U}}{dt} = -\eta\frac{\partial \widehat{L}(\boldsymbol{U})}{\partial \boldsymbol{U}}, \qquad \boldsymbol{U}(0) \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_n \otimes \boldsymbol{I}_p). \tag{3}$$

Both η and σ^2 may depend on p, m, n. Define the loss values at time t: $\widehat{L}(t) = \widehat{L}(\boldsymbol{U}(t))$, $L(t) = L(\boldsymbol{U}(t))$. We will be interested in their evolution in the limit of large p, n, m. As we shall see in Section 4, this evolution depends qualitatively on the *joint* scaling of p, n, m, and σ^2, suggesting existence of several *learning regimes* in the high-dimensional limit. In Section 3, we identify five basic learning regimes, each associated with a face of a triangular prism. In Section 4, we derive explicit finite-time limiting dynamics in four of these regimes and rigorously prove convergence to these limits. The corresponding solutions have very different nature, therefore deserve to be regarded as genuine learning regimes.

3 A formal phase diagram of learning regimes (B)

General framework. There is no established, canonical way to define a phase diagram for the learning dynamics of a model with several jointly scaled hyperparameters. One can identify particular solvable limits, but this does not by itself specify which limits should count as different phases or whether other regimes have been missed. We propose an operational definition based on the Taylor

coefficients of averaged learning observables at initialization. It builds on the diagrammatic expansion method of Yarotsky et al. (2026), but makes the equivalence relation defining phases explicit and allows several observables, such as train and population losses, to be classified jointly.

Here is the general setting. Fix averaged observables $F_\ell(t;\boldsymbol{\theta})$ and positive hyperparameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)$. Their Taylor coefficients at initialization are the derivatives $D_{\ell,s}(\boldsymbol{\theta}) = \frac{d^s}{dt^s} F_\ell(t;\boldsymbol{\theta})\big|_{t=0}$. We say that these observables are in the *Laurent-polynomial coefficient setting* if, at every positive order $s \geq 1$,

$$D_{\ell,s}(\boldsymbol{\theta}) = \sum_{\boldsymbol{q} \in Q_{\ell,s}} c_{\ell,s,\boldsymbol{q}} \boldsymbol{\theta}^{\boldsymbol{q}}, \qquad Q_{\ell,s} \subset \mathbb{Z}^d \text{ finite}, \quad c_{\ell,s,\boldsymbol{q}} \neq 0. \tag{4}$$

Negative integer powers are permitted, as in factors $p^{-s}m^{-s}$ arising from loss normalization. We omit order zero so that adding a constant to an observable does not change its learning phase. The condition concerns derivatives at the chosen reference time and does not assume that the infinite Taylor series converges.

Let $a \to \infty$ and scale the hyperparameters as $\theta_i(a) \sim \kappa_i a^{\alpha_i}$ with $\kappa_i > 0$. The scaling vector $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)$ assigns growth exponent $\boldsymbol{q} \cdot \boldsymbol{\alpha}$ to the monomial $\boldsymbol{\theta}^{\boldsymbol{q}}$. We define the *leading support* at order s to be the set attaining the largest such exponent:

$$S_{\ell,s}(\boldsymbol{\alpha}) = \operatorname*{Argmax}_{\boldsymbol{q} \in Q_{\ell,s}} \boldsymbol{q} \cdot \boldsymbol{\alpha}. \tag{5}$$

Geometrically, this is the support on the face of $\mathrm{conv}(Q_{\ell,s})$ exposed by $\boldsymbol{\alpha}$. We call the collection $(S_{\ell,s}(\boldsymbol{\alpha}))_{\ell,s\geq 1}$ the *leading-support signature*, and put two scaling vectors in the same *formal phase* exactly when their signatures agree. Agreement through any fixed order K need not imply agreement at higher orders, so a finite truncation does not generally determine the phase. Because we retain monomials rather than the value of their sum, the definition also applies when tied leading terms cancel for special choices of the constants κ_i.

We call a boundary between adjacent formal phases *persistent* if their leading supports differ at infinitely many orders s. Equivalently, however many initial coefficients one discards, some later coefficients still select different monomials. Such a difference cannot be attributed to finitely many exceptional orders; it is an essential distinction between the phases at the level of their formal expansions.

The map from continuous scaling vectors to leading-support signatures turns the classification into a discrete combinatorial problem. For a fixed finite realization, an analytic gradient-flow trajectory is determined by its Taylor germ wherever it can be continued along the real-time solution. This motivates considering the entire coefficient sequence rather than only the initial slope. For averaged losses, however, the formal series need not converge, and coefficient-wise large-system limits need not commute with summation or analytic continuation. Directly constructing and comparing limiting trajectories can therefore be difficult even for the present linear autoencoder, whereas each finite-order coefficient can be analyzed exactly. Our signature diagram offers a systematic classification at that level and suggests distinctions relevant to the dynamics, but we do not claim it exhausts trajectory-level regimes. In particular, long-time behavior may require separate analysis, as in Section 4.

Autoencoder phase diagram. For the Gaussian autoencoder of Section 2, our observables are the averaged train loss $F_{\mathrm{tr}}(t) = \mathbb{E}[\widehat{L}(t)]$ and population loss $F_{\mathrm{pop}}(t) = \mathbb{E}[L(t)]$. The five positive hyperparameters are $\boldsymbol{\theta} = (p, n, m, \sigma, \eta)$, with scaling exponents $\boldsymbol{\alpha} = (\alpha_p, \alpha_n, \alpha_m, \alpha_\sigma, \alpha_\eta)$ and $\alpha_p, \alpha_n, \alpha_m > 0$. The learning rate η multiplies the whole gradient-flow vector field: $F_\ell(t; p, n, m, \sigma, \eta) = F_\ell(\eta t; p, n, m, \sigma, 1)$. Consequently, every order-s coefficient has the same factor η^s; α_η changes the clock but not the leading-support signature; thus η does not affect the phase.

We will use three convenient ratios: $\phi = p/m$ (input-to-sample), $\rho = n\sigma^2$ (relative initialization magnitude), and $\psi = p/n$ (input-to-hidden). Their scaling exponents are

$$\boldsymbol{\gamma} = (\alpha_\phi, \alpha_\rho, \alpha_\psi) = \big(\alpha_p - \alpha_m,\ \alpha_n + 2\alpha_\sigma,\ \alpha_p - \alpha_n\big). \tag{6}$$

Recall that the *normal fan* of a polytope partitions scaling directions according to which face maximizes a linear score on that polytope.

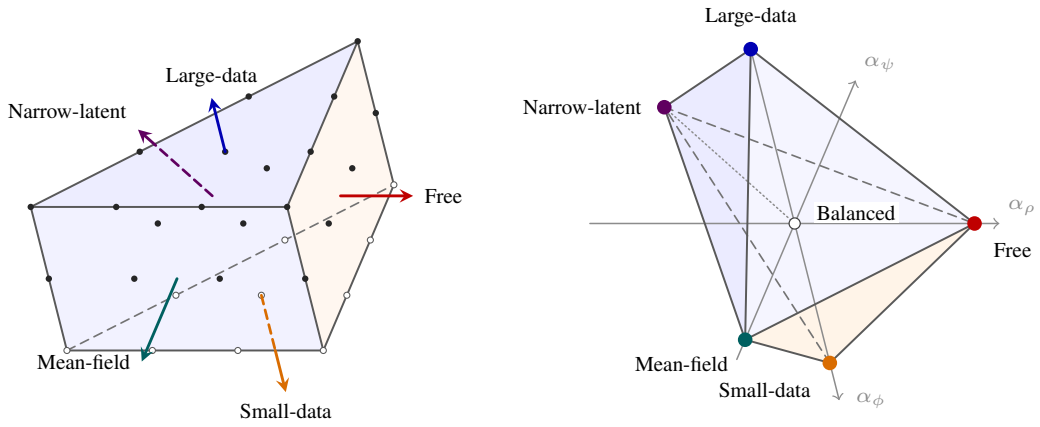


Figure 1: The autoencoder phase diagram in scaling coordinates $\boldsymbol{\gamma} = (\alpha_\phi, \alpha_\rho, \alpha_\psi)$ (**right**), and the dual triangular prisms formed by leading monomial supports $\cup_{\boldsymbol{\alpha}} S_{\mathrm{tr},s((\alpha))}, \cup_{\boldsymbol{\alpha}} S_{\mathrm{pop},s}((\alpha))$ at Taylor order $s = 2$ (**left**). The five colored face normals correspond to the five labeled vertex rays. Hollow dots on the lowest layer occur only for the train loss; the prism for the population loss begins one layer higher.

Theorem 3.1 (Autoencoder phase diagram)**.** *Assume that the training matrix $\boldsymbol{X} \in \mathbb{R}^{p\times m}$ has i.i.d. standard Gaussian entries, and that the initialization is Gaussian as in Eq.* (3)*. Then, at every order $s \geq 1$, the averaged train and population losses have Laurent-polynomial coefficients in $\boldsymbol{\theta}$. Their joint formal phase depends only on $\boldsymbol{\gamma}$ and is the normal fan of a triangular prism. Equivalently, the ray from $\mathbf{0}$ through each nonzero $\boldsymbol{\gamma}$ meets the boundary of the dual triangular bipyramid; its phase is determined by the unique face whose relative interior contains that intersection point. The bipyramid has five vertices, nine edges, and six triangular faces; the balanced point $\boldsymbol{\gamma} = \mathbf{0}$ represents the remaining phase. The five vertex rays are:*

$$\begin{array}{llll} \alpha_\phi < 0, & \alpha_\rho = \alpha_\psi = 0 & & \textit{(“Large-data”)}, \\ \alpha_\psi < 0, & \alpha_\phi = \alpha_\rho = 0 & & \textit{(“Mean-field”)}, \\ \alpha_\rho < 0, & \alpha_\psi = -\alpha_\rho, & \alpha_\phi = 0 & \textit{(“Narrow-latent”)}, \\ \alpha_\phi > 0, & \alpha_\rho = \alpha_\psi = 0 & & \textit{(“Small-data”)}, \\ \alpha_\rho > 0, & \alpha_\phi = \alpha_\psi = 0 & & \textit{(“Free”)}. \end{array} \tag{7}$$

Every transition between adjacent phases is persistent.

Proof sketch. Repeated differentiation of gradient flow and Gaussian Wick contraction give the Laurent-polynomial coefficients. For $s \geq 2$, their potentially leading monomials are indexed by integers $u, v \geq 0$, $u + v \leq s + 1$, and $r \in \{0, \ldots, s\}$ for the train loss, or $r \in \{1, \ldots, s\}$ for the population loss. Up to a term constant across these monomials, the exponent for $\mathbf{k} = (u, v, r)$ is $\mathbf{k} \cdot \mathbf{w}$, where $\mathbf{w} = (\alpha_\rho + \alpha_\psi, \alpha_\rho, -\alpha_\phi)$. Equivalently, the affine coordinates $\mathbf{x} = (u + v, u, r)$ turn this exponent into $\alpha_\rho x_1 + \alpha_\psi x_2 - \alpha_\phi x_3$. The index sets are triangular prisms, all with the same normal fan (see Figure 1). Distinct phase cells select different monomials at every $s \geq 2$, yielding persistent transitions. See the full proof in Appendix B. □

Figure 1 shows the phase diagram. Table 1 expresses the five vertex regimes in terms of the original scales of m, p, n, σ^2 and summarizes their limiting dynamics (see Section C). The relations explain the names: “Large-data” and “Small-data” compare m with p; “Mean-field” and “Narrow-latent” highlight different latent-width scalings; in “Free”, the initialized model dominates the target.

Other phases combine these basic extremes. An open bipyramid edge interpolates between its two endpoint rays and selects the common edge of their prism facets; an open triangular face combines three vertex rays and selects their common prism vertex. At the balanced center, all monomials of the corresponding prisms tie. The “Large-data” and “Small-data” regimes are opposite to each other; this is the only opposite pair among the five vertex regimes.

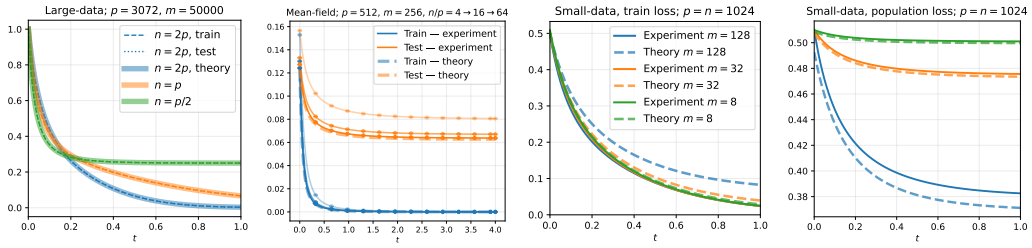


Figure 2: Empirical validation of the limiting predictions (Sec. 4) across different regimes on subsets of CIFAR-10. See D for experimental details and additional plots. *Narrow-latent* is isolated on Fig. 3.

Table 1: Five basic regimes, ordered by complexity from lowest (top row) to highest (bottom row).

Regime	**Scaling**	**Limiting dynamics**	**Solution status**
Large-data	$m \gg p \asymp n \asymp \sigma^{-2}$; $\eta = p$	Spectral logistic flow of $\boldsymbol{A}$ over MP law $\mu_{p/n}$	Closed-form formula via MP generating function
Mean-field	$n \asymp \sigma^{-2} \gg p \asymp m$; $\eta = p$	Decoupled empirical covariance eigenmodes; average over $\nu_{p/m}$	Closed-form integral formula; large-time asymptotics
Narrow-latent	$n \ll \sigma^{-2} \asymp p \asymp m$; $\eta = p$	Row-decoupled dynamics for row-norm a and data alignment r	Finite-dimensional implicit ODE; explicit asymptotics
Small-data	$m \ll p \asymp n \asymp \sigma^{-2}$; $\eta = m$	Active/inactive block dynamics with leakage moments $a, \{\mu_k\}_{k\geq 1}$	Infinite moment hierarchy; numerical truncation
Free	$\sigma^{-2} \ll n \asymp p \asymp m$; $\eta = \sigma^{-2}$	Target term negligible; dynamics approximately deflate initial model	Classified by the prism; trajectory solution open

4 Limiting solutions in extreme regimes (C)

We present explicit solutions for both train and population losses in regimes corresponding to four out of five two-dimensional faces of the prism. These regimes are naturally arranged by the complexity of the final solution, as summarized in Table 1.

As noted in Section 3, the learning rate η only rescales time. We choose it so that the relevant matrix dynamics evolve on times $t = O(1)$. Indeed, Eq. (3) gives $\dot{\boldsymbol{U}} = \frac{\eta}{p}\boldsymbol{U}(2\boldsymbol{S} - \boldsymbol{S}\boldsymbol{A} - \boldsymbol{A}\boldsymbol{S})$ with $\boldsymbol{A} = \boldsymbol{U}^\top \boldsymbol{U}$. In the large-data, mean-field, and narrow-latent regimes, $\|\boldsymbol{S}\|_{\mathrm{op}}$ and $\|\boldsymbol{A}(0)\|_{\mathrm{op}}$ are of order one, giving a characteristic rate η/p. In the small-data regime, the nonzero eigenvalues of $\boldsymbol{S}$ are of order p/m, giving rate η/m. In the free regime, $\|\boldsymbol{A}(0)\|_{\mathrm{op}}$ is of order $p\sigma^2 \gg 1$, giving rate $\eta\sigma^2$. The choices $\eta = p, m$, and σ^{-2} in Table 1 normalize these respective rates.

Recall the parameters $\phi = p/m, \rho = n\sigma^2, \psi = p/n$ that appeared in our phase diagram analysis. Each of the five basic extreme regimes admits a two-parameter refinement by varying these parameters or their combinations. For mean-field and narrow-latent results, we impose the following common spectral assumption. For some deterministic $K < \infty$ and deterministic probability measure ν_ϕ,

$$\mathbb{P}(\|\boldsymbol{S}\|_{\mathrm{op}} \leq K) \longrightarrow 1, \qquad \frac{1}{p}\sum_{i=1}^{p} \delta_{\lambda_i(\boldsymbol{S})} \Longrightarrow \nu_\phi \qquad \text{in prob.} \tag{8}$$

In the large-data regime these follow from the stronger near-isotropy assumption stated in 4.2; the small-data regime requires a different normalization stated in 4.4. All the theorems in the present section prove convergence of losses and other relevant quantitites, uniformly in probability on all compact time intervals. We therefore adopt the following standard stochastic process notation:

$$\boxed{Z = o_{\mathbb{P}}^{\mathrm{ucp}}(1)} \quad \equiv \quad \boxed{\sup_{0\leq t\leq T} \|Z(t)\| \to 0 \ \forall T \in \mathbb{R}_+ \text{ under a given joint } p, n, m \to \infty \text{ limit.}}$$

We check our solutions numerically on subsets of CIFAR10: Fig. 2. In order to ensure data isotropy Eq. (2), we whiten it using the whole training set of 50000 examples, see Appendix D for details. Note that the subset we actually train on still has non-trivial $\boldsymbol{S}$.

4.1 MEAN-FIELD (C.2)

In the corresponding limit, $\boldsymbol{A}(0)$ becomes scalar, and the whole dynamics of $\boldsymbol{A}(t)$ decouple into independent ones for each eigencomponent of $\boldsymbol{S}$:

Theorem 4.1 (Short version of Theorem C.8 + Corollary C.11)**.** *Consider $p, n, m \to \infty$ in such a way that $\psi \to 0$, while ϕ, ρ stay finite. Then under Eq.* (8)*,*

$$\boldsymbol{A}(\cdot) - a(\cdot; \boldsymbol{S}) = o_{\mathbb{P}}^{\mathrm{ucp}}(1), \qquad a(t; \lambda) := \frac{\rho}{\rho + (1-\rho)e^{-4\lambda t}}. \tag{9}$$

Moreover, $\widehat{L}(\cdot) - \widehat{\mathcal{L}}(\cdot) = o_{\mathbb{P}}^{\mathrm{ucp}}(1)$ and (under Eq. (2)*) $L(\cdot) - \mathcal{L}(\cdot) = o_{\mathbb{P}}^{\mathrm{ucp}}(1)$ with*

$$\mathcal{L}(t) = \int \ell(\lambda t)\, d\nu_\phi(\lambda), \qquad \widehat{\mathcal{L}}(t) = \int \lambda \ell(\lambda t)\, d\nu_\phi(\lambda), \qquad \ell(t) := \frac{(1-\rho)^2 e^{-8t}}{2\left(\rho + (1-\rho)e^{-4t}\right)^2}. \tag{10}$$

Moreover, if $\boldsymbol{X}$ has Gaussian entries then the same holds for $\mathbb{E}[L(t)]$ and $\mathbb{E}[\widehat{L}(t)]$.

Gaussian data. Although ν_ϕ becomes the Marchenko-Pastur (MP) distribution, the above integrals cannot be evaluated analytically. Nevertheless, their large-t asymptotics of limit losses $\mathcal{L}$ and $\widehat{\mathcal{L}}$ are tractable. Indeed, since $\ell(t)$ decays exponentially, and ν_ϕ features a spectral gap for any $\phi \neq 1$ (i.e. there is a gap between zero and the bulk of the spectrum), $\widehat{\mathcal{L}}(t)$ also decays exponentially when $\phi \neq 1$. When $\phi < 1$, this also applies to $\mathcal{L}(t)$. For $\phi > 1$, there is a point mass at zero, which does not affect $\widehat{\mathcal{L}}(t)$ due to the factor λ, while it results in a remainder term for $\mathcal{L}(t)$: $\mathcal{L}(\infty) = \frac{(1-\rho)^2}{2}\left(1 - \frac{1}{\phi}\right)_+$. The final population loss is zero when $\rho = 1$. This is due to the fact that $\boldsymbol{U}^\top \boldsymbol{U} \to \boldsymbol{I}_p$ as $n \to \infty$ for fixed p due to LLN, whenever $\sigma^2 = 1/n$. Suppose $\rho \neq 1$; otherwise, $\mathcal{L} = \widehat{\mathcal{L}} \equiv 0$. When $\phi = 1$, the spectral gap disappears; because of this, both $\widehat{\mathcal{L}}(t)$ and $\mathcal{L}(t)$ decay much slower:

$$\mathcal{L}(t) - \mathcal{L}(\infty) \asymp \begin{cases} \frac{1}{t^{1/2}}, & \phi = 1, \\ \frac{e^{-8(1-\sqrt{\phi})^2 t}}{t^{3/2}}, & \phi \neq 1, \end{cases} \qquad \widehat{\mathcal{L}}(t) \asymp \begin{cases} \frac{1}{t^{3/2}}, & \phi = 1, \\ \frac{e^{-8(1-\sqrt{\phi})^2 t}}{t^{3/2}}, & \phi \neq 1 \end{cases} \tag{11}$$

as $t \to \infty$, where the exact expressions are derived in C.2.1. As a result, if we fix large enough t, the training loss as function of ϕ features a maximum near $\phi = 1$, i.e. when the problem becomes critically determined. This is in parallel with double descent phenomena: the population loss peaks at $m = p$ for linear regression trained on a noisy linear teacher. The nature of the peak is different, however: for a linear regression, it is caused by the fact that the model has to spend all its capacity to fit the teacher, thus becoming very sensitive to noise, while in our case, there is no noise in the target; the peak is caused by vanishing non-zero eigenvalues of $\boldsymbol{S}$ which appear when $m = p$ and dramatically slow down the gradient flow.

We emphasize that we *do not* expect these asymptotics to hold exactly for *finite* p, n, m losses $L(t)$ and $\widehat{L}(t)$. This is due to the fact that the above theorem proves convergence *only on compact* time intervals. Therefore we would expect that for large enough p, n, m, the above asymptotic formulas could be numerically close for large t, but not for infinite t.

4.2 LARGE-DATA (C.4)

Here, we adopt a stronger condition for data covariance: $\|\boldsymbol{S} - \boldsymbol{I}_p\|_{\mathrm{op}} \to 0$. Since the empirical covariance becomes the identity, both train and population losses follow the same logistic curve:

Theorem 4.2 (Short version of Theorem C.27)**.** *Consider $p, n, m \to \infty$ in such a way that $\phi \to 0$, while ψ, ρ stay finite. Then under $\|\boldsymbol{S} - \boldsymbol{I}_p\|_{\mathrm{op}} \to 0$ in probability,*

$$\boldsymbol{A}(\cdot) - \Theta(\cdot; \boldsymbol{A}(0)) = o_{\mathbb{P}}^{\mathrm{ucp}}(1), \qquad \Theta(t; s) := \frac{s}{s + (1-s)e^{-4t}}, \tag{12}$$

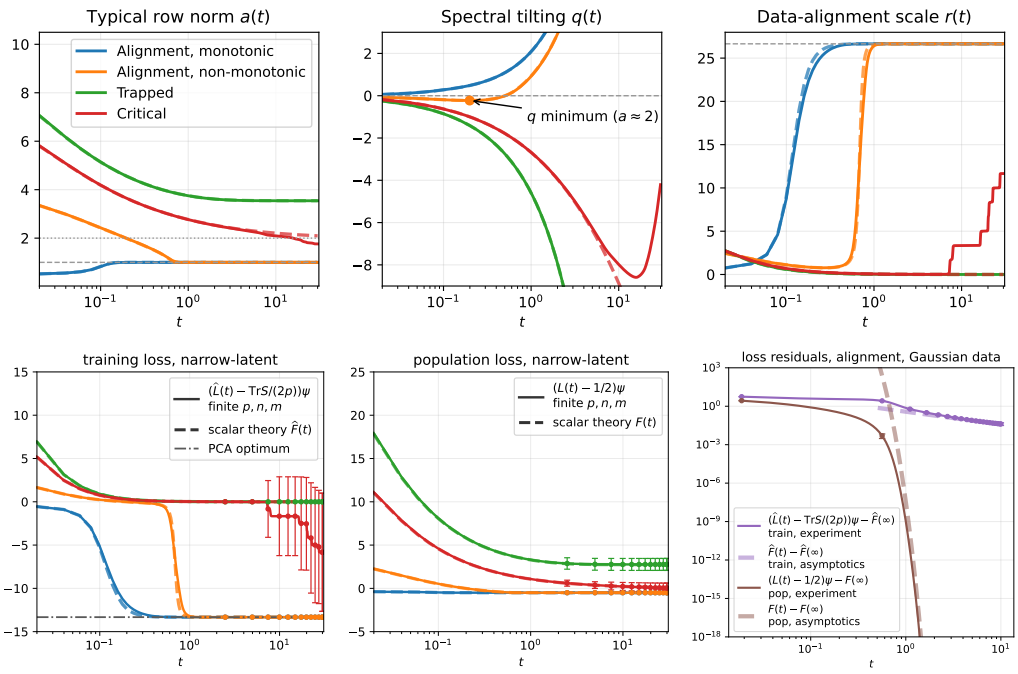


Figure 3: We measure quantities related to *narrow-latent* and compare to their theoretical limits on a CIFAR10 subset, see Section 4.3 for details. The bottom-right plot checks Eq. (20) and uses Gaussian data. See D for implementation details.

$$L(\cdot) - \mathcal{L}(\cdot) = o_{\mathbb{P}}^{\mathrm{ucp}}(1), \qquad \widehat{L}(\cdot) - \mathcal{L}(\cdot) = o_{\mathbb{P}}^{\mathrm{ucp}}(1), \qquad \mathcal{L}(t) = \frac{1}{2}\int (\Theta(t;\rho\lambda) - 1)^2 \, \mu_\psi(d\lambda), \tag{13}$$

where μ_ψ is the MP-distribution. The integral admits a closed formula Eq. (618)*. If, moreover, $\mathbb{E}[\|\boldsymbol{S}\|_{op}^{1+\delta}] = O(1)$ for some $\delta > 0$ then the same holds for $\mathbb{E}[\widehat{L}(t)]$ and $\mathbb{E}[L(t)]$.*

The spectrum of $\boldsymbol{A}(0)/\rho$ follows a MP-distribution μ_ψ. This causes a qualitative difference in limit loss $\mathcal{L}(t)$ behavior for large t between $\psi = 1$ and $\psi \neq 1$, similar to that in the mean-field regime discussed below. The difference is again caused by the spectral gap present when $\psi \neq 1$, but in contrast to the mean-field regime, the loss decays exponentially in both scenarios, albeit with different exponents (see C.4.1 for derivations):

$$\mathcal{L}(t) - \mathcal{L}(\infty) \asymp \begin{cases} e^{-2t}, & \psi = 1, \\ e^{-8t}, & \psi \neq 1, \end{cases} \qquad \mathcal{L}(\infty) = \frac{p - \min(n,p)}{2p} = \frac{1}{2}\left(1 - \frac{1}{\psi}\right)_+, \tag{14}$$

where the second formula follows from the fact that every nonzero eigenvalue of $\boldsymbol{A}(0)$ converges to 1, while zero ones stay zero, while $\operatorname{rank} \boldsymbol{A}(0) = \min(n,p)$. As in the mean-field limit, the same remark on large-t asymptotics of $L(t)$ for finite p, n, m applies here.

4.3 Narrow-latent (C.3)

Define the exponential transform and its logarithmic derivative by

$$M_{\nu_\phi}(q) := \int e^{q\lambda} \, d\nu_\phi(\lambda), \qquad m_{\nu_\phi}(q) := \frac{M'_{\nu_\phi}(q)}{M_{\nu_\phi}(q)}. \tag{15}$$

The theorem below establishes limits of latent-space Gramians $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{U}^\top$, $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\top$. They become scalar (converging, respectively, to a and r defined below), while the model progressively aligns with the data covariance spectrum (the evolution of q defined below):

Theorem 4.3 (Short version of Theorem C.12 + Corollary C.23)**.** *Consider $p, n, m \to \infty$ in such a way that $\psi \to \infty$, while $\phi, \psi\rho$ stay finite. Then under Eq.* (8)*,*

$$\boldsymbol{G}(\cdot) - a(\cdot)\boldsymbol{I} = o_{\mathbb{P}}^{\mathrm{ucp}}(1), \qquad \boldsymbol{H}(\cdot) - r(\cdot)\boldsymbol{I} = o_{\mathbb{P}}^{\mathrm{ucp}}(1), \tag{16}$$

$$\widehat{L}(\cdot)-\frac{\operatorname{Tr}\boldsymbol{S}}{2p}=o_{\mathbb{P}}^{\mathrm{ucp}}(1),\quad \left(\widehat{L}(\cdot)-\frac{\operatorname{Tr}\boldsymbol{S}}{2p}\right)\psi-\widehat{F}(\cdot)=o_{\mathbb{P}}^{\mathrm{ucp}}(1),\quad \widehat{F}(t)=r(t)\left(\frac{a(t)}{2}-1\right), \tag{17}$$

where $r(t)=a(t)m_{\nu_\phi}(q(t))$*, and* $a(t), q(t)$ *are the solution of the ODE system below:*

$$\dot{a}(t)=4a(t)m_{\nu_\phi}(q(t))(1-a(t)),\qquad \dot{q}(t)=2(2-a(t)),\qquad a(0)=\psi\rho,\qquad q(0)=0. \tag{18}$$

Moreover, under the isotropy condition Eq. (2)*,*

$$L(\cdot)-\frac{1}{2}=o_{\mathbb{P}}^{\mathrm{ucp}}(1),\quad \left(L(\cdot)-\frac{1}{2}\right)\psi-F(\cdot)=o_{\mathbb{P}}^{\mathrm{ucp}}(1),\quad F(t)=a(t)\left(\frac{a(t)}{2}-1\right). \tag{19}$$

If, moreover, $\mathbb{E}[\|\boldsymbol{S}\|_{op}^{1+\delta}]=O(1)$ *for some* $\delta>0$ *then (17, 19) hold also for* $\mathbb{E}[\widehat{L}(t)]$ *and* $\mathbb{E}[L(t)]$*.*

The scalars have a direct interpretation. Since the theorem above shows $\boldsymbol{G}=\boldsymbol{U}\boldsymbol{U}^\top\approx a\boldsymbol{I}$, the quantity a is the asymptotic squared norm of a typical latent row of $\boldsymbol{U}$, while the off-diagonal part of $\boldsymbol{G}$ vanishes. Likewise $\boldsymbol{H}=\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\top\approx r\boldsymbol{I}$, so r is the corresponding covariance-weighted row energy, or data-alignment scale. Finally, q encodes spectral-tilting: $q>0$ progressively favors large eigenvalue directions, whereas $q<0$ favors the lower edge and, when $\phi>1$, the null space of $\boldsymbol{S}$.

The limiting ODE Eq. (18) exhibits several qualitatively different behaviors, governed by the initialization magnitude $\kappa:=a(0)=\psi\rho=p\sigma^2$ and by the null-space mass $w_0:=\nu_\phi(\{0\})$

Proposition 4.4 (Qualitative behavior of the narrow-latent limit)**.** *Under* $\nu_\phi\neq\delta_0$*,*

1. ***Monotonic alignment*** *(blue curves on Fig. 3). If* $0<\kappa\leq 2$*, then* $q(t)$ *is nondecreasing and* $q(t)\to+\infty$*,* $a(t)\to 1$*. Thus the dynamics favor the upper edge of the data spectrum.*
2. ***Non-monotonic alignment*** *(orange curves). Suppose* $\kappa>2$ *and* $\frac{\kappa^2}{\kappa-1}w_0^2<4$*. Then* $q(t)$ *initially decreases, so the dynamics move toward the lower spectral edge, but this motion eventually reverses. In particular,* $q(t)$ *has a finite minimum, hence* $q(t)\to+\infty$*,* $a(t)\to 1$*.*
3. ***Null-space trapping*** *(green curves). Suppose* $\kappa>2$ *and* $\frac{\kappa^2}{\kappa-1}w_0^2>4$*. The reversal never occurs:* $q(t)\to-\infty$*,* $r(t)\to 0$*,* $a(t)\to a_*>2$*. The limiting dynamics become trapped in the null space of the empirical covariance rather than aligning with its principal directions.*
4. ***Critical boundary*** *(red curves). At* $\frac{\kappa^2}{\kappa-1}w_0^2=4$ *one has* $q(t)\to-\infty$*,* $a(t)\to 2$*, giving the boundary between alignment and strict trapping.*

In particular, if $w_0=0$, trapping is impossible for any finite κ. The proof, the first integral underlying the threshold above, and detailed large-time asymptotics are given in C.3.6. For Gaussian data, $w_0=\max\left(0,1-\frac{1}{\phi}\right)$, so a nontrivial trapping transition can occur only when $\phi>1$, i.e. when the empirical covariance has a macroscopic null space. In the alignment cases, $q(t)\to+\infty$ concentrates the row-energy distribution near the upper spectral edge. When the eigenvalues relevant to the rank-n PCA optimum converge to the same edge, as in the Gaussian setting considered here, this recovers the PCA solution.

The trapped branch also has a finite-dimensional optimization interpretation. C.3.7 shows that null-space critical points satisfying $\boldsymbol{U}\boldsymbol{U}^\top\succ 2\boldsymbol{I}$ are nonglobal local minima of the tied empirical loss. Thus the strict limit $a_*>2$ lies in this local-minimum region, while the critical value $a_*=2$ lies on its degenerate boundary. This is a landscape interpretation of the large-time limiting system: Theorem 4.3 only establishes convergence of the finite-dimensional dynamics on fixed time intervals, and does not assert convergence of a finite-dimensional trajectory to such a critical point as $t\to\infty$.

A second notable feature is that successful alignment does not imply the same relaxation time for training and population performance. For Gaussian data, in the monotonic-alignment regime with $\kappa\neq 1$, the population correction approaches its limiting value exponentially, whereas the training correction does so only algebraically:

$$F(t)-F(\infty)\asymp t^6e^{-8\lambda_+t},\qquad \widehat{F}(t)-\widehat{F}(\infty)\asymp t^{-1}. \tag{20}$$

Intuitively, once the row norms have nearly equilibrated, the remaining slow motion is primarily spectral alignment. This continues to change the training loss, which is sensitive to the empirical covariance spectrum, while having a much weaker effect on the isotropic population loss. Fig. 3 illustrates both the alignment/trapping alternatives and this separation of time scales. Further asymptotics for all four cases are given in C.3.6.

4.4 SMALL-DATA (C.5)

The relevant data assumption is the active-space near-isometry $\|\boldsymbol{\Gamma}-\boldsymbol{I}_m\|_{\mathrm{op}} \to 0$, where $\boldsymbol{\Gamma} = \frac{1}{p}\boldsymbol{X}^\top\boldsymbol{X}$.

Theorem 4.5 (Short version of Theorem C.31)**.** *Consider $p, n, m \to \infty$ in such a way that $\phi \to \infty$, while ψ, ρ stay finite. Then under $\|\boldsymbol{\Gamma} - \boldsymbol{I}_m\|_{\mathrm{op}} \to 0$ in probability, $\widehat{L}(\cdot) - \widehat{\mathcal{L}}(\cdot) = o_{\mathbb{P}}^{\mathrm{ucp}}(1)$ and $L(\cdot) - \mathcal{L} = o_{\mathbb{P}}^{\mathrm{ucp}}(1)$, where*

$$\mathcal{L} = \frac{(1-\rho)^2 + \psi\rho^2}{2}, \qquad \widehat{\mathcal{L}}(t) = \frac{(1-a(t))^2 + a(t)\mu_1(t)}{2}. \tag{21}$$

where $(a, \mu_{k\geq 1})$ are a unique solution of a hierarchical system of ODEs specified in Theorem C.31. The $O(1/\phi)$ correction for L also converges: $(L(\cdot) - L(0))\phi - C(\cdot) = o_{\mathbb{P}}^{\mathrm{ucp}}(1)$, where $C(t)$ is given by an explicit functional involving a, μ_1, and μ_2. Moreover, if $\boldsymbol{X}$ has Gaussian entries then the same holds for $\mathbb{E}[L(t)]$, $\mathbb{E}[\widehat{L}(t)]$, and $\mathbb{E}[(L(t) - L(0))\phi]$.

A direct implication of the above is that the train loss exhibits $O(1)$ change, while that of the population one is only $O(1/\phi)$. In fact, one can easily check directly that $C'(t) \leq 0$, hence the limit population loss decreases monotonically. As for the other regimes, we check the loss convergence to the theoretical limit numerically on subsets of CIFAR10: Fig. 2, two last plots. Due to the lack of space, we do not discuss the physical meaning of $(a, \mu_{k\geq 1})$ in the main. We also leave a detailed analysis of the limit ODE hierarchy for future work.

5 DISCUSSION, LIMITATIONS, AND FUTURE WORK

Our main contribution is a systematic classification of extreme learning regimes for weight-tied linear autoencoders. The diagrammatic expansion of Section 3 organizes dominant loss terms into a triangular-prism hierarchy, whose five facets correspond to the large-data, small-data, mean-field, narrow-latent, and free regimes. Section 4 provides an independent dynamical justification of this picture: four of these facets lead to qualitatively different limiting systems, ranging from decoupled spectral modes to a nonlinear alignment dynamics and an infinite hierarchy of leakage moments. Thus the prism provides not only a classification of formal loss expansions, but also a map between asymptotic scalings and distinct mechanisms of learning.

Spectral geometry and initialization dependence. A recurring theme is that these mechanisms are controlled by spectral geometry. Closing spectral gaps cause critical slowing at $p/m = 1$ in the mean-field regime and at $p/n = 1$ in the large-data regime. In the narrow-latent regime, the spectral clock $q(t)$ instead dynamically tilts the directions explored by a typical latent row. This produces particularly rich initialization dependence: for sufficiently small initialization the model aligns with leading data directions and approaches PCA, while for larger initialization $q(t)$ can first move toward the lower spectral edge before reversing. With a sufficiently large initialization and a nontrivial data null space, the reversal never occurs and the limiting dynamics become trapped, so that the training loss does not approach its global PCA minimum.

Train–population separation. The solutions show that training and population losses can evolve on parametrically different scales. At the large-data facet the two losses coincide asymptotically, whereas at the opposite small-data facet training changes at order one while the leading population loss is frozen and its first nontrivial change is only of order m/p. In the narrow-latent alignment phase the separation is dynamical: after row norms have nearly equilibrated, the remaining slow motion is mainly spectral alignment, which continues to affect training loss but is largely invisible to isotropic population loss. For Gaussian data this yields the striking asymptotic separation Eq. (20).

Limitations and open problems. The prism classification is complete at the level of dominant expansion monomials, but this does not by itself prove completeness of all possible limiting trajectories. The free facet remains a natural target for analysis. Moreover, our convergence theorems hold uniformly only on fixed time intervals, whereas the large-time asymptotics are taken after the high-dimensional limit; understanding the joint large-dimension/large-time regime is therefore an important open problem. Finally, the cleanest population results rely on isotropy or related spectral assumptions.

AI USE STATEMENT

We used a popular LLM extensively as a research assistant in developing Appendix C. All formal results in Appendix C were formulated and proved with substantial assistance from the LLM. This included assistance in identifying appropriate theorem statements, constructing proof strategies, working out technical arguments, and drafting the resulting proofs. The authors independently inspected, checked, and revised all such arguments before including them in the manuscript and take full responsibility for their correctness. We also used an AI assistant to facilitate coding and as an auxiliary tool for polishing the manuscript.

No LLM was used in developing the prism classification. In particular, the diagrammatic construction, the classification of extreme regimes, and the results of Section 3 and Appendix B were developed independently by the authors.

REPRODUCIBILITY STATEMENT

The main theoretical results are stated with their assumptions in the main text, and complete proofs are provided in the appendices. For the numerical experiments, we include Jupyter Notebooks used to generate all experimental results and figures in the paper.

REFERENCES

Advani, M. S., Saxe, A. M., and Sompolinsky, H. (2020). High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446.

Baldi, P. (2012). Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 37–49. JMLR Workshop and Conference Proceedings.

Baldi, P. and Hornik, K. (1989). Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Bourlard, H. and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4):291–294.

Chizat, L., Oyallon, E., and Bach, F. (2019). On lazy training in supervised differentiable programming. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Cui, H. and Zdeborová, L. (2023). High-dimensional asymptotics of denoising autoencoders. *Advances in Neural Information Processing Systems*, 36:11850–11890.

Golikov, E. (2020a). Dynamically stable infinite-width limits of neural classifiers. *arXiv preprint arXiv:2006.06574*.

Golikov, E. (2020b). Towards a general theory of infinite-width limits of neural classifiers. In *International Conference on Machine Learning*, pages 3617–3626. PMLR.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural Tangent Kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580.

Kunin, D., Bloom, J., Goeva, A., and Seed, C. (2019). Loss landscapes of regularized linear autoencoders. In *International conference on machine learning*, pages 3560–3569. PMLR.

Lampinen, A. K. and Ganguli, S. (2019). An analytic theory of generalization dynamics and transfer learning in deep linear networks. In *International Conference on Learning Representations*.

Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. (2018). Deep neural networks as Gaussian processes. In *International Conference on Learning Representations*.

Nguyen, P.-M. (2021). Analysis of feature learning in weight-tied autoencoders via the mean field lens. *arXiv preprint arXiv:2102.08373*.

Nguyen, T. V., Wong, R. K., and Hegde, C. (2019). On the dynamics of gradient descent for autoencoders. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2858–2867. PMLR.

Pretorius, A., Kroon, S., and Kamper, H. (2018). Learning dynamics of linear denoising autoencoders. In *International Conference on Machine Learning*, pages 4141–4150. PMLR.

Refinetti, M. and Goldt, S. (2022). The dynamics of representation learning in shallow, non-linear autoencoders. In *International Conference on Machine Learning*, pages 18499–18519. PMLR.

Rotskoff, G. M. and Vanden-Eijnden, E. (2018). Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *stat*, 1050:22.

Saxe, A. M., McClelland, J. L., and Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

Saxe, A. M., McClelland, J. L., and Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546.

Sirignano, J. and Spiliopoulos, K. (2020). Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752.

Tu, Z., Aranguri Diaz, S. T., and Jacot, A. (2024). Mixed dynamics in linear networks: Unifying the lazy and active regimes. *Advances in Neural Information Processing Systems*, 37:106059–106104.

Woodworth, B., Gunasekar, S., Lee, J. D., Moroshko, E., Savarese, P., Golan, I., Soudry, D., and Srebro, N. (2020). Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pages 3635–3673. PMLR.

Yang, G. and Hu, E. J. (2021). Tensor Programs IV: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pages 11727–11737. PMLR.

Yarotsky, D., Golikov, E., and Gusev, Y. (2026). Gradient flow through diagram expansions: Learning regimes and explicit solutions. In *Forty-third International Conference on Machine Learning*.

CONTENTS

1 Introduction **1**

2 Problem statement **2**

3 A formal phase diagram of learning regimes (B) **2**

4 Limiting solutions in extreme regimes (C) **5**

4.1 Mean-field (C.2) 6

4.2 Large-data (C.4) 6

4.3 Narrow-latent (C.3) 7

4.4 Small-data (C.5) 9

5 Discussion, limitations, and future work **9**

Bibliography **10**

A Related works **14**

A.1 Linear autoencoders, PCA, and optimization landscapes 14

A.2 Training and generalization dynamics 14

A.3 Large-size limits and learning-regime classification 14

A.4 High-dimensional generalization and spectral criticality 15

B Proof of the autoencoder phase diagram (Theorem 3.1) **15**

B.1 Laurent coefficients and diagram expansions 15

B.2 Monomials that can be leading 17

B.3 The normal fan and persistent transitions 18

C Solutions of extreme regimes (Section 4) **19**

C.1 Common notation, stability principles, and preliminary results 19

C.2 Mean-field (Section 4.1) 24

C.2.1 Large time behavior 30

C.3 Narrow-latent (Section 4.3) 32

C.3.1 Exact identities and a priori bounds 35

C.3.2 The empirical scalar system and proxy dynamics 35

C.3.3 Self-consistency of the proxy and stability of the exact flow 37

C.3.4 Passage to a general limiting spectral measure 41

C.3.5 Consequences for population and train losses 45

C.3.6 Large-time behavior of the limiting system 48

C.3.7 Landscape interpretation of null-space trapping 56

C.3.8 Role of Gaussianity and brief extension to anisotropic data 59

C.4 Large-data (Section 4.2) 60
C.4.1 Large-time behavior 66
C.5 Small-data (Section 4.4) 69
C.5.1 Exact orthogonal-data reduction 71
C.5.2 The exact matrix moment hierarchy 72
C.5.3 Uniform a priori bounds 73
C.5.4 A Banach algebra for the hierarchy 73
C.5.5 Scalarization of the random initial condition 74
C.5.6 Propagation of scalarity 76
C.5.7 Identification of the scalar hierarchy 77
C.5.8 The active spectral measure and uniqueness 78
C.5.9 Train loss for orthogonal data 79
C.5.10 Population loss and its first correction for orthogonal data 79
C.5.11 Transfer from orthogonalized to general near-isometric data 80
C.5.12 Train loss for general near-isometric data 82
C.5.13 Population correction for general near-isometric data 82
C.5.14 Instantiation for isotropic Gaussian training data 83
C.5.15 Remarks on anisotropic data 85

D Experimental details **85**
D.1 Recurring experimental setups 86
D.2 Large-data experiments 86
D.3 Mean-field experiments 87
D.4 Narrow-latent experiments 88
D.5 Small-data experiments 90

A Related works

A.1 Linear autoencoders, PCA, and optimization landscapes

Prior work. Linear autoencoders provide one of the classical connections between neural networks and principal component analysis. Early works characterized their optimal solutions in terms of singular-value decomposition and principal subspaces (Bourlard and Kamp, 1988; Baldi and Hornik, 1989), with later work giving a broader analytical treatment of linear autoencoders (Baldi, 2012). More recently, Kunin et al. (2019) analyzed the critical-point structure of regularized linear autoencoders and showed how regularization resolves part of the non-identifiability of the latent representation.

Our contribution. Rather than characterizing optima or the loss landscape, we study the complete gradient-flow trajectory of a weight-tied linear autoencoder from random initialization. Our asymptotic analysis simultaneously varies the input dimension, latent dimension, training-set size, and initialization magnitude. This places the usual undercomplete and overcomplete distinction inside a larger family of high-dimensional learning regimes and allows us to study training and population losses throughout learning, rather than only the final solution.

A.2 Training and generalization dynamics

Prior work. Exact learning dynamics have been derived for several linear or otherwise tractable neural-network models. Saxe et al. (2013; 2019) obtained exact solutions for deep linear networks under suitably aligned initializations. For linear autoencoders, Pretorius et al. (2018) integrated the gradient flow on a finite training set, again relying on an aligned initialization. Nguyen et al. (2019) studied gradient descent in weight-sharing autoencoders for several structured generative models. For nonlinear shallow autoencoders, Nguyen (2021) and Refinetti and Goldt (2022) derived high-dimensional dynamics in, respectively, wide mean-field and narrow-latent settings, using online SGD.

Related work has also studied train and test dynamics in supervised linear models. In particular, Lampinen and Ganguli (2019) and Advani et al. (2020) derive analytically tractable generalization dynamics for high-dimensional linear networks. Finally, Cui and Zdeborová (2023) studies finite-sample autoencoder generalization, but at the level of the empirical risk minimizer rather than its training trajectory.

Our contribution. We derive limiting train and population trajectories for a finite, repeatedly used training set and random isotropic initialization. We obtain rigorous finite-time limits in four qualitatively different regimes. These solutions also expose phenomena that are invisible from the final estimator alone: train and population losses may evolve on parametrically different time scales, and in the small-data regime training changes by $O(1)$ while the population loss changes only at order m/p.

A.3 Large-size limits and learning-regime classification

Prior work. Different infinite-size scalings can lead to qualitatively different neural network dynamics. Classical examples include mean-field and NTK/lazy limits, as well as intermediate or feature-learning scalings. Several works have systematically compared or classified such limits (Golikov, 2020b;a; Yang and Hu, 2021). Woodworth et al. (2020) emphasized in particular that initialization scale can control transitions between kernel and rich regimes. Most such classifications send network width to infinity while the remaining problem dimensions are fixed.

Most closely related to our approach, Yarotsky et al. (2026) proposed a diagram-expansion method in which learning regimes are identified from dominant families of terms in the loss expansion, and applied it to tensor factorization models.

Our contribution. We extend this classification problem to a setting in which the input dimension p, latent dimension n, training-set size m, and initialization scale σ^2 may all vary jointly. Extending the diagrammatic formalism with data edges allows train and population losses to be treated separately. For the weight-tied linear autoencoder, the resulting dominant monomials form a finite hierarchy

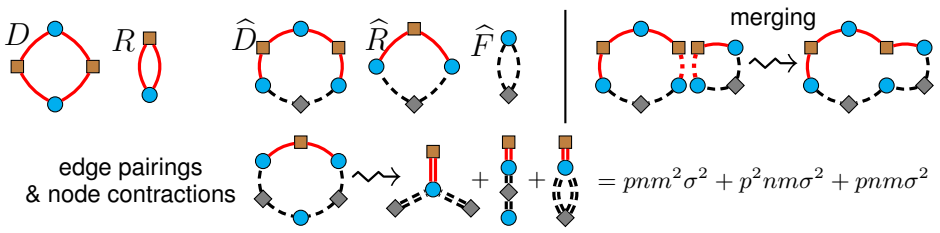


Figure 4: Diagrams used to represent the averaged derivatives $D_{\mathrm{pop},s}$ and $D_{\mathrm{tr},s}$. **Top row: Left:** Base diagrams $D, R, \widehat{D}, \widehat{R}, \widehat{F}$ (Eqs. (23), (24)). Blue circles, brown squares, and grey diamonds denote p-, n-, and m-nodes, respectively. Red and dashed black edges denote $\boldsymbol{U}$, and $\boldsymbol{X}$, respectively. **Right:** Merging two diagrams. For each pair of matched $\boldsymbol{U}$-edges, the diagrams are attached and both edges removed. **Bottom row:** Three possible edge pairings with associated node contractions for the ring diagram $\mathrm{Tr}[U^\top U(XX^\top)^2]$.

naturally represented by a triangular prism. Its five facets give the large-data, small-data, mean-field, narrow-latent, and free basic regimes, while edges and vertices represent combinations of these extremes.

A.4 High-dimensional generalization and spectral criticality

Prior work. High-dimensional learning can exhibit sharp behavior near interpolation thresholds. The double-descent literature highlighted anomalous generalization near the point where model and sample dimensions become comparable (Belkin et al., 2019), while exact analyses of high-dimensional ridgeless regression characterize this phenomenon using random-matrix spectra (Hastie et al., 2022). Closely related dynamically, Advani et al. (2020) showed how the spectrum of the empirical covariance controls train and generalization-error trajectories in high-dimensional linear models.

Our contribution. We find related spectral critical phenomena in an unsupervised reconstruction problem, but with a different mechanism. In the mean-field regime, the empirical covariance gap closes at $p/m = 1$, changing the long-time relaxation and producing different asymptotic rates for train and population losses. In the large-data regime, an analogous slowdown occurs at $p/n = 1$ when the spectrum of the random initial Gram matrix reaches zero. The narrow-latent solution further shows that different spectral edges can govern train and population relaxation, leading to parametrically separated time scales. Thus spectral geometry controls not only final generalization but also the time scales on which optimization and generalization occur.

B Proof of the autoencoder phase diagram (Theorem 3.1)

B.1 Laurent coefficients and diagram expansions

We prove the coefficient assertion of Theorem 3.1 and introduce the diagrammatic notation used in the next subsection. Throughout, $\boldsymbol{X}$ has independent $\mathcal{N}(0,1)$ entries and is independent of the Gaussian initialization $\boldsymbol{U}(0)$; expectations include both sources of randomness. We write $L_{\mathrm{pop}} = L$ and $L_{\mathrm{tr}} = \widehat{L}$.

Lemma B.1. *For every $s \geq 1$, the averaged derivatives $D_{\mathrm{pop},s}$ and $D_{\mathrm{tr},s}$ of Eq. (4) are Laurent polynomials in (p, n, m, σ, η). Every monomial has degree s in η and even degree in σ.*

Proof. Base diagrams. Expanding the losses in traces gives

$$L = \frac{1}{2} + \frac{H_{\mathrm{pop}}}{p}, \qquad \widehat{L} = \frac{\widehat{F}}{2pm} + \frac{T}{pm}, \tag{22}$$

where $H_{\mathrm{pop}} = \frac{1}{2}D - R$, $T = \frac{1}{2}\widehat{D} - \widehat{R}$, and

$$D = \mathrm{Tr}[(\boldsymbol{U}^\top \boldsymbol{U})^2], \qquad R = \mathrm{Tr}[\boldsymbol{U}^\top \boldsymbol{U}], \tag{23}$$

$$\widehat{D} = \mathrm{Tr}[(\boldsymbol{U}^\top \boldsymbol{U})^2 \boldsymbol{X}\boldsymbol{X}^\top], \qquad \widehat{R} = \mathrm{Tr}[\boldsymbol{U}^\top \boldsymbol{U}\boldsymbol{X}\boldsymbol{X}^\top], \qquad \widehat{F} = \mathrm{Tr}[\boldsymbol{X}\boldsymbol{X}^\top]. \tag{24}$$

A *diagram* records the summation indices in such a trace: vertices have types p, n, or m, and an edge labeled $\boldsymbol{U}$ or $\boldsymbol{X}$ represents an entry U_{ai} or $X_{i\mu}$, respectively. Thus $\boldsymbol{U}$-edges join n- to p-vertices and $\boldsymbol{X}$-edges join p- to m-vertices. For example,

$$\widehat{R} = \sum_{a=1}^{n} \sum_{i,j=1}^{p} \sum_{\mu=1}^{m} U_{ai} U_{aj} X_{j\mu} X_{i\mu}. \tag{25}$$

Its diagram is a four-edge ring with successive vertex types p, n, p, m. The other traces in Eqs. (23) and (24) also give rings, possibly with more vertices or with no m-vertices. The value of a diagram is the sum, over all assignments to its vertices, of the product of its edge entries. These five base diagrams and the operations below are illustrated in Fig. 4.

Merging and differentiation. For functions represented by diagrams, write $G_1 \star G_2 = \langle \nabla_{\boldsymbol{U}} G_1, \nabla_{\boldsymbol{U}} G_2 \rangle$. To evaluate this product graphically, choose one $\boldsymbol{U}$-edge in each diagram, remove the two chosen edges, identify their n-endpoints and identify their p-endpoints, then sum the resulting diagrams over all choices of edges. Each choice counts with its multiplicity. A merger of two rings is again a ring; the operation is bilinear, but need not be associative.

Since $\widehat{F}$ is independent of $\boldsymbol{U}$, gradient flow yields

$$\frac{d^s L_\ell}{dt^s} = -\eta \left\langle \nabla_{\boldsymbol{U}} \frac{d^{s-1} L_\ell}{dt^{s-1}}, \nabla_{\boldsymbol{U}} \widehat{L} \right\rangle, \qquad s \geq 1. \tag{26}$$

Set $G_{\mathrm{pop}}^{[0]} = H_{\mathrm{pop}}$, $G_{\mathrm{tr}}^{[0]} = T$, and recursively $G_\ell^{[s]} = G_\ell^{[s-1]} \star T$. Thus the order of the successive mergers is fixed. Eqs. (22) and (26) give

$$D_{\mathrm{pop},s} = \frac{(-\eta)^s}{p^{s+1} m^s} \mathbb{E}\big[G_{\mathrm{pop}}^{[s]}(\boldsymbol{U}(0), \boldsymbol{X})\big], \tag{27}$$

$$D_{\mathrm{tr},s} = \frac{(-\eta)^s}{p^{s+1} m^{s+1}} \mathbb{E}\big[G_{\mathrm{tr}}^{[s]}(\boldsymbol{U}(0), \boldsymbol{X})\big]. \tag{28}$$

For later use, we note that the edge types can be arranged in two contiguous arcs. Each of the four rings $D, R, \widehat{D}, \widehat{R}$ has a cyclic sequence of edge labels of the form $\mathsf{U}^{2h}\mathsf{X}^{2k}$, where U and X denote edge types, $k = 0$ for D, R, and $k = 1$ for $\widehat{D}, \widehat{R}$. If two rings have this form, merge $\boldsymbol{U}$-edges at opposite ends of their $\boldsymbol{X}$-arcs (choosing any $\boldsymbol{U}$-edge when an $\boldsymbol{X}$-arc is empty). The resulting ring again has a contiguous $\boldsymbol{U}$-arc and a contiguous $\boldsymbol{X}$-arc: the selected $\boldsymbol{U}$-edges disappear and the two $\boldsymbol{X}$-arcs join around the cyclic boundary. Induction gives such a two-arc ring for any selected sequence of base factors. This specifies the merger choices used to establish realizability in Appendix B.2.

Wick pairing. To average a diagram, pair its $\boldsymbol{U}$-edges among themselves and its $\boldsymbol{X}$-edges among themselves. A paired pair of edges forces their corresponding endpoints to have equal indices; it contributes σ^2 for a pair of $\boldsymbol{U}$-edges and 1 for a pair of $\boldsymbol{X}$-edges. Each pair becomes one edge in the contracted graph. If N_p, N_n, N_m vertices of the respective types remain and the original diagram has N_U $\boldsymbol{U}$-edges, the pairing contributes

$$p^{N_p} n^{N_n} m^{N_m} \sigma^{N_U}. \tag{29}$$

For example, the three pairings of the ring representing $\mathrm{Tr}[\boldsymbol{U}^\top \boldsymbol{U}(\boldsymbol{X}\boldsymbol{X}^\top)^2]$ are shown in Fig. 4. The contracted graph remains connected, since it is obtained by identifying vertices of the original connected ring.

There are finitely many mergers and pairings at each fixed s. Every resulting term has the form Eq. (29), with an even N_U; the denominators in Eqs. (27) and (28) introduce only integer negative powers of p, m. This proves the claimed Laurent-polynomial form and the common factor η^s.

Finally, averaging may be interchanged with each fixed-order derivative at $t = 0$. Indeed, with $\boldsymbol{A} = \boldsymbol{U}^\top \boldsymbol{U}$ and $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^\top/m$, the flow satisfies

$$\frac{d}{dt}\|\boldsymbol{U}(t)\|_F^2 = \frac{4\eta}{p}\,\mathrm{Tr}[\boldsymbol{S}(\boldsymbol{A} - \boldsymbol{A}^2)] \le \frac{\eta}{p}\,\mathrm{Tr}\,\boldsymbol{S},$$

because $\boldsymbol{A} - \boldsymbol{A}^2 \preceq \boldsymbol{I}/4$. On a bounded positive time interval, the weight norm and every fixed-order loss derivative therefore admit polynomial bounds in the Gaussian initial variables. Their finite moments justify differentiation under the expectation. □

B.2 Monomials that can be leading

Let $Q_{\ell,s}$ be the exponent set of $D_{\ell,s}$ in Eq. (4). Its five coordinates correspond, in order, to p, n, m, σ, η. We say that $\boldsymbol{q}' \in Q_{\ell,s}$ *dominates* $\boldsymbol{q} \in Q_{\ell,s}$ if

$$q'_\sigma = q_\sigma, \qquad q'_\eta = q_\eta, \qquad q'_j \ge q_j \quad (j \in \{p, n, m\}),$$

with at least one strict inequality. An exponent is *Pareto-optimal* if no other exponent in $Q_{\ell,s}$ dominates it; the set of all such exponents is the *Pareto set* $P_{\ell,s}$. We compare only equal powers of σ, since α_σ may have either sign; the powers of η are automatically equal at each order s. Because $\alpha_p, \alpha_n, \alpha_m > 0$, dominance implies $\boldsymbol{q}' \cdot \boldsymbol{\alpha} > \boldsymbol{q} \cdot \boldsymbol{\alpha}$. Hence every leading exponent belongs to $P_{\ell,s}$.

Proposition B.2 (Leading monomials)**.** *For every $s \ge 1$, the train and population Pareto sets consist exactly of the vectors*

$$\boldsymbol{q}_s(u, v, r) = \big(u - r,\ v + 1,\ r - s,\ 2(u + v + 1),\ s\big), \qquad u, v \ge 0, \quad u + v \le s + 1, \tag{30}$$

where $r \in \{0, \dots, s\}$ for the train loss and $r \in \{1, \dots, s\}$ for the population loss. Every listed vector occurs with a nonzero coefficient and belongs to the leading support for some admissible scaling vector.

Proof. Consider the expansions Eqs. (27) and (28), and let s_D count the factors of type D or $\widehat{D}$ among the $s+1$ factors in a term. Put $\delta_{\mathrm{tr}} = 1$ and $\delta_{\mathrm{pop}} = 0$. Any resulting ring has $s_D + 1$ n-vertices, $s + \delta_\ell$ m-vertices, $2(s_D + 1)$ $\boldsymbol{U}$-edges, and $2(s + \delta_\ell)$ $\boldsymbol{X}$-edges. These counts hold regardless of the merger choices.

For a Wick pairing, denote by N_p, N_n, N_m the numbers of vertices after contraction. Set

$$v = N_n - 1, \qquad u = s_D - v, \qquad r = N_m - \delta_\ell.$$

The vertex counts give $u, v \ge 0$, $u + v = s_D \le s + 1$, and the stated ranges of r. The contracted graph is connected and has $E = (s_D + 1) + (s + \delta_\ell)$ edges, one for each paired pair. Thus it has at most $E + 1$ vertices, and

$$N_p \le E + 1 - N_n - N_m = s + u - r + 1. \tag{31}$$

By Eqs. (27) and (28), the exponent of this term is

$$\boldsymbol{q} = \big(N_p - s - 1,\ v + 1,\ r - s,\ 2(u + v + 1),\ s\big).$$

It is dominated by $\boldsymbol{q}_s(u, v, r)$ if that vector is realized.

We now construct a pairing with equality in Eq. (31). Use the two-arc merger choices described in Appendix B.1. In the resulting ring, the $\boldsymbol{X}$-arc contains $s + \delta_\ell$ m-vertices and $s + \delta_\ell - 1$ internal p-vertices. Pair the two $\boldsymbol{X}$-edges incident to each of $s - r$ selected internal p-vertices. Each such pairing identifies neighboring m-vertices and leaves its p-vertex as a petal. The remaining m-vertices form $r + \delta_\ell$ groups, each with two unpaired $\boldsymbol{X}$-edges at its ends.

Likewise, the $\boldsymbol{U}$-arc has $s_D + 1$ n-vertices and s_D internal p-vertices. Pair both $\boldsymbol{U}$-edges at each of $u = s_D - v$ selected internal p-vertices. This leaves $v + 1$ groups of n-vertices and u further petals. Finally, pair the two unpaired edges incident to each remaining m- or n-group. These pairings preserve the groups of m- and n-vertices and identify all p-vertices of the residual ring into one. The contracted diagram consequently has

$$N_m = r + \delta_\ell, \qquad N_n = v + 1, \qquad N_p = (s - r) + u + 1 = s + u - r + 1,$$

as required.

The coefficient of this monomial cannot vanish by cancellation. For fixed s_D, every merger term has the same sign, since it contains exactly $s + 1 - s_D$ factors of type R or $\widehat{R}$; the Wick weights and merger multiplicities are positive. Its σ-exponent $2(s_D + 1)$ fixes s_D, so terms with opposite signs cannot contribute to the same exponent vector. Hence all listed vectors occur.

Every unlisted exponent is dominated by a listed one, as shown above. Listed vectors with distinct $s_D = u + v$ have different σ-exponents; at fixed s_D, they satisfy $q_p + q_n + q_m = s_D + 1 - s$ and are therefore mutually non-dominating. This proves the claimed Pareto sets.

To see that all listed vectors can actually lead, choose $\alpha_p = \alpha_n = \alpha_m = c > 0$ and $\alpha_\sigma = -c/2$, with arbitrary α_η. Every listed vector then has scaling exponent $-sc + s\alpha_\eta$, and every dominated vector has a strictly smaller exponent. Thus all listed vectors belong to the leading support at this balanced scaling. □

B.3 The normal fan and persistent transitions

Proof of Theorem 3.1. The Laurent-polynomial assertion follows from Lemma B.1. By Proposition B.2, every exponent outside the Pareto sets is strictly dominated for admissible $\boldsymbol{\alpha}$. It therefore cannot tie for the maximum, and we may determine the leading supports using only the vectors $\boldsymbol{q}_s(u, v, r)$.

Recall that $\alpha_\phi = \alpha_p - \alpha_m$, $\alpha_\rho = \alpha_n + 2\alpha_\sigma$, and $\alpha_\psi = \alpha_p - \alpha_n$. Substituting Eq. (30) gives

$$\begin{aligned} \boldsymbol{q}_s(u, v, r) \cdot \boldsymbol{\alpha} &= \alpha_\rho - s\alpha_m + s\alpha_\eta + u(\alpha_\rho + \alpha_\psi) + v\alpha_\rho - r\alpha_\phi \\ &= \alpha_\rho - s\alpha_m + s\alpha_\eta + \alpha_\rho x_1 + \alpha_\psi x_2 - \alpha_\phi x_3, \end{aligned} \tag{32}$$

where $\mathbf{x} = (x_1, x_2, x_3) = (u + v, u, r)$. The first three terms on the right are common to all monomials at order s. Thus the leading supports depend only on $\boldsymbol{\gamma} = (\alpha_\phi, \alpha_\rho, \alpha_\psi)$; in particular, α_η does not affect them.

In the $\mathbf{x}$ coordinates, the convex hulls of the train and population Pareto sets are

$$\Pi_{\mathrm{tr},s} = \big\{\mathbf{x} : 0 \le x_2 \le x_1 \le s + 1, \quad 0 \le x_3 \le s\big\}, \tag{33}$$

$$\Pi_{\mathrm{pop},s} = \big\{\mathbf{x} : 0 \le x_2 \le x_1 \le s + 1, \quad 1 \le x_3 \le s\big\}. \tag{34}$$

The Pareto monomials correspond to all lattice points in these polytopes. By Eq. (32), their leading supports are the lattice points on the faces exposed by the linear score $\alpha_\rho x_1 + \alpha_\psi x_2 - \alpha_\phi x_3$.

For $s \ge 2$, both polytopes are triangular prisms with the same normal fan. At $s = 1$, the train polytope is still a prism, whereas the population polytope has $x_3 = 1$ throughout. Its support depends only on α_ρ, α_ψ and introduces no additional distinctions beyond those made by the train support. Consequently, the joint formal phases are the cells of the common prism normal fan.

The five prism facets are $x_3 = \max x_3$, $x_3 = \min x_3$, $x_2 = 0$, $x_1 = x_2$, and $x_1 = s + 1$. Their outward normal rays, expressed in $\boldsymbol{\gamma}$ coordinates, are respectively the "Large-data", "Small-data", "Mean-field", "Narrow-latent", and "Free" rays in Eq. (7). For example, the outward score normal to $x_1 = x_2$ is $-x_1 + x_2$, which gives $\alpha_\rho < 0$ and $\alpha_\psi = -\alpha_\rho$.

Every direction $\boldsymbol{\gamma}$ is attainable with $\alpha_p, \alpha_n, \alpha_m > 0$. Indeed, given $\boldsymbol{\gamma}$, choose M sufficiently large and set

$$\alpha_p = M, \qquad \alpha_m = M - \alpha_\phi, \qquad \alpha_n = M - \alpha_\psi, \qquad \alpha_\sigma = \frac{\alpha_\rho - \alpha_n}{2}.$$

Hence the admissibility conditions do not remove any cell of the normal fan.

To obtain the dual description, translate either full-dimensional prism so that the origin lies in its interior. Its polar, formed using the dual score coordinates $\boldsymbol{\lambda} = (\alpha_\rho, \alpha_\psi, -\alpha_\phi)$, is a triangular bipyramid. Under the invertible change from $\boldsymbol{\lambda}$ to $\boldsymbol{\gamma}$, its five vertices lie on the rays in Eq. (7). Directions in the relative interior of each normal cone expose one fixed prism face; equivalently, their rays meet the relative interior of one fixed face of the bipyramid. This proves the stated phase description, including its five vertices, nine edges, and six triangular faces. At $\boldsymbol{\gamma} = \mathbf{0}$, every Pareto monomial is leading, giving the balanced phase.

Finally, distinct cells of this normal fan expose different faces of the train prism at every $s \geq 2$. Their train leading supports therefore differ at infinitely many orders. In particular, every transition between adjacent formal phases is persistent. This conclusion concerns supports of formal coefficients and does not assume that tied monomials have a nonzero sum after particular scaling constants are substituted. □

C SOLUTIONS OF EXTREME REGIMES (SECTION 4)

C.1 COMMON NOTATION, STABILITY PRINCIPLES, AND PRELIMINARY RESULTS

This subsection fixes notation used throughout Appendix C and isolates the finite-dimensional facts shared by the four solved regimes. The common object is the empirical covariance $\boldsymbol{S}$, while the state of the model can be represented either by the input-space Gram matrix $\boldsymbol{A} = \boldsymbol{U}^\top \boldsymbol{U}$ or, in the narrow-latent regime, by the latent-space Gramians built from $\boldsymbol{U}$. The proofs below separate the deterministic stability of these matrix ODEs from the random-matrix statements used to identify the relevant proxy trajectories.

Throughout Appendix C, $N \in \mathbb{N}$ denotes a generic asymptotic sequence index. Thus the dimensions and parameters are sequences

$$p = p_N, \qquad n = n_N, \qquad m = m_N, \qquad \sigma = \sigma_N, \tag{35}$$

and the subscript N is suppressed whenever this causes no ambiguity. In the mean-field and large-data subsections we sometimes use p itself as the sequence index, which is only a notational shorthand for the same asymptotic convention.

All four regimes use the same Gaussian weight initialization, independent of the training matrix:

$$\boldsymbol{U}_N(0) = \sigma_N \boldsymbol{Z}_N, \qquad (\boldsymbol{Z}_N)_{\alpha i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \qquad \boldsymbol{Z}_N \text{ independent of } \boldsymbol{X}_N. \tag{36}$$

The regime-specific scalings below specify the relevant rescaled variance, for example $n_N \sigma_N^2$ or $p_N \sigma_N^2$, rather than redefining the initialization.

For a training matrix $\boldsymbol{X} \in \mathbb{R}^{p \times m}$, weight matrix $\boldsymbol{U} \in \mathbb{R}^{n \times p}$, and population covariance $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$, define

$$\boldsymbol{S} := \frac{\boldsymbol{X}\boldsymbol{X}^\top}{m}, \qquad \boldsymbol{A} := \boldsymbol{U}^\top \boldsymbol{U}, \qquad \boldsymbol{E} := \boldsymbol{I}_p - \boldsymbol{A}, \qquad \gamma := \frac{\eta}{p}. \tag{37}$$

The empirical and population losses are

$$\widehat{L}(\boldsymbol{U}) = \frac{1}{2p} \operatorname{Tr}(\boldsymbol{E}\boldsymbol{S}\boldsymbol{E}), \qquad L(\boldsymbol{U}) = \frac{1}{2p} \operatorname{Tr}(\boldsymbol{E}\boldsymbol{\Sigma}\boldsymbol{E}). \tag{38}$$

In the isotropic case $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, this reduces to

$$L(\boldsymbol{U}) = \frac{1}{2p} \operatorname{Tr}(\boldsymbol{E}^2). \tag{39}$$

Common proof architecture. All four solved regimes are perturbative stability arguments around a proxy that becomes exactly solvable in the corresponding extreme limit. There are two recurring mechanisms. In the mean-field regime, the perturbation is on the state side: $\boldsymbol{A}(0)$ is close to a scalar matrix, and stability with respect to the initial Gram matrix transfers the explicit commuting solution to the true trajectory. The narrow-latent regime uses the same state-side idea after passing to latent-space Gramians: a scalarized proxy is self-consistent because weighted Gaussian Gramians concentrate, and a finite-time stability estimate transfers the proxy to the exact flow. In the large-data regime, the perturbation is on the data side: $\boldsymbol{S}$ is close to $\boldsymbol{I}_p$, so stability with respect to $\boldsymbol{S}$ reduces the dynamics to the population-covariance flow. In the small-data regime, a rotation to the training span turns the normalized sample Gram matrix $\boldsymbol{\Gamma} = p^{-1}\boldsymbol{X}^\top \boldsymbol{X}$ into the data-side parameter; the exactly orthogonal case $\boldsymbol{\Gamma} = \boldsymbol{I}_m$ produces a closed hierarchy of matrix sequences $\mathbf{Q}$, and stability with respect to $\boldsymbol{\Gamma}$ transfers that hierarchy to near-isometric data.

Thus each proof has the same three layers: identify an exactly tractable proxy, prove that either the state or the data geometry is close to the proxy in operator norm, and propagate that closeness on every fixed time interval by a dimension-free stability estimate. The random-matrix input is regime-specific; the deterministic ODE facts are collected next.

Standing spectral assumptions on the data. The four proofs use the same basic spectral regularity principle, after the normalization appropriate to the regime. For a sequence of positive semidefinite data operators $\boldsymbol{S}_N \in \mathbb{R}^{p_N \times p_N}$, write

$$\mu_{\boldsymbol{S}_N} := \frac{1}{p_N} \sum_{i=1}^{p_N} \delta_{\lambda_i(\boldsymbol{S}_N)}. \tag{40}$$

The two standing spectral hypotheses are that, for some deterministic $K < \infty$ and deterministic probability measure ν_ϕ,

$$\mathbb{P}(\|\boldsymbol{S}_N\|_{\mathrm{op}} \le K) \longrightarrow 1, \tag{41}$$

and

$$\mu_{\boldsymbol{S}_N} \Longrightarrow \nu_\phi \qquad \text{in probability,} \tag{42}$$

where the second convergence is weak convergence of probability measures. Together, equation 41–equation 42 imply

$$\operatorname{supp} \nu_\phi \subset [0, K]. \tag{43}$$

The operator-norm bound is the input needed for dimension-free finite-time ODE estimates; the spectral limit is needed only when a random finite-size spectral functional is replaced by a deterministic limiting formula.

In the mean-field and narrow-latent regimes these assumptions are imposed directly on the empirical covariance

$$\boldsymbol{S}_N = \frac{1}{m_N} \boldsymbol{X}_N \boldsymbol{X}_N^\top. \tag{44}$$

In the large-data regime we assume the stronger near-isotropy condition

$$\|\boldsymbol{S}_N - \boldsymbol{I}_{p_N}\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{45}$$

This automatically implies equation 41 and equation 42, with

$$\nu_\phi = \delta_1. \tag{46}$$

The small-data regime requires one normalization caveat. There $\phi_N = p_N/m_N \to \infty$, so the unscaled empirical covariance $\boldsymbol{S}_N = \boldsymbol{X}_N \boldsymbol{X}_N^\top / m_N$ has operator norm of order ϕ_N and is not bounded. The dynamically relevant normalized data operator is

$$\widetilde{\boldsymbol{S}}_N := \frac{1}{\phi_N} \boldsymbol{S}_N = \frac{1}{p_N} \boldsymbol{X}_N \boldsymbol{X}_N^\top, \qquad \boldsymbol{\Gamma}_N := \frac{1}{p_N} \boldsymbol{X}_N^\top \boldsymbol{X}_N. \tag{47}$$

The nonzero eigenvalues of $\widetilde{\boldsymbol{S}}_N$ and $\boldsymbol{\Gamma}_N$ coincide. The small-data assumption is the active-space near-isometry

$$\|\boldsymbol{\Gamma}_N - \boldsymbol{I}_{m_N}\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{48}$$

Thus the active spectral law converges to δ_1 and $\|\widetilde{\boldsymbol{S}}_N\|_{\mathrm{op}} \to 1$ in probability. If one instead uses the ambient $p_N \times p_N$ empirical spectral distribution of $\widetilde{\boldsymbol{S}}_N$, its limit is δ_0, because $m_N/p_N \to 0$ and only m_N eigenvalues can be nonzero. Thus the spectral-limit convention is: a general ν_ϕ in mean-field and narrow-latent, $\nu_\phi = \delta_1$ in large-data, and the automatically determined active law δ_1 (or ambient law δ_0) in small-data. Equivalently, all four regimes use a bounded normalized data operator and a limiting spectral law; only the normalization differs in the small-data corner.

Proposition C.1 (Common gradient-flow identities)**.** *For every fixed symmetric $\boldsymbol{S} \succeq 0$, the gradient flow of the empirical loss in equation 38 is*

$$\dot{\boldsymbol{U}} = \gamma \boldsymbol{U} \left[\boldsymbol{S}(\boldsymbol{I}_p - \boldsymbol{A}) + (\boldsymbol{I}_p - \boldsymbol{A})\boldsymbol{S} \right]. \tag{49}$$

Equivalently, if

$$\mathcal{G}_{\boldsymbol{S}}(\boldsymbol{A}) := \boldsymbol{S}(\boldsymbol{I}_p - \boldsymbol{A}) + (\boldsymbol{I}_p - \boldsymbol{A})\boldsymbol{S} = 2\boldsymbol{S} - \boldsymbol{S}\boldsymbol{A} - \boldsymbol{A}\boldsymbol{S}, \tag{50}$$

then $\dot{\boldsymbol{U}} = \gamma \boldsymbol{U} \mathcal{G}_{\boldsymbol{S}}(\boldsymbol{A})$.

Proof. Since $d\boldsymbol{E} = -d\boldsymbol{A}$,

$$d\widehat{L} = -\frac{1}{2p}\,\mathrm{Tr}[d\boldsymbol{A}(\boldsymbol{S}\boldsymbol{E} + \boldsymbol{E}\boldsymbol{S})]\,. \tag{51}$$

Using

$$d\boldsymbol{A} = d\boldsymbol{U}^\top \boldsymbol{U} + \boldsymbol{U}^\top d\boldsymbol{U} \tag{52}$$

and the symmetry of $\boldsymbol{S}\boldsymbol{E} + \boldsymbol{E}\boldsymbol{S}$ gives

$$\frac{\partial \widehat{L}}{\partial \boldsymbol{U}} = -\frac{1}{p}\boldsymbol{U}(\boldsymbol{S}\boldsymbol{E} + \boldsymbol{E}\boldsymbol{S}). \tag{53}$$

Multiplying by the learning rate η yields equation 49. □

Lemma C.2 (Closed Gram dynamics and a priori bound)**.** *Let $\boldsymbol{U}$ solve equation 49 with $\gamma \geq 0$, and set $\boldsymbol{A} = \boldsymbol{U}^\top \boldsymbol{U}$. Then $\boldsymbol{A}$ satisfies the closed matrix ODE*

$$\dot{\boldsymbol{A}} = \gamma \mathcal{F}_{\boldsymbol{S}}(\boldsymbol{A}), \qquad \mathcal{F}_{\boldsymbol{S}}(\boldsymbol{A}) := 2(\boldsymbol{S}\boldsymbol{A} + \boldsymbol{A}\boldsymbol{S}) - \boldsymbol{S}\boldsymbol{A}^2 - \boldsymbol{A}^2\boldsymbol{S} - 2\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}. \tag{54}$$

If $\boldsymbol{A}(0) \succeq 0$, the solution exists globally, remains positive semidefinite, and obeys

$$0 \preceq \boldsymbol{A}(t), \qquad \|\boldsymbol{A}(t)\|_{\mathrm{op}} = \|\boldsymbol{U}(t)\|_{\mathrm{op}}^2 \leq R_0 := \max\{1, \|\boldsymbol{A}(0)\|_{\mathrm{op}}\}, \qquad t \geq 0. \tag{55}$$

Proof. Set

$$\mathcal{G}_{\boldsymbol{S}}(\boldsymbol{A}) := \boldsymbol{S}(\boldsymbol{I}_p - \boldsymbol{A}) + (\boldsymbol{I}_p - \boldsymbol{A})\boldsymbol{S} = 2\boldsymbol{S} - \boldsymbol{S}\boldsymbol{A} - \boldsymbol{A}\boldsymbol{S}. \tag{56}$$

Since $\boldsymbol{A}$ and $\boldsymbol{S}$ are symmetric, $\mathcal{G}_{\boldsymbol{S}}(\boldsymbol{A})$ is symmetric as well. Differentiating $\boldsymbol{A} = \boldsymbol{U}^\top \boldsymbol{U}$ and using $\dot{\boldsymbol{U}} = \gamma \boldsymbol{U} \mathcal{G}_{\boldsymbol{S}}(\boldsymbol{A})$ gives

$$\begin{aligned} \dot{\boldsymbol{A}} &= \dot{\boldsymbol{U}}^\top \boldsymbol{U} + \boldsymbol{U}^\top \dot{\boldsymbol{U}} \\ &= \gamma\left[\mathcal{G}_{\boldsymbol{S}}(\boldsymbol{A})\boldsymbol{A} + \boldsymbol{A}\mathcal{G}_{\boldsymbol{S}}(\boldsymbol{A})\right]. \end{aligned} \tag{57}$$

Expanding the two products in equation 57 yields

$$\dot{\boldsymbol{A}} = \gamma\big[2(\boldsymbol{S}\boldsymbol{A} + \boldsymbol{A}\boldsymbol{S}) - \boldsymbol{S}\boldsymbol{A}^2 - \boldsymbol{A}^2\boldsymbol{S} - 2\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}\big], \tag{58}$$

which is exactly equation 54.

Positive semidefiniteness is automatic along the $\boldsymbol{U}$-flow, since

$$\boldsymbol{A}(t) = \boldsymbol{U}(t)^\top \boldsymbol{U}(t) \succeq 0. \tag{59}$$

It remains to prove the dimension-free upper bound. Define

$$\alpha(t) := \lambda_{\max}(\boldsymbol{A}(t)). \tag{60}$$

For a differentiable curve of real symmetric matrices, the upper right Dini derivative of the largest eigenvalue is

$$D^+\alpha(t) = \max_{\substack{\boldsymbol{v} \in E_{\max}(\boldsymbol{A}(t)) \\ \|\boldsymbol{v}\|_2 = 1}} \boldsymbol{v}^\top \dot{\boldsymbol{A}}(t)\boldsymbol{v}, \tag{61}$$

where $E_{\max}(\boldsymbol{A})$ denotes the top eigenspace of $\boldsymbol{A}$. For completeness, we verify this formula. Let $h_j \downarrow 0$ be a sequence realizing the upper right limsup, and let $\boldsymbol{v}_j$ be a unit top eigenvector of $\boldsymbol{A}(t+h_j)$. Passing to a subsequence if necessary, compactness of the unit sphere gives

$$\boldsymbol{v}_j \longrightarrow \boldsymbol{v}, \qquad \boldsymbol{v} \in E_{\max}(\boldsymbol{A}(t)). \tag{62}$$

Since

$$\boldsymbol{v}_j^\top \boldsymbol{A}(t)\boldsymbol{v}_j \leq \alpha(t), \tag{63}$$

we have

$$\begin{aligned} \limsup_{h \downarrow 0} \frac{\alpha(t+h) - \alpha(t)}{h} &\leq \lim_{j \to \infty} \boldsymbol{v}_j^\top \frac{\boldsymbol{A}(t+h_j) - \boldsymbol{A}(t)}{h_j} \boldsymbol{v}_j \\ &= \boldsymbol{v}^\top \dot{\boldsymbol{A}}(t)\boldsymbol{v}. \end{aligned} \tag{64}$$

Conversely, for any fixed unit $\boldsymbol{v} \in E_{\max}(\boldsymbol{A}(t))$, the variational characterization of the largest eigenvalue gives

$$\alpha(t+h) - \alpha(t) \geq \boldsymbol{v}^\top \big(\boldsymbol{A}(t+h) - \boldsymbol{A}(t)\big)\boldsymbol{v}. \tag{65}$$

Therefore

$$\liminf_{h\downarrow 0} \frac{\alpha(t+h)-\alpha(t)}{h} \geq \boldsymbol{v}^\top \dot{\boldsymbol{A}}(t)\boldsymbol{v}. \tag{66}$$

Taking the maximum over the top eigenspace proves equation 61.

Now fix a unit $\boldsymbol{v} \in E_{\max}(\boldsymbol{A}(t))$, so that

$$\boldsymbol{A}(t)\boldsymbol{v} = \alpha(t)\boldsymbol{v}. \tag{67}$$

Using equation 57 and the symmetry of $\mathcal{G}_{\boldsymbol{S}}(\boldsymbol{A})$,

$$\boldsymbol{v}^\top \dot{\boldsymbol{A}}\boldsymbol{v} = 2\gamma\alpha(t)\, \boldsymbol{v}^\top \mathcal{G}_{\boldsymbol{S}}(\boldsymbol{A})\boldsymbol{v}. \tag{68}$$

Moreover,

$$\begin{aligned} \boldsymbol{v}^\top \mathcal{G}_{\boldsymbol{S}}(\boldsymbol{A})\boldsymbol{v} &= \boldsymbol{v}^\top \boldsymbol{S}(\boldsymbol{I}_p - \boldsymbol{A})\boldsymbol{v} + \boldsymbol{v}^\top(\boldsymbol{I}_p - \boldsymbol{A})\boldsymbol{S}\boldsymbol{v} \\ &= 2(1-\alpha(t))\boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{v}. \end{aligned} \tag{69}$$

Combining equation 61, equation 68, and equation 69 yields

$$D^+\alpha(t) = \max_{\substack{\boldsymbol{v}\in E_{\max}(\boldsymbol{A}(t)) \\ \|\boldsymbol{v}\|_2=1}} 4\gamma\alpha(t)(1-\alpha(t))\boldsymbol{v}^\top \boldsymbol{S}\boldsymbol{v}. \tag{70}$$

Because $\gamma \geq 0$ and $\boldsymbol{S} \succeq 0$, the right-hand side is nonpositive whenever $\alpha(t) \geq 1$. The Dini-derivative comparison principle therefore prevents α from crossing upward through the level $\max\{1, \alpha(0)\}$, and hence

$$\alpha(t) \leq \max\{1, \alpha(0)\} \qquad \text{for every } t \geq 0. \tag{71}$$

Since $\boldsymbol{A}(t) \succeq 0$, its operator norm equals its largest eigenvalue, so

$$\|\boldsymbol{A}(t)\|_{\mathrm{op}} \leq R_0 := \max\{1, \|\boldsymbol{A}(0)\|_{\mathrm{op}}\}. \tag{72}$$

Finally,

$$\|\boldsymbol{U}(t)\|_{\mathrm{op}}^2 = \|\boldsymbol{U}(t)^\top \boldsymbol{U}(t)\|_{\mathrm{op}} = \|\boldsymbol{A}(t)\|_{\mathrm{op}} \leq R_0. \tag{73}$$

In every fixed finite dimension the vector field in equation 49 is polynomial, hence locally Lipschitz, in the entries of $\boldsymbol{U}$. A finite maximal existence time could therefore occur only through blow-up of $\|\boldsymbol{U}(t)\|_{\mathrm{op}}$, which is excluded by equation 73. Thus the solution exists globally, and equation 55 follows. □

Lemma C.3 (Finite-time stability on bounded sets)**.** *Let Y and $\overline{Y}$ be trajectories in a finite-dimensional normed space satisfying*

$$\dot{Y} = F(Y;\theta) + R(t), \qquad \dot{\overline{Y}} = F(\overline{Y};\overline{\theta}), \tag{74}$$

and suppose both trajectories stay in a set on which

$$\|F(Y;\theta) - F(\widetilde{Y};\overline{\theta})\| \leq L\|Y - \widetilde{Y}\| + M\|\theta - \overline{\theta}\|. \tag{75}$$

Then, for every $t \geq 0$,

$$\|Y(t) - \overline{Y}(t)\| \leq e^{Lt}\left[\|Y(0) - \overline{Y}(0)\| + \int_0^t \left(M\|\theta - \overline{\theta}\| + \|R(s)\|\right) ds\right]. \tag{76}$$

Proof. Subtract the two integral equations and use equation 75. Gronwall's inequality gives equation 76. □

Lemma C.4 (Stability with respect to the Gram state)**.** *Let $\boldsymbol{A}(t)$ and $\boldsymbol{B}(t)$ solve*

$$\dot{\boldsymbol{A}} = \gamma\mathcal{F}_{\boldsymbol{S}}(\boldsymbol{A}), \qquad \dot{\boldsymbol{B}} = \gamma\mathcal{F}_{\boldsymbol{S}}(\boldsymbol{B}), \tag{77}$$

with positive semidefinite initial conditions. Set

$$K_{\boldsymbol{S}} := \|\boldsymbol{S}\|_{\mathrm{op}}, \qquad R := \max\{1, \|\boldsymbol{A}(0)\|_{\mathrm{op}}, \|\boldsymbol{B}(0)\|_{\mathrm{op}}\}. \tag{78}$$

Then

$$\|\boldsymbol{A}(t) - \boldsymbol{B}(t)\|_{\mathrm{op}} \leq \exp(\gamma K_{\boldsymbol{S}}(4+8R)t)\, \|\boldsymbol{A}(0) - \boldsymbol{B}(0)\|_{\mathrm{op}}. \tag{79}$$

Proof. Lemma C.2 bounds both trajectories by R. If $\boldsymbol{D} = \boldsymbol{A} - \boldsymbol{B}$, then

$$\|\mathcal{F}_{\boldsymbol{S}}(\boldsymbol{A}) - \mathcal{F}_{\boldsymbol{S}}(\boldsymbol{B})\|_{\mathrm{op}} \le K_{\boldsymbol{S}}(4 + 8R)\|\boldsymbol{D}\|_{\mathrm{op}}. \tag{80}$$

Indeed,

$$\boldsymbol{A}^2 - \boldsymbol{B}^2 = \boldsymbol{A}\boldsymbol{D} + \boldsymbol{D}\boldsymbol{B}, \qquad \boldsymbol{A}\boldsymbol{S}\boldsymbol{A} - \boldsymbol{B}\boldsymbol{S}\boldsymbol{B} = \boldsymbol{D}\boldsymbol{S}\boldsymbol{A} + \boldsymbol{B}\boldsymbol{S}\boldsymbol{D}, \tag{81}$$

and the bound follows by submultiplicativity. Applying Lemma C.3 with no parameter perturbation and no residual gives equation 79. □

Proposition C.5 (Stability with respect to the data covariance)**.** *Let $\boldsymbol{S}, \boldsymbol{T} \succeq 0$, let $\boldsymbol{A}_{\boldsymbol{S}}$ and $\boldsymbol{A}_{\boldsymbol{T}}$ solve*

$$\dot{\boldsymbol{A}}_{\boldsymbol{S}} = \gamma\mathcal{F}_{\boldsymbol{S}}(\boldsymbol{A}_{\boldsymbol{S}}), \qquad \dot{\boldsymbol{A}}_{\boldsymbol{T}} = \gamma\mathcal{F}_{\boldsymbol{T}}(\boldsymbol{A}_{\boldsymbol{T}}), \qquad \boldsymbol{A}_{\boldsymbol{S}}(0) = \boldsymbol{A}_{\boldsymbol{T}}(0) = \boldsymbol{A}_0 \succeq 0. \tag{82}$$

Set

$$R := \max\{1, \|\boldsymbol{A}_0\|_{\mathrm{op}}\}, \qquad K := \max\{\|\boldsymbol{S}\|_{\mathrm{op}}, \|\boldsymbol{T}\|_{\mathrm{op}}\}. \tag{83}$$

Then, for every $T_0 < \infty$,

$$\sup_{0\le t\le T_0} \|\boldsymbol{A}_{\boldsymbol{S}}(t) - \boldsymbol{A}_{\boldsymbol{T}}(t)\|_{\mathrm{op}} \le C_{\gamma,T_0,K,R}\|\boldsymbol{S} - \boldsymbol{T}\|_{\mathrm{op}}, \tag{84}$$

where

$$C_{\gamma,T_0,K,R} := (4R + 4R^2)\frac{e^{\gamma K(4+8R)T_0} - 1}{K(4 + 8R)}, \tag{85}$$

with the continuous value $\gamma T_0(4R + 4R^2)$ when $K = 0$. In particular, if $\gamma = 1$, $\boldsymbol{T} = \boldsymbol{I}_p$, and $\|\boldsymbol{S} - \boldsymbol{I}_p\|_{\mathrm{op}} \le 1$, then

$$\sup_{0\le t\le T_0} \|\boldsymbol{A}_{\boldsymbol{S}}(t) - \boldsymbol{A}_{\boldsymbol{I}}(t)\|_{\mathrm{op}} \le K_{T_0,R}\|\boldsymbol{S} - \boldsymbol{I}_p\|_{\mathrm{op}}, \tag{86}$$

where one may take

$$K_{T_0,R} := (4R + 4R^2)\frac{e^{(8+16R)T_0} - 1}{8 + 16R}. \tag{87}$$

Proof. On the ball $\|\boldsymbol{A}\|_{\mathrm{op}} \le R$,

$$\|\mathcal{F}_{\boldsymbol{S}}(\boldsymbol{A}) - \mathcal{F}_{\boldsymbol{S}}(\boldsymbol{B})\|_{\mathrm{op}} \le K(4 + 8R)\|\boldsymbol{A} - \boldsymbol{B}\|_{\mathrm{op}}, \tag{88}$$

while direct subtraction of the two covariance parameters gives

$$\|\mathcal{F}_{\boldsymbol{S}}(\boldsymbol{A}) - \mathcal{F}_{\boldsymbol{T}}(\boldsymbol{A})\|_{\mathrm{op}} \le (4R + 4R^2)\|\boldsymbol{S} - \boldsymbol{T}\|_{\mathrm{op}}. \tag{89}$$

Lemma C.2 keeps both trajectories in this ball. Lemma C.3, applied with parameter $\theta = \boldsymbol{S}$, yields equation 84; the final specialization uses $K \le 2$. □

Lemma C.6 (Stability of population and train losses)**.** *Let $\boldsymbol{A}, \boldsymbol{B} \succeq 0$ with*

$$\|\boldsymbol{A}\|_{\mathrm{op}} \le R, \qquad \|\boldsymbol{B}\|_{\mathrm{op}} \le R. \tag{90}$$

For isotropic population covariance, define

$$L_p(\boldsymbol{A}) := \frac{1}{2p}\operatorname{Tr}(\boldsymbol{I}_p - \boldsymbol{A})^2, \qquad \widehat{L}_{p,\boldsymbol{S}}(\boldsymbol{A}) := \frac{1}{2p}\operatorname{Tr}\left[\boldsymbol{S}(\boldsymbol{I}_p - \boldsymbol{A})^2\right]. \tag{91}$$

Then

$$|L_p(\boldsymbol{A}) - L_p(\boldsymbol{B})| \le (1 + R)\|\boldsymbol{A} - \boldsymbol{B}\|_{\mathrm{op}}, \tag{92}$$

and, for $\boldsymbol{S} \succeq 0$,

$$|\widehat{L}_{p,\boldsymbol{S}}(\boldsymbol{A}) - \widehat{L}_{p,\boldsymbol{S}}(\boldsymbol{B})| \le \|\boldsymbol{S}\|_{\mathrm{op}}(1 + R)\|\boldsymbol{A} - \boldsymbol{B}\|_{\mathrm{op}}. \tag{93}$$

Moreover,

$$|\widehat{L}_{p,\boldsymbol{S}}(\boldsymbol{A}) - L_p(\boldsymbol{A})| \le \frac{1}{2}(1 + R)^2\|\boldsymbol{S} - \boldsymbol{I}_p\|_{\mathrm{op}}. \tag{94}$$

Consequently, under the specialization in Proposition C.5,

$$\sup_{0\le t\le T_0} |L_p(\boldsymbol{A}_{\boldsymbol{S}}(t)) - L_p(\boldsymbol{A}_{\boldsymbol{I}}(t))| \le (1 + R)K_{T_0,R}\|\boldsymbol{S} - \boldsymbol{I}_p\|_{\mathrm{op}}, \tag{95}$$

and

$$\sup_{0\le t\le T_0} |\widehat{L}_{p,\boldsymbol{S}}(\boldsymbol{A}_{\boldsymbol{S}}(t)) - L_p(\boldsymbol{A}_{\boldsymbol{I}}(t))| \le \left[\frac{1}{2}(1 + R)^2 + (1 + R)K_{T_0,R}\right]\|\boldsymbol{S} - \boldsymbol{I}_p\|_{\mathrm{op}}. \tag{96}$$

Proof. Set $\boldsymbol{E_A} = \boldsymbol{I}_p - \boldsymbol{A}$ and $\boldsymbol{E_B} = \boldsymbol{I}_p - \boldsymbol{B}$. Then

$$\boldsymbol{E_A}^2 - \boldsymbol{E_B}^2 = \boldsymbol{E_A}(\boldsymbol{E_A} - \boldsymbol{E_B}) + (\boldsymbol{E_A} - \boldsymbol{E_B})\boldsymbol{E_B}, \tag{97}$$

so

$$\|\boldsymbol{E_A}^2 - \boldsymbol{E_B}^2\|_{\mathrm{op}} \le 2(1+R)\|\boldsymbol{A} - \boldsymbol{B}\|_{\mathrm{op}}. \tag{98}$$

The first two inequalities follow from $|\operatorname{Tr} \boldsymbol{M}| \le p\|\boldsymbol{M}\|_{\mathrm{op}}$. For the covariance perturbation,

$$\widehat{L}_{p,\boldsymbol{S}}(\boldsymbol{A}) - L_p(\boldsymbol{A}) = \frac{1}{2p} \operatorname{Tr}\left[(\boldsymbol{S} - \boldsymbol{I}_p)(\boldsymbol{I}_p - \boldsymbol{A})^2\right]. \tag{99}$$

Since $\boldsymbol{Q} := (\boldsymbol{I}_p - \boldsymbol{A})^2 \succeq 0$,

$$|\operatorname{Tr}(\boldsymbol{MQ})| \le \|\boldsymbol{M}\|_{\mathrm{op}} \operatorname{Tr} \boldsymbol{Q}, \qquad \operatorname{Tr} \boldsymbol{Q} \le p(1+R)^2, \tag{100}$$

which proves equation 94. The last two inequalities combine the preceding estimates with Proposition C.5. □

Lemma C.7 (Commuting logistic flow). *Assume $\boldsymbol{A}_0, \boldsymbol{S} \succeq 0$ commute, and let*

$$\dot{\boldsymbol{A}} = \gamma \mathcal{F}_{\boldsymbol{S}}(\boldsymbol{A}), \qquad \boldsymbol{A}(0) = \boldsymbol{A}_0. \tag{101}$$

Then $\boldsymbol{A}(t)$ continues to commute with $\boldsymbol{S}$. In a common orthonormal eigenbasis,

$$\boldsymbol{S} = \boldsymbol{V} \operatorname{diag}(\lambda_1, \dots, \lambda_p)\boldsymbol{V}^\top, \qquad \boldsymbol{A}_0 = \boldsymbol{V} \operatorname{diag}(a_1(0), \dots, a_p(0))\boldsymbol{V}^\top, \tag{102}$$

the eigenvalues solve

$$\dot{a}_i(t) = 4\gamma\lambda_i a_i(t)(1 - a_i(t)), \qquad a_i(t) = \frac{a_i(0)}{a_i(0) + (1 - a_i(0))e^{-4\gamma\lambda_i t}}. \tag{103}$$

Zero eigenvalues of $\boldsymbol{A}_0$ remain zero.

Proof. On commuting matrices,

$$\mathcal{F}_{\boldsymbol{S}}(\boldsymbol{A}) = 4\boldsymbol{S}\boldsymbol{A}(\boldsymbol{I}_p - \boldsymbol{A}). \tag{104}$$

Therefore the common eigenbasis does not rotate, and the matrix ODE reduces to the scalar logistic equations in equation 103. Separation of variables gives the stated formula, with the cases $a_i(0) \in \{0, 1\}$ obtained by continuity. Uniqueness of the matrix ODE completes the argument. □

C.2 Mean-field (Section 4.1)

We now specialize the common setup of Appendix C.1 to the mean-field corner. All occurrences of $\boldsymbol{S}_p, \boldsymbol{A}_p, \boldsymbol{E}_p$, the train and population losses, and the gradient flow use the notation and identities of equation 37–equation 49; in particular the learning-rate scaling $\eta = p$ gives $\gamma = 1$. We write $L_p(t) := L(\boldsymbol{U}_p(t))$ and $\widehat{L}_p(t) := \widehat{L}(\boldsymbol{U}_p(t))$.

The regime-specific assumptions are

$$p, n, m \longrightarrow \infty, \qquad \frac{p}{n} \longrightarrow 0, \qquad \rho_p := n\sigma_p^2 \longrightarrow \rho \in (0, \infty), \tag{105}$$

These are the only mean-field-specific dimension and variance scalings. For the population-loss statement we first impose isotropy,

$$\boldsymbol{\Sigma}_p := \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{I}_p, \tag{106}$$

and return to anisotropic population covariance at the end of the subsection.

The data covariance $\boldsymbol{S}_p$ is assumed to satisfy the two standing spectral hypotheses equation 41 and equation 42, with limiting law ν_ϕ. The operator-norm bound gives finite-time stability, while $\mu_{\boldsymbol{S}_p} \Rightarrow \nu_\phi$ identifies the deterministic spectral averages in the losses.

For $\lambda \ge 0$ and $t \ge 0$, define

$$a_t(\lambda) := \frac{\rho}{\rho + (1 - \rho)e^{-4\lambda t}}, \tag{107}$$

and

$$e_t(\lambda) := 1 - a_t(\lambda) = \frac{1 - \rho}{\rho e^{4\lambda t} + 1 - \rho}. \tag{108}$$

We also define

$$\ell_t(\lambda) := \frac{1}{2} e_t(\lambda)^2 = \frac{1}{2}\left(\frac{1 - \rho}{\rho e^{4\lambda t} + 1 - \rho}\right)^2. \tag{109}$$

Main theorem.

Theorem C.8 (Finite-time mean-field limit)**.** *Assume equation 105 and the common spectral hypotheses equation 41–equation 42. Then, for every fixed $T < \infty$,*

$$\sup_{0\le t\le T} \|\boldsymbol{A}_p(t) - a_t(\boldsymbol{S}_p)\|_{\mathrm{op}} \xrightarrow[p\to\infty]{\mathbb{P}} 0, \tag{110}$$

where $a_t(\boldsymbol{S}_p)$ is defined by functional calculus.

Under the population-isotropy assumption equation 106, it follows that

$$\sup_{0\le t\le T} \left| L_p(t) - \frac{1}{2}\int e_t(\lambda)^2 \, d\nu_\phi(\lambda) \right| \xrightarrow[p\to\infty]{\mathbb{P}} 0, \tag{111}$$

and

$$\sup_{0\le t\le T} \left| \widehat{L}_p(t) - \frac{1}{2}\int \lambda e_t(\lambda)^2 \, d\nu_\phi(\lambda) \right| \xrightarrow[p\to\infty]{\mathbb{P}} 0. \tag{112}$$

Proof plan. First, Gaussian singular-value concentration shows that the random initial Gram matrix is close to $\rho\boldsymbol{I}_p$ in operator norm. Second, Lemma C.4 propagates this initial perturbation on every fixed time interval. Third, the scalar initial condition commutes with $\boldsymbol{S}_p$, so Lemma C.7 gives the proxy by functional calculus. Finally, Lemma C.6 transfers the matrix approximation to the two losses, and weak spectral convergence of $\boldsymbol{S}_p$ identifies their deterministic limits.

Common deterministic input. The closed Gram dynamics, global a priori bound, commuting logistic solution, and finite-time stability with respect to the Gram state are exactly Lemmas C.2, C.7, and C.4 from Appendix C.1.

Concentration of the initial Gram matrix.

Lemma C.9. *Under the standing initialization in Eq.* (36) *and equation 105,*

$$\|\boldsymbol{A}_p(0) - \rho\boldsymbol{I}_p\|_{\mathrm{op}} \xrightarrow[p\to\infty]{\mathbb{P}} 0. \tag{113}$$

More precisely,

$$\|\boldsymbol{A}_p(0) - \rho\boldsymbol{I}_p\|_{\mathrm{op}} = O_{\mathbb{P}}\left(\sqrt{\frac{p}{n}}\right) + |\rho_p - \rho|. \tag{114}$$

Proof. By the standing initialization Eq. (36),

$$\boldsymbol{A}_p(0) = \frac{\rho_p}{n}\boldsymbol{Z}_p^\top \boldsymbol{Z}_p. \tag{115}$$

A standard Gaussian singular-value estimate states that if $\boldsymbol{Z}$ is an $n \times p$ matrix with i.i.d. standard Gaussian entries, then, for every $r \ge 0$,

$$\mathbb{P}\left(s_{\max}(\boldsymbol{Z}) > \sqrt{n} + \sqrt{p} + r\right) \le e^{-r^2/2}, \tag{116}$$

and, whenever $\sqrt{n} \ge \sqrt{p} + r$,

$$\mathbb{P}\left(s_{\min}(\boldsymbol{Z}) < \sqrt{n} - \sqrt{p} - r\right) \le e^{-r^2/2}. \tag{117}$$

Taking $r = \sqrt{p}$ shows that, with probability at least $1 - 2e^{-p/2}$ for all sufficiently large p,

$$1 - 2\sqrt{\frac{p}{n}} \le \frac{s_{\min}(\boldsymbol{Z}_p)}{\sqrt{n}} \le \frac{s_{\max}(\boldsymbol{Z}_p)}{\sqrt{n}} \le 1 + 2\sqrt{\frac{p}{n}}. \tag{118}$$

Consequently,

$$\left\| \frac{1}{n}\boldsymbol{Z}_p^\top \boldsymbol{Z}_p - \boldsymbol{I}_p \right\|_{\mathrm{op}} = O_{\mathbb{P}}\left(\sqrt{\frac{p}{n}}\right). \tag{119}$$

Using equation 115,

$$\|\boldsymbol{A}_p(0) - \rho\boldsymbol{I}_p\|_{\mathrm{op}} \le \rho_p \left\| \frac{1}{n}\boldsymbol{Z}_p^\top \boldsymbol{Z}_p - \boldsymbol{I}_p \right\|_{\mathrm{op}} + |\rho_p - \rho|. \tag{120}$$

Equations (105), (119) and (120) prove the claims. □

Proof of the matrix trajectory limit.

Proof of equation 110. Let

$$\bar{\boldsymbol{A}}_p(t) := a_t(\boldsymbol{S}_p). \tag{121}$$

By Lemma C.7, $\bar{\boldsymbol{A}}_p$ is the exact solution of the same Gram ODE as $\boldsymbol{A}_p$, with initial condition

$$\bar{\boldsymbol{A}}_p(0) = \rho \boldsymbol{I}_p. \tag{122}$$

Fix $T < \infty$. By Lemma C.9,

$$\|\boldsymbol{A}_p(0) - \rho \boldsymbol{I}_p\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{123}$$

In particular, for any deterministic $R_0 > \max\{1, \rho\}$,

$$\mathbb{P}\big(\|\boldsymbol{A}_p(0)\|_{\mathrm{op}} \le R_0\big) \longrightarrow 1. \tag{124}$$

Intersecting this event with the event in equation 41, Lemma C.4 gives

$$\sup_{0\le t\le T} \|\boldsymbol{A}_p(t) - \bar{\boldsymbol{A}}_p(t)\|_{\mathrm{op}} \le \exp\big(K(4+8R_0)T\big)\,\|\boldsymbol{A}_p(0) - \rho \boldsymbol{I}_p\|_{\mathrm{op}}. \tag{125}$$

The right-hand side converges to zero in probability by equation 123, while the event on which the estimate holds has probability tending to one. This proves equation 110. □

The estimate equation 125 also makes explicit why the result is a finite-time statement. For fixed T the stability factor is harmless, but the argument by itself gives no useful control when T is allowed to diverge with the dimensions.

Transfer from the Gram trajectory to the losses. The needed population- and train-loss Lipschitz estimates are collected in Lemma C.6.

Applying Lemma C.6 to $\boldsymbol{A} = \boldsymbol{A}_p(t)$ and $\boldsymbol{B} = \bar{\boldsymbol{A}}_p(t)$, and using equation 110, gives

$$\sup_{0\le t\le T} \left| L_p(t) - \frac{1}{2p} \operatorname{Tr}\big[\boldsymbol{I}_p - a_t(\boldsymbol{S}_p)\big]^2 \right| \xrightarrow{\mathbb{P}} 0, \tag{126}$$

and

$$\sup_{0\le t\le T} \left| \widehat{L}_p(t) - \frac{1}{2p} \operatorname{Tr}\Big[\boldsymbol{S}_p\big(\boldsymbol{I}_p - a_t(\boldsymbol{S}_p)\big)^2\Big] \right| \xrightarrow{\mathbb{P}} 0. \tag{127}$$

Since $a_t(\boldsymbol{S}_p)$ is a spectral function of $\boldsymbol{S}_p$, the comparison losses can be written exactly as

$$\frac{1}{2p} \operatorname{Tr}\big[\boldsymbol{I}_p - a_t(\boldsymbol{S}_p)\big]^2 = \int \ell_t(\lambda)\, d\mu_{\boldsymbol{S}_p}(\lambda), \tag{128}$$

and

$$\frac{1}{2p} \operatorname{Tr}\Big[\boldsymbol{S}_p\big(\boldsymbol{I}_p - a_t(\boldsymbol{S}_p)\big)^2\Big] = \int \lambda \ell_t(\lambda)\, d\mu_{\boldsymbol{S}_p}(\lambda). \tag{129}$$

Thus, before taking any random-matrix limit, the full finite-dimensional dynamics have already been reduced, up to an error tending to zero uniformly on $[0, T]$, to spectral averages over the empirical covariance $\boldsymbol{S}_p$.

Uniform convergence of the spectral averages.

Lemma C.10. *Under equation 41 and equation 42, for every fixed $T < \infty$,*

$$\sup_{0\le t\le T} \left| \int \ell_t(\lambda)\, d\mu_{\boldsymbol{S}_p}(\lambda) - \int \ell_t(\lambda)\, d\nu_\phi(\lambda) \right| \xrightarrow{\mathbb{P}} 0, \tag{130}$$

and

$$\sup_{0\le t\le T} \left| \int \lambda \ell_t(\lambda)\, d\mu_{\boldsymbol{S}_p}(\lambda) - \int \lambda \ell_t(\lambda)\, d\nu_\phi(\lambda) \right| \xrightarrow{\mathbb{P}} 0. \tag{131}$$

Proof. We prove a slightly more general elementary statement. Let

$$h : [0, T] \times [0, K] \longrightarrow \mathbb{R} \tag{132}$$

be continuous. Since its domain is compact, h is uniformly continuous. Hence, for every $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|t - s| < \delta \quad \Longrightarrow \quad \sup_{\lambda \in [0,K]} |h(t, \lambda) - h(s, \lambda)| < \frac{\varepsilon}{3}. \tag{133}$$

Choose a finite δ-net $t_1, \ldots, t_N$ of $[0, T]$. For every fixed j, weak convergence in probability in equation 42 gives

$$\int h(t_j, \lambda)\, d\mu_{\boldsymbol{S}_p}(\lambda) - \int h(t_j, \lambda)\, d\nu_\phi(\lambda) \xrightarrow{\mathbb{P}} 0. \tag{134}$$

Since N is finite,

$$\max_{1 \le j \le N} \left| \int h(t_j, \lambda)\, d\mu_{\boldsymbol{S}_p}(\lambda) - \int h(t_j, \lambda)\, d\nu_\phi(\lambda) \right| \xrightarrow{\mathbb{P}} 0. \tag{135}$$

On the event $\|\boldsymbol{S}_p\|_{\mathrm{op}} \le K$, for arbitrary $t \in [0, T]$ choose t_j with $|t - t_j| < \delta$. Then

$$\begin{aligned} \left| \int h(t, \lambda)\, d(\mu_{\boldsymbol{S}_p} - \nu_\phi)(\lambda) \right| \le & \int |h(t, \lambda) - h(t_j, \lambda)|\, d\mu_{\boldsymbol{S}_p}(\lambda) \\ & + \left| \int h(t_j, \lambda)\, d(\mu_{\boldsymbol{S}_p} - \nu_\phi)(\lambda) \right| \\ & + \int |h(t_j, \lambda) - h(t, \lambda)|\, d\nu_\phi(\lambda). \end{aligned} \tag{136}$$

The first and third terms in equation 136 are each less than $\varepsilon/3$ by equation 133, while the middle term is controlled uniformly over the finite net by equation 135. Since the event $\|\boldsymbol{S}_p\|_{\mathrm{op}} \le K$ has probability tending to one, it follows that

$$\sup_{0 \le t \le T} \left| \int h(t, \lambda)\, d\mu_{\boldsymbol{S}_p}(\lambda) - \int h(t, \lambda)\, d\nu_\phi(\lambda) \right| \xrightarrow{\mathbb{P}} 0. \tag{137}$$

Taking

$$h(t, \lambda) = \ell_t(\lambda) \tag{138}$$

and

$$h(t, \lambda) = \lambda \ell_t(\lambda) \tag{139}$$

proves equation 130 and equation 131. □

Combining equation 126, equation 127, and Lemma C.10 proves equation 111 and equation 112, and hence completes the proof of Theorem C.8.

Gaussian data and the Marchenko–Pastur formula. We now specialize Theorem C.8 to the data model used in the main text:

$$(\boldsymbol{X}_p)_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), \qquad \frac{p}{m} \longrightarrow \phi \in (0, \infty). \tag{140}$$

The Marchenko–Pastur theorem and the corresponding extreme-eigenvalue convergence imply

$$\mu_{\boldsymbol{S}_p} \Longrightarrow \mu_\phi \qquad \text{almost surely,} \tag{141}$$

and

$$\|\boldsymbol{S}_p\|_{\mathrm{op}} \longrightarrow (1 + \sqrt{\phi})^2 \qquad \text{almost surely.} \tag{142}$$

Here

$$d\mu_\phi(\lambda) = \left(1 - \frac{1}{\phi}\right)_+ \delta_0(d\lambda) + \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\phi\lambda} \mathbf{1}_{[\lambda_-, \lambda_+]}(\lambda)\, d\lambda, \tag{143}$$

with

$$\lambda_\pm = (1 \pm \sqrt{\phi})^2. \tag{144}$$

Thus the assumptions of Theorem C.8 hold after choosing any fixed

$$K > (1 + \sqrt{\phi})^2. \tag{145}$$

Therefore, for every fixed $T < \infty$,

$$\sup_{0 \le t \le T} |L_p(t) - \mathcal{L}(t)| \xrightarrow{\mathbb{P}} 0, \tag{146}$$

and

$$\sup_{0 \le t \le T} \left| \widehat{L}_p(t) - \widehat{\mathcal{L}}(t) \right| \xrightarrow{\mathbb{P}} 0, \tag{147}$$

where

$$\mathcal{L}(t) = \frac{1}{2} \int \left(\frac{1-\rho}{\rho e^{4\lambda t} + 1 - \rho} \right)^2 d\mu_\phi(\lambda), \tag{148}$$

and

$$\widehat{\mathcal{L}}(t) = \frac{1}{2} \int \lambda \left(\frac{1-\rho}{\rho e^{4\lambda t} + 1 - \rho} \right)^2 d\mu_\phi(\lambda). \tag{149}$$

These are the mean-field formulas stated in Section 4.1.

The preceding result is stronger than convergence of the averaged losses, since it gives convergence in probability of the entire loss trajectories on compact time intervals. The expectation-level limits used in the main text follow by uniform integrability.

Corollary C.11 (Convergence of averaged losses)**.** *Under the standing initialization Eq.* (36)*, equation 105, and equation 140, for every fixed $T < \infty$,*

$$\mathbb{E}\left[\sup_{0 \le t \le T} |L_p(t) - \mathcal{L}(t)| \right] \longrightarrow 0, \tag{150}$$

and

$$\mathbb{E}\left[\sup_{0 \le t \le T} |\widehat{L}_p(t) - \widehat{\mathcal{L}}(t)| \right] \longrightarrow 0. \tag{151}$$

In particular, for every fixed $t \ge 0$,

$$\mathbb{E}[L_p(t)] \longrightarrow \mathcal{L}(t), \qquad \mathbb{E}[\widehat{L}_p(t)] \longrightarrow \widehat{\mathcal{L}}(t). \tag{152}$$

Proof. By Lemma C.2, with

$$M_p := \max\{1, \|\boldsymbol{A}_p(0)\|_{\mathrm{op}}\}, \tag{153}$$

one has uniformly in time

$$\|\boldsymbol{A}_p(t)\|_{\mathrm{op}} \le M_p. \tag{154}$$

Hence

$$\sup_{t \ge 0} L_p(t) \le \frac{1}{2}(1 + M_p)^2, \tag{155}$$

and

$$\sup_{t \ge 0} \widehat{L}_p(t) \le \frac{1}{2} \|\boldsymbol{S}_p\|_{\mathrm{op}} (1 + M_p)^2. \tag{156}$$

The Gaussian singular-value bounds used in Lemma C.9, applied to $\boldsymbol{Z}_p$, imply that for every fixed $q < \infty$,

$$\sup_p \mathbb{E}[M_p^q] < \infty. \tag{157}$$

The analogous Gaussian singular-value bound applied to $\boldsymbol{X}_p$, together with $p/m \to \phi$, gives

$$\sup_p \mathbb{E}\left[\|\boldsymbol{S}_p\|_{\mathrm{op}}^q \right] < \infty \qquad \text{for every fixed } q < \infty. \tag{158}$$

Equations (155) to (158) imply uniform integrability of the two suprema appearing in equation 150 and equation 151. Their convergence to zero in probability follows from equation 146 and equation 147, together with boundedness of the deterministic limit trajectories. Uniform integrability therefore upgrades convergence in probability to convergence in L^1. □

Universality within the isotropic setting. The dynamical part of Theorem C.8 does not use Gaussianity of the data. In particular, for any isotropic training-data ensemble satisfying equation 41 and equation 42, the finite-time limiting losses are

$$\mathcal{L}_{\nu_\phi}(t) = \frac{1}{2}\int \left(\frac{1-\rho}{\rho e^{4\lambda t}+1-\rho}\right)^2 d\nu_\phi(\lambda), \tag{159}$$

and

$$\widehat{\mathcal{L}}_{\nu_\phi}(t) = \frac{1}{2}\int \lambda \left(\frac{1-\rho}{\rho e^{4\lambda t}+1-\rho}\right)^2 d\nu_\phi(\lambda). \tag{160}$$

Thus the Marchenko–Pastur law is not intrinsic to the mean-field dynamical reduction: it enters only through the limiting spectrum of the empirical covariance. Data ensembles in the Marchenko–Pastur universality class give exactly equation 148–equation 149, whereas ensembles with a different limiting covariance spectrum give the same spectral formula with μ_ϕ replaced by the corresponding measure ν_ϕ.

Remark on non-isotropic population covariance. We finally indicate how the result changes if the population covariance is not the identity. Let

$$\boldsymbol{\Sigma}_p := \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] \succeq 0 \tag{161}$$

be arbitrary. The empirical gradient flow equation 49 depends only on the training covariance $\boldsymbol{S}_p$ and is therefore unchanged. Consequently, under the standing initialization Eq. (36), equation 105, and equation 41, the matrix trajectory conclusion

$$\sup_{0\le t\le T} \|\boldsymbol{A}_p(t) - a_t(\boldsymbol{S}_p)\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0 \tag{162}$$

continues to hold without any population-isotropy assumption.

The train loss is also unchanged and is still controlled solely by the ordinary empirical spectral distribution of $\boldsymbol{S}_p$. The population loss, however, is now

$$L_p^{(\Sigma)}(t) = \frac{1}{2p}\operatorname{Tr}\left[(\boldsymbol{I}_p - \boldsymbol{A}_p(t))\boldsymbol{\Sigma}_p(\boldsymbol{I}_p - \boldsymbol{A}_p(t))\right] = \frac{1}{2p}\operatorname{Tr}\left[\boldsymbol{\Sigma}_p(\boldsymbol{I}_p - \boldsymbol{A}_p(t))^2\right]. \tag{163}$$

Thus ordinary spectral convergence of $\boldsymbol{S}_p$ is no longer sufficient to determine the population limit.

To make the required additional information explicit, let

$$\boldsymbol{S}_p = \sum_{i=1}^{p} \lambda_{i,p} q_{i,p} q_{i,p}^\top \tag{164}$$

be an orthonormal spectral decomposition, and define the $\boldsymbol{\Sigma}_p$-weighted empirical spectral measure

$$\omega_p := \frac{1}{p}\sum_{i=1}^{p} \big(q_{i,p}^\top \boldsymbol{\Sigma}_p q_{i,p}\big)\delta_{\lambda_{i,p}}. \tag{165}$$

Equivalently, for every bounded measurable function f,

$$\int f(\lambda)\, d\omega_p(\lambda) = \frac{1}{p}\operatorname{Tr}\big[\boldsymbol{\Sigma}_p f(\boldsymbol{S}_p)\big]. \tag{166}$$

Suppose, in addition to the assumptions used for equation 162, that

$$\frac{1}{p}\operatorname{Tr}\boldsymbol{\Sigma}_p = O(1), \tag{167}$$

and that, for a deterministic finite measure ω supported on $[0, K]$,

$$\omega_p \Longrightarrow \omega \qquad \text{in probability.} \tag{168}$$

Then

$$\sup_{0\le t\le T}\left|L_p^{(\Sigma)}(t) - \frac{1}{2}\int e_t(\lambda)^2\, d\omega(\lambda)\right| \xrightarrow{\mathbb{P}} 0. \tag{169}$$

Indeed, for $\boldsymbol{A}, \boldsymbol{B} \succeq 0$ with

$$\|\boldsymbol{A}\|_{\mathrm{op}}, \|\boldsymbol{B}\|_{\mathrm{op}} \leq R, \tag{170}$$

one has

$$\left| L_p^{(\Sigma)}(\boldsymbol{A}) - L_p^{(\Sigma)}(\boldsymbol{B}) \right| \leq \frac{\operatorname{Tr} \boldsymbol{\Sigma}_p}{p} (1+R) \|\boldsymbol{A} - \boldsymbol{B}\|_{\mathrm{op}}. \tag{171}$$

To see this, set $\boldsymbol{E}_{\boldsymbol{A}} = \boldsymbol{I}_p - \boldsymbol{A}$ and $\boldsymbol{E}_{\boldsymbol{B}} = \boldsymbol{I}_p - \boldsymbol{B}$. Since $\boldsymbol{\Sigma}_p \succeq 0$,

$$\left| \operatorname{Tr} \left[\boldsymbol{\Sigma}_p (\boldsymbol{E}_{\boldsymbol{A}}^2 - \boldsymbol{E}_{\boldsymbol{B}}^2) \right] \right| \leq \operatorname{Tr}(\boldsymbol{\Sigma}_p) \, \|\boldsymbol{E}_{\boldsymbol{A}}^2 - \boldsymbol{E}_{\boldsymbol{B}}^2\|_{\mathrm{op}}, \tag{172}$$

while

$$\|\boldsymbol{E}_{\boldsymbol{A}}^2 - \boldsymbol{E}_{\boldsymbol{B}}^2\|_{\mathrm{op}} \leq 2(1+R) \|\boldsymbol{A} - \boldsymbol{B}\|_{\mathrm{op}}. \tag{173}$$

Equations (172) and (173) give equation 171 after multiplication by $1/(2p)$.

For the comparison trajectory,

$$\frac{1}{2p} \operatorname{Tr} \left[\boldsymbol{\Sigma}_p \big(\boldsymbol{I}_p - a_t(\boldsymbol{S}_p) \big)^2 \right] = \frac{1}{2} \int e_t(\lambda)^2 \, d\omega_p(\lambda). \tag{174}$$

The convergence in equation 169 then follows by combining equation 162, equation 167, equation 171, and a finite-net argument identical to the proof of Lemma C.10.

When $\boldsymbol{\Sigma}_p = \boldsymbol{I}_p$, the weights in equation 165 are all equal to one, and hence

$$\omega_p = \mu_{\boldsymbol{S}_p}. \tag{175}$$

Thus equation 169 reduces to Theorem C.8. In the genuinely anisotropic case, by contrast, the ordinary limiting eigenvalue distribution ν_ϕ of $\boldsymbol{S}_p$ does not in general determine ω: the latter also records the alignment between the eigenvectors of the empirical covariance and the population covariance $\boldsymbol{\Sigma}_p$. Thus anisotropy leaves the leading mean-field matrix dynamics unchanged but introduces an additional weighted spectral object in the population loss.

The results of this subsection concern the dimension limit on fixed compact time intervals. They do not assert uniformity as $T \to \infty$ and therefore do not, by themselves, justify interchange of the large-dimension and large-time limits analyzed separately in Appendix C.2.1.

C.2.1 Large time behavior

For every positive empirical covariance eigenvalue $\lambda > 0$, we have

$$a_\lambda(t) = \frac{\rho}{\rho + (1-\rho)e^{-4\lambda t}} \to 1 \qquad \text{as } t \to \infty. \tag{176}$$

Therefore

$$\ell(\lambda t) \to 0 \qquad \text{for every } \lambda > 0. \tag{177}$$

If $\phi \leq 1$, the limiting empirical covariance has no atom at zero, and hence $\lim_{t\to\infty} L(t) = 0$. If $\phi > 1$, then μ_ϕ has an atom at $\lambda = 0$ of mass $1 - \frac{1}{\phi}$. Those zero-eigenvalue directions are never trained by the empirical loss, so their reconstruction factor remains equal to its initialization value ρ. Hence

$$\ell(0) = \frac{1}{2}(1-\rho)^2. \tag{178}$$

Therefore

$$\mathcal{L}(\infty) = \frac{1}{2}(1-\rho)^2 \left(1 - \frac{1}{\phi}\right)_+. \tag{179}$$

On the other hand, the empirical train loss contains an additional factor of λ. Therefore the zero-eigenvalue directions do not contribute to the empirical loss, and

$$\lim_{t\to\infty} \widehat{\mathcal{L}}(t) = 0. \tag{180}$$

We now compute the exact train loss asymptotics for large t. For $\phi = 1$, since $\lambda_- = 0$ and $\lambda_+ = 4$, using a Laplace approximation,

$$\begin{aligned}\widehat{\mathcal{L}}(t) &= \frac{1}{2}\int_0^4 \lambda \left(\frac{1-\rho}{\rho e^{4\lambda t}+1-\rho}\right)^2 \frac{\sqrt{(4-\lambda)\lambda}}{2\pi\lambda}\, d\lambda \\ &= \frac{1}{4\pi}\int_0^4 \left(\frac{1-\rho}{\rho e^{4\lambda t}+1-\rho}\right)^2 \sqrt{(4-\lambda)\lambda}\, d\lambda \\ &\sim \frac{1}{2\pi}\int_0^\infty \left(\frac{1-\rho}{\rho e^{4\lambda t}+1-\rho}\right)^2 \sqrt{\lambda}\, d\lambda\end{aligned} \tag{181}$$

as $t \to \infty$. This integral could be evaluated explicitly:

$$\widehat{\mathcal{L}}(t) \sim \frac{\mathrm{Li}_{1/2}\left(\frac{\rho-1}{\rho}\right) - \mathrm{Li}_{3/2}\left(\frac{\rho-1}{\rho}\right)}{32\sqrt{\pi}t^{3/2}}, \qquad t\to\infty, \qquad \phi = 1. \tag{182}$$

In contrast, when $\phi \neq 1$, $\lambda_- > 0$, which yields

$$\begin{aligned}\widehat{\mathcal{L}}(t) &= \frac{1}{2}\int_{\lambda_-}^{\lambda_+} \lambda \left(\frac{1-\rho}{\rho e^{4\lambda t}+1-\rho}\right)^2 \frac{\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}}{2\pi\phi\lambda}\, d\lambda \\ &= \frac{1}{4\pi\phi}\int_{\lambda_-}^{\lambda_+} \left(\frac{1-\rho}{\rho e^{4\lambda t}+1-\rho}\right)^2 \sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}\, d\lambda \\ &\sim \frac{\sqrt{\lambda_+-\lambda_-}}{4\pi\phi}\left(\frac{1-\rho}{\rho}\right)^2 \int_{\lambda_-}^{\lambda_+} e^{-8\lambda t}\sqrt{\lambda-\lambda_-}\, d\lambda\end{aligned} \tag{183}$$

as $t \to \infty$. This integral could be evaluated explicitly:

$$\widehat{\mathcal{L}}(t) \sim \left(\frac{1-\rho}{\rho}\right)^2 \frac{e^{-8(1-\sqrt{\phi})^2 t}}{64\sqrt{2\pi}\phi^{3/4}t^{3/2}}, \qquad t\to\infty, \qquad \phi \neq 1. \tag{184}$$

Putting all together,

$$\boxed{\widehat{\mathcal{L}}(t) \sim \frac{1}{64\sqrt{2\pi}\phi^{3/4}t^{3/2}} \begin{cases} 2\sqrt{2}\left[\mathrm{Li}_{1/2} - \mathrm{Li}_{3/2}\right]\left(\frac{\rho-1}{\rho}\right), & \phi = 1, \\ \left(\frac{1-\rho}{\rho}\right)^2 e^{-8(1-\sqrt{\phi})^2 t}, & \phi \neq 1 \end{cases} \qquad t\to\infty.} \tag{185}$$

We now perform a similar computation for the population loss. For $\phi = 1$, again using a Laplace approximation,

$$\begin{aligned}\mathcal{L}(t) - \mathcal{L}(\infty) &= \frac{1}{2}\int_0^4 \left(\frac{1-\rho}{\rho e^{4\lambda t}+1-\rho}\right)^2 \frac{\sqrt{(4-\lambda)\lambda}}{2\pi\lambda}\, d\lambda \\ &= \frac{1}{4\pi}\int_0^4 \left(\frac{1-\rho}{\rho e^{4\lambda t}+1-\rho}\right)^2 \sqrt{\frac{4-\lambda}{\lambda}}\, d\lambda \\ &\sim \frac{1}{2\pi}\int_0^\infty \left(\frac{1-\rho}{\rho e^{4\lambda t}+1-\rho}\right)^2 \frac{d\lambda}{\sqrt{\lambda}} \\ &= \frac{\mathrm{Li}_{-1/2}\left(\frac{\rho-1}{\rho}\right) - \mathrm{Li}_{1/2}\left(\frac{\rho-1}{\rho}\right)}{4\sqrt{\pi}t^{1/2}}\end{aligned} \tag{186}$$

as $t \to \infty$. When $\phi \neq 1$, $\lambda_- > 0$, which yields similarly to the train loss case,

$$\begin{aligned}\mathcal{L}(t) - \mathcal{L}(\infty) &= \frac{1}{2}\int_{\lambda_-}^{\lambda_+} \left(\frac{1-\rho}{\rho e^{4\lambda t} + 1 - \rho}\right)^2 \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\phi\lambda}\, d\lambda \\ &= \frac{1}{4\pi\phi}\int_{\lambda_-}^{\lambda_+} \left(\frac{1-\rho}{\rho e^{4\lambda t} + 1 - \rho}\right)^2 \sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}\, \frac{d\lambda}{\lambda} \\ &\sim \frac{\sqrt{\lambda_+ - \lambda_-}}{4\pi\phi}\left(\frac{1-\rho}{\rho}\right)^2 \int_{\lambda_-}^{\lambda_+} e^{-8\lambda t}\sqrt{\lambda - \lambda_-}\, \frac{d\lambda}{\lambda} \\ &= \left(\frac{1-\rho}{\rho}\right)^2 \frac{e^{-8(1-\sqrt{\phi})^2 t}}{64\sqrt{2\pi}\phi^{3/4}\left(1-\sqrt{\phi}\right)^2 t^{3/2}}\end{aligned} \tag{187}$$

as $t \to \infty$. Putting all together,

$$\boxed{\mathcal{L}(t) - \mathcal{L}(\infty) \sim \frac{1}{32\sqrt{\pi}\phi^{3/4}} \begin{cases} \frac{8}{t^{1/2}}\left[\mathrm{Li}_{-1/2} - \mathrm{Li}_{1/2}\right]\left(\frac{\rho-1}{\rho}\right), & \phi = 1, \\ \left(\frac{1-\rho}{\rho}\right)^2 \frac{e^{-8(1-\sqrt{\phi})^2 t}}{\left(1-\sqrt{\phi}\right)^2 (2t)^{3/2}}, & \phi \neq 1 \end{cases} \quad t \to \infty.} \tag{188}$$

C.3 Narrow-latent (Section 4.3)

This subsection uses the common notation, losses, and gradient-flow identity from Appendix C.1 and records only the additional objects specific to the narrow-latent reduction. We take $\eta = p$, hence $\gamma = 1$, and

$$p = p_N \longrightarrow \infty, \qquad n = n_N \longrightarrow \infty, \qquad m = m_N \longrightarrow \infty, \qquad \frac{n}{p} \longrightarrow 0, \qquad \kappa_N := p_N \sigma_N^2 \longrightarrow \kappa \in (0, \infty). \tag{189}$$

Thus

$$\psi_N := \frac{p_N}{n_N} \longrightarrow \infty, \qquad \rho_N := n_N \sigma_N^2, \qquad \kappa_N = \psi_N \rho_N. \tag{190}$$

The training samples are centered and isotropic,

$$\mathbb{E}\boldsymbol{x} = 0, \qquad \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{I}_p. \tag{191}$$

The empirical covariance $\boldsymbol{S}_N$ is the common matrix equation 44. We assume the standing spectral conditions equation 41 and equation 42, with limiting law ν_ϕ; in this subsection we denote the common deterministic spectral bound K by L. Spectral boundedness controls the matrix ODE on compact time intervals and the weak spectral limit determines the scalar limiting coefficients.

The regime-specific state variables are the two latent-space Gramians

$$\boldsymbol{G}_N(t) := \boldsymbol{U}_N(t)\boldsymbol{U}_N(t)^\top, \qquad \boldsymbol{H}_N(t) := \boldsymbol{U}_N(t)\boldsymbol{S}_N\boldsymbol{U}_N(t)^\top. \tag{192}$$

Specializing the common flow equation 49 to $\gamma = 1$ gives the exact latent-space form

$$\dot{\boldsymbol{U}}_N = 2\boldsymbol{U}_N\boldsymbol{S}_N - \boldsymbol{H}_N\boldsymbol{U}_N - \boldsymbol{G}_N\boldsymbol{U}_N\boldsymbol{S}_N. \tag{193}$$

By equation 41–equation 42, the limiting law satisfies $\operatorname{supp}\nu_\phi \subset [0, L]$.

Define the exponential transform by

$$M_{\nu_\phi}(q) := \int e^{q\lambda}\, d\nu_\phi(\lambda). \tag{194}$$

and its logarithmic derivative by

$$m_{\nu_\phi}(q) := \frac{M'_{\nu_\phi}(q)}{M_{\nu_\phi}(q)}. \tag{195}$$

The limiting scalar dynamics are

$$\begin{aligned}\dot{a}(t) &= 4a(t)m_{\nu_\phi}(q(t))(1 - a(t)), \\ \dot{q}(t) &= 2(2 - a(t)), \\ a(0) &= \kappa, \qquad q(0) = 0,\end{aligned} \tag{196}$$

and

$$r(t) := a(t) m_{\nu_\phi}(q(t)). \tag{197}$$

These three scalars have a direct interpretation. Since the theorem below shows $\boldsymbol{G}_N = \boldsymbol{U}_N \boldsymbol{U}_N^\top \approx a\boldsymbol{I}_n$, the quantity $a(t)$ is the asymptotic squared norm of a typical latent row of $\boldsymbol{U}_N(t)$, while the off-diagonal part of $\boldsymbol{G}_N$ vanishes. Likewise $\boldsymbol{H}_N = \boldsymbol{U}_N \boldsymbol{S}_N \boldsymbol{U}_N^\top \approx r\boldsymbol{I}_n$, so $r(t)$ is the corresponding covariance-weighted row energy, or data-alignment scale. Their ratio has the spectral meaning

$$\frac{r(t)}{a(t)} = m_{\nu_\phi}(q(t)) = \int \lambda \, d\nu_{\phi_{q(t)}}(\lambda), \tag{198}$$

where ν_{ϕ_q} is the exponentially tilted law defined below. Finally, $q(t)$ is the spectral-tilting clock. In the proxy solution one has $q = 2R$ and a spectral component with data eigenvalue λ is multiplied by $e^{q\lambda/2}$; hence positive q progressively favors large-eigenvalue directions, whereas negative q favors the lower edge and, when present, the null space.

Main theorem.

Theorem C.12 (Narrow-latent limit for a general spectral measure)**.** *Assume equation 189 and the common spectral hypotheses equation 41–equation 42. Let a, q, r be defined by equation 196 and equation 197. Then, for every fixed $T < \infty$,*

$$\sup_{0 \le t \le T} \|\boldsymbol{G}_N(t) - a(t)\boldsymbol{I}_n\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0 \tag{199}$$

and

$$\sup_{0 \le t \le T} \|\boldsymbol{H}_N(t) - r(t)\boldsymbol{I}_n\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{200}$$

Proof plan.

Step 1: express the losses through $\boldsymbol{G}$ and $\boldsymbol{H}$. Population isotropy implies

$$L(\boldsymbol{U}) = \frac{1}{2p} \operatorname{Tr}(\boldsymbol{I}_p - \boldsymbol{U}^\top \boldsymbol{U})^2. \tag{201}$$

Using cyclicity of trace,

$$L(\boldsymbol{U}) = \frac{1}{2} - \frac{1}{p} \operatorname{Tr} \boldsymbol{G} + \frac{1}{2p} \operatorname{Tr} \boldsymbol{G}^2. \tag{202}$$

Similarly, the empirical loss satisfies

$$\widehat{L}(\boldsymbol{U}) = \frac{1}{2p} \operatorname{Tr} \boldsymbol{S} - \frac{1}{p} \operatorname{Tr} \boldsymbol{H} + \frac{1}{2p} \operatorname{Tr}(\boldsymbol{G}\boldsymbol{H}). \tag{203}$$

Consequently, the losses are determined at the $1/\psi_N$ scale once $\boldsymbol{G}$ and $\boldsymbol{H}$ are known to leading order.

Step 2: identify the desired scalarization. We shall construct scalar functions $a_N(t)$ and $r_N(t)$ and prove, uniformly on every fixed interval $0 \le t \le T$,

$$\boldsymbol{G}(t) = a_N(t)\boldsymbol{I}_n + o_{\mathbb{P}}(1) \quad \text{in operator norm,} \tag{204}$$

and

$$\boldsymbol{H}(t) = r_N(t)\boldsymbol{I}_n + o_{\mathbb{P}}(1) \quad \text{in operator norm.} \tag{205}$$

This is the rigorous formulation of asymptotic row orthogonality and self-averaging in the narrow-latent regime.

Step 3: construct a scalarized proxy dynamics. If one formally replaces

$$\boldsymbol{G}(t) \approx a_N(t)\boldsymbol{I}_n, \qquad \boldsymbol{H}(t) \approx r_N(t)\boldsymbol{I}_n \tag{206}$$

in Eq. (193), then one obtains

$$\dot{\overline{\boldsymbol{U}}} = (2 - a_N)\overline{\boldsymbol{U}}\boldsymbol{S} - r_N\overline{\boldsymbol{U}}. \tag{207}$$

This equation can be solved explicitly:

$$\overline{\boldsymbol{U}}(t) = e^{-B_N(t)}\boldsymbol{U}(0)e^{R_N(t)\boldsymbol{S}}, \tag{208}$$

where

$$\dot{R}_N(t) = 2 - a_N(t), \qquad \dot{B}_N(t) = r_N(t). \tag{209}$$

The central dynamical task is then to prove

$$\sup_{0\le t\le T} \|\boldsymbol{U}(t) - \overline{\boldsymbol{U}}(t)\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{210}$$

Step 4: prove self-consistency of the proxy. For the explicit proxy $\overline{\boldsymbol{U}}$, define

$$\overline{\boldsymbol{G}} = \overline{\boldsymbol{U}}\overline{\boldsymbol{U}}^\top, \qquad \overline{\boldsymbol{H}} = \overline{\boldsymbol{U}}\boldsymbol{S}\overline{\boldsymbol{U}}^\top. \tag{211}$$

Conditional on $\boldsymbol{S}$, these are weighted Gaussian sample-covariance matrices. Since $n/p \to 0$, we shall prove

$$\sup_{0\le t\le T} \|\overline{\boldsymbol{G}}(t) - a_N(t)\boldsymbol{I}_n\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0 \tag{212}$$

and

$$\sup_{0\le t\le T} \|\overline{\boldsymbol{H}}(t) - r_N(t)\boldsymbol{I}_n\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{213}$$

These estimates imply that the proxy solves the exact matrix ODE up to a vanishing residual. Lemma C.3 then proves equation 210. Equations (212) to (210) together yield the desired scalarization equation 204–equation 205.

Thus, logically, Step 4 supplies the estimate that closes the comparison introduced in Step 3.

Step 5: pass from the empirical scalar dynamics to a deterministic spectral limit. The finite-N functions a_N, r_N depend on the empirical spectrum of $\boldsymbol{S}_N$. If

$$\mu_{\boldsymbol{S}_N} = \frac{1}{p}\sum_{j=1}^{p} \delta_{\lambda_j(\boldsymbol{S}_N)} \xrightarrow{\mathbb{P}} \nu_\phi \tag{214}$$

for some deterministic probability measure ν_ϕ, then we shall prove

$$\sup_{0\le t\le T} |a_N(t) - a(t)| \xrightarrow{\mathbb{P}} 0 \tag{215}$$

and

$$\sup_{0\le t\le T} |r_N(t) - r(t)| \xrightarrow{\mathbb{P}} 0, \tag{216}$$

where a, r are determined by the exponential transform of ν_ϕ. No Marchenko–Pastur assumption is required for this step.

Combining all five steps gives the chain

$$\begin{aligned} \boldsymbol{G}, \boldsymbol{H} &\approx \overline{\boldsymbol{G}}, \overline{\boldsymbol{H}} \\ &\approx a_N\boldsymbol{I}_n, r_N\boldsymbol{I}_n \\ &\approx a\boldsymbol{I}_n, r\boldsymbol{I}_n, \end{aligned} \tag{217}$$

which inserted into equation 202–equation 203 produces the limiting loss formulas.

We now carry out these steps.

C.3.1 Exact identities and a priori bounds

The common Gram-flow bound, Lemma C.2, applies to equation 49 with $\gamma = 1$ and gives global existence together with the dimension-free estimate equation 55.

The loss identities announced in Step 1 are also exact.

Lemma C.13 (Exact loss representation)**.** *Under the isotropy assumption equation 191,*

$$L(\boldsymbol{U}) = \frac{1}{2} - \frac{1}{p}\operatorname{Tr}\boldsymbol{G} + \frac{1}{2p}\operatorname{Tr}\boldsymbol{G}^2. \tag{218}$$

For every data matrix $\boldsymbol{X}$,

$$\widehat{L}(\boldsymbol{U}) = \frac{1}{2p}\operatorname{Tr}\boldsymbol{S} - \frac{1}{p}\operatorname{Tr}\boldsymbol{H} + \frac{1}{2p}\operatorname{Tr}(\boldsymbol{G}\boldsymbol{H}). \tag{219}$$

Proof. Isotropy gives

$$\begin{aligned} L(\boldsymbol{U}) &= \frac{1}{2p}\mathbb{E}\left[\boldsymbol{x}^\top(\boldsymbol{I}_p - \boldsymbol{U}^\top\boldsymbol{U})^2\boldsymbol{x}\right] \\ &= \frac{1}{2p}\operatorname{Tr}(\boldsymbol{I}_p - \boldsymbol{U}^\top\boldsymbol{U})^2. \end{aligned} \tag{220}$$

Expanding and using

$$\operatorname{Tr}(\boldsymbol{U}^\top\boldsymbol{U}) = \operatorname{Tr}(\boldsymbol{U}\boldsymbol{U}^\top) = \operatorname{Tr}\boldsymbol{G} \tag{221}$$

and

$$\operatorname{Tr}(\boldsymbol{U}^\top\boldsymbol{U})^2 = \operatorname{Tr}(\boldsymbol{U}\boldsymbol{U}^\top)^2 = \operatorname{Tr}\boldsymbol{G}^2 \tag{222}$$

proves equation 218.

For the empirical loss,

$$\widehat{L}(\boldsymbol{U}) = \frac{1}{2p}\operatorname{Tr}\left[(\boldsymbol{I}_p - \boldsymbol{U}^\top\boldsymbol{U})\boldsymbol{S}(\boldsymbol{I}_p - \boldsymbol{U}^\top\boldsymbol{U})\right]. \tag{223}$$

Expanding the product and using cyclicity of trace,

$$\operatorname{Tr}(\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{S}) = \operatorname{Tr}(\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\top) = \operatorname{Tr}\boldsymbol{H} \tag{224}$$

and

$$\operatorname{Tr}(\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\top\boldsymbol{U}) = \operatorname{Tr}(\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\top) = \operatorname{Tr}(\boldsymbol{G}\boldsymbol{H}), \tag{225}$$

which proves equation 219. □

C.3.2 The empirical scalar system and proxy dynamics

We now implement Steps 2 and 3.

For a fixed realization of $\boldsymbol{S}$, define

$$M_N(q) = M_{\boldsymbol{S}}(q) = \frac{1}{p}\operatorname{Tr}\left(e^{q\boldsymbol{S}}\right) \tag{226}$$

and

$$m_N(q) = \frac{M_N'(q)}{M_N(q)} = \frac{\operatorname{Tr}(\boldsymbol{S}e^{q\boldsymbol{S}})}{\operatorname{Tr}(e^{q\boldsymbol{S}})}. \tag{227}$$

The scalar pair (a_N, q_N) is defined by

$$\begin{aligned} &\dot{a}_N(t) = 4a_N(t)m_N(q_N(t))(1 - a_N(t)), \\ &\dot{q}_N(t) = 2(2 - a_N(t)), \\ &a_N(0) = \kappa_N, \qquad q_N(0) = 0. \end{aligned} \tag{228}$$

We then set

$$r_N(t) = a_N(t)m_N(q_N(t)), \tag{229}$$

$$R_N(t) = \frac{1}{2}q_N(t), \tag{230}$$

and

$$B_N(t) = \int_0^t r_N(s)\,ds. \tag{231}$$

Lemma C.14 (Scalar identities and bounds)**.** *On the event*

$$\|\boldsymbol{S}\|_{\mathrm{op}} \le L, \tag{232}$$

the solution of equation 228 exists globally and satisfies

$$\min\{1, \kappa_N\} \le a_N(t) \le \max\{1, \kappa_N\}, \qquad t \ge 0. \tag{233}$$

Moreover,

$$a_N(t) = \kappa_N e^{-2B_N(t)} M_N(q_N(t)) \tag{234}$$

and

$$r_N(t) = \kappa_N e^{-2B_N(t)} M_N'(q_N(t)). \tag{235}$$

For every fixed $T < \infty$, there exists a deterministic $C_T < \infty$ such that, with probability tending to one,

$$\sup_{0 \le t \le T} (|a_N(t)| + |q_N(t)| + |r_N(t)| + |B_N(t)|) \le C_T. \tag{236}$$

Proof. Since $\boldsymbol{S} \succeq 0$,

$$0 \le m_N(q) \le \|\boldsymbol{S}\|_{\mathrm{op}} \le L. \tag{237}$$

Hence

$$a_N < 1 \implies \dot{a}_N \ge 0, \qquad a_N > 1 \implies \dot{a}_N \le 0, \tag{238}$$

which proves equation 233.

Define

$$\Phi_N(t) = \log a_N(t) + 2B_N(t) - \log M_N(q_N(t)). \tag{239}$$

Then

$$\begin{aligned} \dot{\Phi}_N(t) &= 4m_N(q_N)(1 - a_N) + 2a_N m_N(q_N) \\ &\quad - 2(2 - a_N)m_N(q_N) \\ &= 0. \end{aligned} \tag{240}$$

Since

$$M_N(0) = 1, \qquad \Phi_N(0) = \log \kappa_N, \tag{241}$$

we obtain equation 234. Multiplying by $m_N = M_N'/M_N$ gives equation 235.

Finally, equation 233 and equation 237 imply

$$|\dot{q}_N(t)| \le 2\,(2 + \max\{1, \kappa_N\}) \tag{242}$$

and

$$0 \le r_N(t) \le L \max\{1, \kappa_N\}. \tag{243}$$

Integration gives equation 236. □

Define the proxy trajectory

$$\overline{\boldsymbol{U}}_N(t) = e^{-B_N(t)} \boldsymbol{U}(0) e^{R_N(t)\boldsymbol{S}}. \tag{244}$$

Differentiation gives

$$\dot{\overline{\boldsymbol{U}}}_N = (2 - a_N)\overline{\boldsymbol{U}}_N \boldsymbol{S} - r_N \overline{\boldsymbol{U}}_N. \tag{245}$$

Thus $\overline{\boldsymbol{U}}_N$ is exactly the dynamics obtained by replacing the true Gram matrices $\boldsymbol{G}, \boldsymbol{H}$ by $a_N \boldsymbol{I}_n, r_N \boldsymbol{I}_n$.

C.3.3 SELF-CONSISTENCY OF THE PROXY AND STABILITY OF THE EXACT FLOW

We now carry out Step 4. Conditional on $\boldsymbol{S}$, the matrices in this subsection are functions only of the independent Gaussian initialization $\boldsymbol{Z}$.

We first prove a uniform concentration result.

Lemma C.15 (Uniform weighted Gaussian Gram concentration)**.** *Let $\boldsymbol{Z}_N \in \mathbb{R}^{n_N \times p_N}$ have independent standard Gaussian entries, and suppose*

$$\frac{n_N}{p_N} \longrightarrow 0. \tag{246}$$

Let $\boldsymbol{C}_N(t)$, $0 \le t \le T$, be deterministic symmetric matrices satisfying

$$\sup_N \sup_{0\le t\le T} \|\boldsymbol{C}_N(t)\|_{\mathrm{op}} \le K \tag{247}$$

and

$$\sup_N \sup_{\substack{s,t\in[0,T]\\ s\neq t}} \frac{\|\boldsymbol{C}_N(t)-\boldsymbol{C}_N(s)\|_{\mathrm{op}}}{|t-s|} \le K_1. \tag{248}$$

Then

$$\sup_{0\le t\le T} \left\| \frac{1}{p_N}\boldsymbol{Z}_N\boldsymbol{C}_N(t)\boldsymbol{Z}_N^\top - \frac{\operatorname{Tr}\boldsymbol{C}_N(t)}{p_N}\boldsymbol{I}_{n_N}\right\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{249}$$

Proof. Fix first t and write

$$\boldsymbol{D}_N = \frac{1}{p}\boldsymbol{Z}\boldsymbol{C}\boldsymbol{Z}^\top - \frac{\operatorname{Tr}\boldsymbol{C}}{p}\boldsymbol{I}_n. \tag{250}$$

For a deterministic unit vector $\boldsymbol{v} \in \mathbb{R}^n$,

$$\boldsymbol{Z}^\top\boldsymbol{v} \overset{d}{=} \boldsymbol{g}, \qquad \boldsymbol{g} \sim N(0, \boldsymbol{I}_p), \tag{251}$$

and therefore

$$\boldsymbol{v}^\top\boldsymbol{D}_N\boldsymbol{v} = \frac{1}{p}\left(\boldsymbol{g}^\top\boldsymbol{C}\boldsymbol{g} - \operatorname{Tr}\boldsymbol{C}\right). \tag{252}$$

Writing the eigenvalues of $\boldsymbol{C}$ as $\lambda_1, \dots, \lambda_p$,

$$\boldsymbol{g}^\top\boldsymbol{C}\boldsymbol{g} - \operatorname{Tr}\boldsymbol{C} = \sum_{i=1}^p \lambda_i(\boldsymbol{g}_i^2 - 1). \tag{253}$$

For

$$|s| \le \frac{1}{4K}, \tag{254}$$

the Gaussian moment-generating function and the elementary inequality for $-u - \frac{1}{2}\log(1-2u)$ give

$$\log \mathbb{E}\exp\left[s(\boldsymbol{g}^\top\boldsymbol{C}\boldsymbol{g} - \operatorname{Tr}\boldsymbol{C})\right] \le 2s^2\|\boldsymbol{C}\|_F^2 \le 2s^2pK^2. \tag{255}$$

A Chernoff bound therefore yields a constant $c > 0$ such that

$$\mathbb{P}\left(|\boldsymbol{g}^\top\boldsymbol{C}\boldsymbol{g} - \operatorname{Tr}\boldsymbol{C}| > p\varepsilon\right) \le 2\exp\left[-cp\min\left\{\frac{\varepsilon^2}{K^2}, \frac{\varepsilon}{K}\right\}\right]. \tag{256}$$

Let $\mathcal{N}$ be a $1/4$-net of the unit sphere $\mathbb{S}^{n-1}$ satisfying

$$|\mathcal{N}| \le 9^n. \tag{257}$$

For every symmetric matrix $\boldsymbol{D}$,

$$\|\boldsymbol{D}\|_{\mathrm{op}} \le 2\max_{\boldsymbol{v}\in\mathcal{N}} |\boldsymbol{v}^\top\boldsymbol{D}\boldsymbol{v}|. \tag{258}$$

Hence

$$\mathbb{P}\left(\|\boldsymbol{D}_N\|_{\mathrm{op}} > 2\varepsilon\right) \le 2\exp\left[n\log 9 - c_\varepsilon p\right], \tag{259}$$

which tends to zero because $n/p \to 0$.

To obtain uniformity in time, define

$$\boldsymbol{D}_N(t) = \frac{1}{p}\boldsymbol{Z}\boldsymbol{C}_N(t)\boldsymbol{Z}^\top - \frac{\operatorname{Tr}\boldsymbol{C}_N(t)}{p}\boldsymbol{I}_n. \tag{260}$$

For $s, t \in [0, T]$,

$$\|\boldsymbol{D}_N(t) - \boldsymbol{D}_N(s)\|_{\mathrm{op}} \le \left(\frac{\|\boldsymbol{Z}\|_{\mathrm{op}}^2}{p} + 1\right) K_1|t - s|. \tag{261}$$

The preceding fixed-matrix argument applied to $\boldsymbol{C} = \boldsymbol{I}_p$ gives

$$\frac{\|\boldsymbol{Z}\|_{\mathrm{op}}^2}{p} = O_{\mathbb{P}}(1). \tag{262}$$

We now make the passage from fixed time to uniform time explicit. Fix $\varepsilon > 0$ and $\delta > 0$. By equation 262, there exists a deterministic $M < \infty$ such that, for all sufficiently large N,

$$\mathbb{P}\left(\left(\frac{\|\boldsymbol{Z}\|_{\mathrm{op}}^2}{p} + 1\right) K_1 > M\right) < \delta. \tag{263}$$

Choose a deterministic mesh

$$0 = t_0 < t_1 < \cdots < t_J = T, \qquad \max_{0 \le j < J}(t_{j+1} - t_j) \le \frac{\varepsilon}{4M}. \tag{264}$$

The number J depends only on T, M, ε, and in particular is independent of N. A union bound and equation 259 therefore give

$$\mathbb{P}\left(\max_{0 \le j \le J} \|\boldsymbol{D}_N(t_j)\|_{\mathrm{op}} > \frac{\varepsilon}{2}\right) \longrightarrow 0. \tag{265}$$

On the complement of the event in equation 263, every $t \in [0, T]$ has a mesh point t_j for which equation 261 implies

$$\|\boldsymbol{D}_N(t) - \boldsymbol{D}_N(t_j)\|_{\mathrm{op}} \le \frac{\varepsilon}{4}. \tag{266}$$

Consequently,

$$\limsup_{N \to \infty} \mathbb{P}\left(\sup_{0 \le t \le T} \|\boldsymbol{D}_N(t)\|_{\mathrm{op}} > \varepsilon\right) \le \delta. \tag{267}$$

Since $\delta > 0$ is arbitrary, this proves equation 249. □

For the proxy trajectory,

$$\overline{\boldsymbol{G}}_N = \overline{\boldsymbol{U}}_N\overline{\boldsymbol{U}}_N^\top = \frac{1}{p}\boldsymbol{Z}\boldsymbol{C}_{G,N}(t)\boldsymbol{Z}^\top, \tag{268}$$

where

$$\boldsymbol{C}_{G,N}(t) = \kappa_N e^{-2B_N(t)} e^{q_N(t)\boldsymbol{S}}. \tag{269}$$

By equation 234,

$$\frac{1}{p}\operatorname{Tr}\boldsymbol{C}_{G,N}(t) = a_N(t). \tag{270}$$

Similarly,

$$\overline{\boldsymbol{H}}_N = \overline{\boldsymbol{U}}_N\boldsymbol{S}\overline{\boldsymbol{U}}_N^\top = \frac{1}{p}\boldsymbol{Z}\boldsymbol{C}_{H,N}(t)\boldsymbol{Z}^\top, \tag{271}$$

where

$$\boldsymbol{C}_{H,N}(t) = \kappa_N e^{-2B_N(t)}\boldsymbol{S}e^{q_N(t)\boldsymbol{S}}, \tag{272}$$

and

$$\frac{1}{p}\operatorname{Tr}\boldsymbol{C}_{H,N}(t) = r_N(t). \tag{273}$$

On the event $\|\boldsymbol{S}\|_{\mathrm{op}} \le L$, Lemma C.14 implies uniform bounds on q_N, B_N, a_N, r_N on every fixed time interval. Consequently, for every fixed T,

$$\sup_{0 \le t \le T}(\|\boldsymbol{C}_{G,N}(t)\|_{\mathrm{op}} + \|\boldsymbol{C}_{H,N}(t)\|_{\mathrm{op}}) \le C_T \tag{274}$$

with probability tending to one. Differentiating equation 269–equation 272 and using the same bounds gives

$$\sup_{\substack{s,t\in[0,T]\\ s\neq t}} \frac{\|\boldsymbol{C}_{G,N}(t)-\boldsymbol{C}_{G,N}(s)\|_{\mathrm{op}}+\|\boldsymbol{C}_{H,N}(t)-\boldsymbol{C}_{H,N}(s)\|_{\mathrm{op}}}{|t-s|}\le C_T \tag{275}$$

with probability tending to one. Conditional on $\boldsymbol{S}$, we may therefore apply Lemma C.15 to obtain

$$\sup_{0\le t\le T}\|\overline{\boldsymbol{G}}_N(t)-a_N(t)\boldsymbol{I}_n\|_{\mathrm{op}}\xrightarrow{\mathbb{P}}0 \tag{276}$$

and

$$\sup_{0\le t\le T}\|\overline{\boldsymbol{H}}_N(t)-r_N(t)\boldsymbol{I}_n\|_{\mathrm{op}}\xrightarrow{\mathbb{P}}0. \tag{277}$$

We now return to the comparison introduced in Step 3.

Theorem C.16 (Quenched narrow-latent approximation)**.** *Assume equation 189 and equation 41. Then, for every fixed $T<\infty$,*

$$\sup_{0\le t\le T}\|\boldsymbol{U}_N(t)-\overline{\boldsymbol{U}}_N(t)\|_{\mathrm{op}}\xrightarrow{\mathbb{P}}0. \tag{278}$$

Consequently,

$$\sup_{0\le t\le T}\|\boldsymbol{G}_N(t)-a_N(t)\boldsymbol{I}_n\|_{\mathrm{op}}\xrightarrow{\mathbb{P}}0 \tag{279}$$

and

$$\sup_{0\le t\le T}\|\boldsymbol{H}_N(t)-r_N(t)\boldsymbol{I}_n\|_{\mathrm{op}}\xrightarrow{\mathbb{P}}0. \tag{280}$$

Proof. Define the exact vector field

$$F_{\boldsymbol{S}}(\boldsymbol{V})=2\boldsymbol{V}\boldsymbol{S}-(\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^\top)\boldsymbol{V}-(\boldsymbol{V}\boldsymbol{V}^\top)\boldsymbol{V}\boldsymbol{S}. \tag{281}$$

By equation 245,

$$\begin{aligned}F_{\boldsymbol{S}}(\overline{\boldsymbol{U}}_N)-\dot{\overline{\boldsymbol{U}}}_N=&-(\overline{\boldsymbol{H}}_N-r_N\boldsymbol{I}_n)\overline{\boldsymbol{U}}_N\\&-(\overline{\boldsymbol{G}}_N-a_N\boldsymbol{I}_n)\overline{\boldsymbol{U}}_N\boldsymbol{S}.\end{aligned} \tag{282}$$

Thus

$$\begin{aligned}\|F_{\boldsymbol{S}}(\overline{\boldsymbol{U}}_N)-\dot{\overline{\boldsymbol{U}}}_N\|_{\mathrm{op}}\le&\|\overline{\boldsymbol{H}}_N-r_N\boldsymbol{I}_n\|_{\mathrm{op}}\|\overline{\boldsymbol{U}}_N\|_{\mathrm{op}}\\&+\|\overline{\boldsymbol{G}}_N-a_N\boldsymbol{I}_n\|_{\mathrm{op}}\|\overline{\boldsymbol{U}}_N\|_{\mathrm{op}}\|\boldsymbol{S}\|_{\mathrm{op}}.\end{aligned} \tag{283}$$

Since

$$\|\overline{\boldsymbol{U}}_N(t)\|_{\mathrm{op}}^2=\|\overline{\boldsymbol{G}}_N(t)\|_{\mathrm{op}}, \tag{284}$$

Equation (276) and the uniform boundedness of a_N imply

$$\sup_{0\le t\le T}\|\overline{\boldsymbol{U}}_N(t)\|_{\mathrm{op}}=O_{\mathbb{P}}(1). \tag{285}$$

Hence

$$\varepsilon_N(T):=\sup_{0\le t\le T}\|F_{\boldsymbol{S}}(\overline{\boldsymbol{U}}_N(t))-\dot{\overline{\boldsymbol{U}}}_N(t)\|_{\mathrm{op}}\xrightarrow{\mathbb{P}}0. \tag{286}$$

It remains to prove stability of the exact ODE. Suppose

$$\|\boldsymbol{V}\|_{\mathrm{op}}\le M,\qquad\|\boldsymbol{W}\|_{\mathrm{op}}\le M,\qquad\|\boldsymbol{S}\|_{\mathrm{op}}\le L. \tag{287}$$

Then

$$\|\boldsymbol{V}\boldsymbol{V}^\top-\boldsymbol{W}\boldsymbol{W}^\top\|_{\mathrm{op}}\le 2M\|\boldsymbol{V}-\boldsymbol{W}\|_{\mathrm{op}} \tag{288}$$

and

$$\|\boldsymbol{V}\boldsymbol{S}\boldsymbol{V}^\top-\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}^\top\|_{\mathrm{op}}\le 2LM\|\boldsymbol{V}-\boldsymbol{W}\|_{\mathrm{op}}. \tag{289}$$

Expanding equation 281 therefore gives

$$\|F_{\boldsymbol{S}}(\boldsymbol{V})-F_{\boldsymbol{S}}(\boldsymbol{W})\|_{\mathrm{op}}\le\left(2L+6LM^2\right)\|\boldsymbol{V}-\boldsymbol{W}\|_{\mathrm{op}}. \tag{290}$$

Set

$$\boldsymbol{E}_N(t) = \boldsymbol{U}_N(t) - \overline{\boldsymbol{U}}_N(t) \tag{291}$$

and introduce the proxy residual

$$\boldsymbol{R}_N(t) := F_{\boldsymbol{S}}(\overline{\boldsymbol{U}}_N(t)) - \dot{\overline{\boldsymbol{U}}}_N(t). \tag{292}$$

By definition of $\varepsilon_N(T)$,

$$\sup_{0\le t\le T} \|\boldsymbol{R}_N(t)\|_{\mathrm{op}} \le \varepsilon_N(T). \tag{293}$$

Moreover, the exact and proxy trajectories start from the same matrix, so

$$\boldsymbol{E}_N(0) = 0. \tag{294}$$

Subtracting their differential equations gives the exact error equation

$$\dot{\boldsymbol{E}}_N(t) = F_{\boldsymbol{S}}(\boldsymbol{U}_N(t)) - F_{\boldsymbol{S}}(\overline{\boldsymbol{U}}_N(t)) + \boldsymbol{R}_N(t). \tag{295}$$

To make the bounded-set argument explicit, choose deterministic $L, M < \infty$ and consider

$$\mathcal{A}_N(L,M) := \left\{ \|\boldsymbol{S}_N\|_{\mathrm{op}} \le L,\ \sup_{0\le t\le T} \|\boldsymbol{U}_N(t)\|_{\mathrm{op}} \le M,\ \sup_{0\le t\le T} \|\overline{\boldsymbol{U}}_N(t)\|_{\mathrm{op}} \le M \right\}. \tag{296}$$

The standing spectral bound controls the first term. For the exact trajectory, Lemma C.2 gives

$$\sup_{0\le t\le T} \|\boldsymbol{U}_N(t)\|_{\mathrm{op}}^2 \le \max\left\{1, \|\boldsymbol{U}_N(0)\|_{\mathrm{op}}^2\right\}, \tag{297}$$

and the common Gaussian initialization with $n/p \to 0$ makes the right-hand side tight. The proxy term is tight by Eq. (285). Hence L and M may be chosen so that

$$\mathbb{P}\big(\mathcal{A}_N(L,M)\big) \longrightarrow 1. \tag{298}$$

On $\mathcal{A}_N(L,M)$, the Lipschitz estimate Eq. (290) holds for the two trajectories at every $t \in [0,T]$. Integrating Eq. (295), using Eq. (293), and taking operator norms therefore gives

$$\|\boldsymbol{E}_N(t)\|_{\mathrm{op}} \le C_T \int_0^t \|\boldsymbol{E}_N(s)\|_{\mathrm{op}}\, ds + t\varepsilon_N(T), \qquad 0 \le t \le T, \tag{299}$$

where one may take $C_T = 2L + 6LM^2$. Gronwall's inequality now yields

$$\sup_{0\le t\le T} \|\boldsymbol{E}_N(t)\|_{\mathrm{op}} \le T e^{C_T T} \varepsilon_N(T) \tag{300}$$

on $\mathcal{A}_N(L,M)$. Combining Eqs. (286), (298) and (300) gives

$$\sup_{0\le t\le T} \|\boldsymbol{E}_N(t)\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0, \tag{301}$$

which proves Eq. (278).

Finally,

$$\|\boldsymbol{U}_N\boldsymbol{U}_N^\top - \overline{\boldsymbol{U}}_N\overline{\boldsymbol{U}}_N^\top\|_{\mathrm{op}} \le \big(\|\boldsymbol{U}_N\|_{\mathrm{op}} + \|\overline{\boldsymbol{U}}_N\|_{\mathrm{op}}\big)\, \|\boldsymbol{U}_N - \overline{\boldsymbol{U}}_N\|_{\mathrm{op}} \tag{302}$$

and

$$\begin{aligned} &\|\boldsymbol{U}_N\boldsymbol{S}\boldsymbol{U}_N^\top - \overline{\boldsymbol{U}}_N\boldsymbol{S}\overline{\boldsymbol{U}}_N^\top\|_{\mathrm{op}} \\ &\quad\le \|\boldsymbol{S}\|_{\mathrm{op}} \big(\|\boldsymbol{U}_N\|_{\mathrm{op}} + \|\overline{\boldsymbol{U}}_N\|_{\mathrm{op}}\big)\, \|\boldsymbol{U}_N - \overline{\boldsymbol{U}}_N\|_{\mathrm{op}}. \end{aligned} \tag{303}$$

Combining these estimates with equation 276–equation 277 proves equation 279–equation 280. □

Theorem C.16 completes Steps 2–4 of the roadmap. It is a quenched statement: conditional on a data covariance $\boldsymbol{S}$ with bounded spectrum, the random Gaussian initialization produces asymptotically scalar hidden-space Gram matrices. No limiting spectral distribution, and in particular no Marchenko–Pastur law, has yet been used.

C.3.4 PASSAGE TO A GENERAL LIMITING SPECTRAL MEASURE

We now implement Step 5 in detail.

The deterministic limiting spectral measure ν_ϕ and the weak-convergence assumption are those stated at the beginning of this subsection.

$$M'_{\nu_\phi}(q) = \int \lambda e^{q\lambda}\, d\nu_\phi(\lambda), \tag{304}$$

$$M''_{\nu_\phi}(q) = \int \lambda^2 e^{q\lambda}\, d\nu_\phi(\lambda), \tag{305}$$

We first show that the empirical transforms converge uniformly on compact q-intervals.

Lemma C.17 (Uniform convergence of the spectral transforms)**.** *For every fixed $Q < \infty$,*

$$\sup_{|q|\le Q} |M_N(q) - M_{\nu_\phi}(q)| \xrightarrow{\mathbb{P}} 0, \tag{306}$$

$$\sup_{|q|\le Q} |M'_N(q) - M'_{\nu_\phi}(q)| \xrightarrow{\mathbb{P}} 0, \tag{307}$$

$$\sup_{|q|\le Q} |M''_N(q) - M''_{\nu_\phi}(q)| \xrightarrow{\mathbb{P}} 0, \tag{308}$$

and

$$\sup_{|q|\le Q} |m_N(q) - m_{\nu_\phi}(q)| \xrightarrow{\mathbb{P}} 0. \tag{309}$$

Proof. Fix $Q < \infty$ and work on the event

$$\mathcal{E}_N = \{\|\boldsymbol{S}_N\|_{\mathrm{op}} \le L\}. \tag{310}$$

Its probability tends to one.

For every fixed $q \in [-Q, Q]$, the functions

$$f_{0,q}(\lambda) = e^{q\lambda}, \qquad f_{1,q}(\lambda) = \lambda e^{q\lambda}, \qquad f_{2,q}(\lambda) = \lambda^2 e^{q\lambda} \tag{311}$$

are bounded and continuous on $[0, L]$. Weak convergence therefore gives

$$M_N(q) \xrightarrow{\mathbb{P}} M_{\nu_\phi}(q), \tag{312}$$

$$M'_N(q) \xrightarrow{\mathbb{P}} M'_{\nu_\phi}(q), \tag{313}$$

and

$$M''_N(q) \xrightarrow{\mathbb{P}} M''_{\nu_\phi}(q). \tag{314}$$

On $\mathcal{E}_N$, the mean-value theorem gives

$$|M_N(q) - M_N(\widetilde{q})| \le L e^{QL} |q - \widetilde{q}|, \tag{315}$$

$$|M'_N(q) - M'_N(\widetilde{q})| \le L^2 e^{QL} |q - \widetilde{q}|, \tag{316}$$

and

$$|M''_N(q) - M''_N(\widetilde{q})| \le L^3 e^{QL} |q - \widetilde{q}|. \tag{317}$$

The same bounds hold for the limiting transforms.

Fix $\varepsilon > 0$ and choose a deterministic finite grid

$$-Q = q_0 < q_1 < \cdots < q_J = Q \tag{318}$$

whose mesh is sufficiently small that the Lipschitz errors in equation 315–equation 317 are all at most $\varepsilon/4$. For every $q \in [-Q, Q]$, let q_j be a nearest grid point. Then, for example,

$$\begin{aligned} |M_N(q) - M_{\nu_\phi}(q)| \le\ & |M_N(q) - M_N(q_j)| \\ & + |M_N(q_j) - M_{\nu_\phi}(q_j)| \\ & + |M_{\nu_\phi}(q_j) - M_{\nu_\phi}(q)|. \end{aligned} \tag{319}$$

Thus

$$\sup_{|q|\le Q} |M_N(q) - M_{\nu_\phi}(q)| \le \max_{0\le j\le J} |M_N(q_j) - M_{\nu_\phi}(q_j)| + \frac{\varepsilon}{2} \tag{320}$$

on $\mathcal{E}_N$. Since the grid is finite, equation 312 and a union bound imply equation 306. The same argument proves equation 307 and equation 308.

To control the ratio, note that for $|q| \le Q$,

$$M_N(q) \ge e^{-QL} \tag{321}$$

on $\mathcal{E}_N$, and

$$M_{\nu_\phi}(q) \ge e^{-QL}. \tag{322}$$

Therefore

$$\begin{aligned} |m_N(q) - m_{\nu_\phi}(q)| \le e^{QL}|M'_N(q) - M'_{\nu_\phi}(q)| \\ + Le^{3QL}|M_N(q) - M_{\nu_\phi}(q)|. \end{aligned} \tag{323}$$

Taking the supremum and using equation 306–equation 307 proves equation 309. □

The quantity m_{ν_ϕ} has a useful interpretation. Define the exponentially tilted probability measure

$$d\nu_{\phi_q}(\lambda) = \frac{e^{q\lambda}}{M_{\nu_\phi}(q)}\, d\nu_\phi(\lambda). \tag{324}$$

Then

$$m_{\nu_\phi}(q) = \int \lambda\, d\nu_{\phi_q}(\lambda). \tag{325}$$

Furthermore,

$$m'_{\nu_\phi}(q) = \frac{M''_{\nu_\phi}(q)M_{\nu_\phi}(q) - M'_{\nu_\phi}(q)^2}{M_{\nu_\phi}(q)^2} = \mathrm{Var}_{\nu_{\phi_q}}(\lambda) \ge 0. \tag{326}$$

Hence

$$0 \le m_{\nu_\phi}(q) \le L \tag{327}$$

and

$$0 \le m'_{\nu_\phi}(q) \le L^2. \tag{328}$$

$$R(t) = \frac{1}{2}q(t), \tag{329}$$

and

$$B(t) = \int_0^t r(s)\, ds. \tag{330}$$

We first obtain a common compact set containing the empirical and limiting trajectories.

Lemma C.18 (Compactness of the scalar trajectories)**.** *For every fixed $T < \infty$, there exist deterministic constants $a_- > 0$, $a_+ < \infty$, and $Q_T < \infty$ such that, with probability tending to one,*

$$a_N(t) \in [a_-, a_+], \qquad |q_N(t)| \le Q_T, \qquad 0 \le t \le T, \tag{331}$$

while deterministically

$$a(t) \in [a_-, a_+], \qquad |q(t)| \le Q_T, \qquad 0 \le t \le T. \tag{332}$$

Proof. Since $\kappa_N \to \kappa \in (0, \infty)$, there exist deterministic constants

$$0 < \kappa_- < \kappa_+ < \infty \tag{333}$$

such that

$$\kappa_- \le \kappa_N \le \kappa_+ \tag{334}$$

for all sufficiently large N.

The barrier argument used in Lemma C.14 gives

$$\min\{1, \kappa_N\} \le a_N(t) \le \max\{1, \kappa_N\} \tag{335}$$

and

$$\min\{1,\kappa\} \le a(t) \le \max\{1,\kappa\}. \tag{336}$$

Thus a_- and a_+ may be chosen independently of N.

Moreover,

$$|\dot{q}_N(t)| \le 2(2+a_+) \tag{337}$$

and

$$|\dot{q}(t)| \le 2(2+a_+). \tag{338}$$

Since both $q_N(0)$ and $q(0)$ vanish, one may take

$$Q_T = 2(2+a_+)T. \tag{339}$$

□

We can now derive an explicit perturbation estimate.

Proposition C.19 (Convergence of the scalar dynamics)**.** *For every fixed $T < \infty$,*

$$\sup_{0\le t\le T} \left(|a_N(t) - a(t)| + |q_N(t) - q(t)|\right) \xrightarrow{\mathbb{P}} 0. \tag{340}$$

Proof. Let Q_T be as in Lemma C.18, and define

$$\delta_N(T) = \sup_{|q|\le Q_T} |m_N(q) - m_{\nu_\phi}(q)|. \tag{341}$$

By Lemma C.17,

$$\delta_N(T) \xrightarrow{\mathbb{P}} 0. \tag{342}$$

Set

$$\Delta a_N(t) = a_N(t) - a(t) \tag{343}$$

and

$$\Delta q_N(t) = q_N(t) - q(t). \tag{344}$$

Subtracting the q-equations gives

$$\dot{\Delta} q_N(t) = -2\Delta a_N(t), \tag{345}$$

hence

$$|\Delta q_N(t)| \le 2\int_0^t |\Delta a_N(s)|\, ds. \tag{346}$$

Define

$$F_N(a,q) = 4a(1-a)m_N(q) \tag{347}$$

and

$$F(a,q) = 4a(1-a)m_{\nu_\phi}(q). \tag{348}$$

Then

$$\dot{\Delta} a_N(t) = F_N(a_N(t), q_N(t)) - F(a(t), q(t)). \tag{349}$$

Adding and subtracting $F(a_N, q_N)$,

$$\begin{aligned} |\dot{\Delta} a_N| \le\ & |F_N(a_N, q_N) - F(a_N, q_N)| \\ & + |F(a_N, q_N) - F(a, q)|. \end{aligned} \tag{350}$$

Since $a_N \in [a_-, a_+]$, there exists $C_a < \infty$ such that

$$|a_N(1-a_N)| \le C_a. \tag{351}$$

Therefore

$$|F_N(a_N, q_N) - F(a_N, q_N)| \le 4C_a \delta_N(T). \tag{352}$$

The derivatives of the limiting vector field are

$$\partial_a F(a,q) = 4(1-2a)m_{\nu_\phi}(q) \tag{353}$$

and

$$\partial_q F(a,q) = 4a(1-a)m'_{\nu_\phi}(q). \tag{354}$$

Using equation 327, equation 328, and the compactness of the a coordinate, there is a deterministic $C_T < \infty$ such that

$$|F(a_N, q_N) - F(a,q)| \le C_T \left(|\Delta a_N| + |\Delta q_N|\right). \tag{355}$$

Hence

$$|\dot{\Delta} a_N(t)| \le 4C_a \delta_N(T) + C_T \left(|\Delta a_N(t)| + |\Delta q_N(t)|\right). \tag{356}$$

Set

$$D_N(t) = |\Delta a_N(t)| + |\Delta q_N(t)|. \tag{357}$$

Since

$$D_N(0) = |\kappa_N - \kappa|, \tag{358}$$

integration of equation 345 and equation 356 gives

$$D_N(t) \le |\kappa_N - \kappa| + 4C_a T \delta_N(T) + C'_T \int_0^t D_N(s)\, ds \tag{359}$$

for a deterministic constant $C'_T < \infty$. Therefore

$$\sup_{0\le t\le T} D_N(t) \le e^{C'_T T} \left[|\kappa_N - \kappa| + 4C_a T \delta_N(T)\right]. \tag{360}$$

The right-hand side tends to zero in probability, proving equation 340. □

The remaining scalar quantities converge as well.

Proposition C.20 (Convergence of r_N, R_N, B_N)**.** *For every fixed $T < \infty$,*

$$\sup_{0\le t\le T} |r_N(t) - r(t)| \xrightarrow{\mathbb{P}} 0, \tag{361}$$

$$\sup_{0\le t\le T} |R_N(t) - R(t)| \xrightarrow{\mathbb{P}} 0, \tag{362}$$

and

$$\sup_{0\le t\le T} |B_N(t) - B(t)| \xrightarrow{\mathbb{P}} 0. \tag{363}$$

Proof. From

$$r_N = a_N m_N(q_N), \qquad r = a m_{\nu_\phi}(q), \tag{364}$$

we have

$$\begin{aligned} |r_N - r| \le L|a_N - a| + a_+ |m_N(q_N) - m_{\nu_\phi}(q_N)| \\ + a_+ |m_{\nu_\phi}(q_N) - m_{\nu_\phi}(q)|. \end{aligned} \tag{365}$$

By equation 341, equation 328, and Proposition C.19,

$$|r_N - r| \le L|a_N - a| + a_+ \delta_N(T) + a_+ L^2 |q_N - q|, \tag{366}$$

which proves equation 361.

Since

$$R_N = \frac{1}{2} q_N, \qquad R = \frac{1}{2} q, \tag{367}$$

Equation (362) follows immediately.

Finally,

$$|B_N(t) - B(t)| \le \int_0^t |r_N(s) - r(s)|\, ds \tag{368}$$

and hence

$$\sup_{0\le t\le T} |B_N(t) - B(t)| \le T \sup_{0\le t\le T} |r_N(t) - r(t)|, \tag{369}$$

which proves equation 363. □

The limiting system admits the self-consistent representation

$$a(t) = \kappa e^{-2B(t)} M_{\nu_\phi}(q(t)) \tag{370}$$

and

$$r(t) = \kappa e^{-2B(t)} M'_{\nu_\phi}(q(t)). \tag{371}$$

Indeed, if

$$\Phi(t) = \log a(t) + 2B(t) - \log M_{\nu_\phi}(q(t)), \tag{372}$$

then

$$\begin{aligned} \dot{\Phi}(t) &= 4m_{\nu_\phi}(q)(1-a) + 2am_{\nu_\phi}(q) \\ &\quad - 2(2-a)m_{\nu_\phi}(q) \\ &= 0. \end{aligned} \tag{373}$$

Since

$$\Phi(0) = \log \kappa, \tag{374}$$

Eqs. (370) to (371) follow.

Equivalently, using $q = 2R$,

$$a(t) = \kappa e^{-2B(t)} M_{\nu_\phi}(2R(t)), \tag{375}$$

$$r(t) = \kappa e^{-2B(t)} M'_{\nu_\phi}(2R(t)), \tag{376}$$

$$\dot{R}(t) = 2 - a(t), \tag{377}$$

and

$$\dot{B}(t) = r(t). \tag{378}$$

Combining the quenched theorem with the scalar convergence proves the main finite-time result stated at the beginning of this subsection.

Proof of Theorem C.12. By Theorem C.16,

$$\sup_{0\le t\le T} \|\boldsymbol{G}_N(t) - a_N(t)\boldsymbol{I}_n\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{379}$$

Hence

$$\begin{aligned} \sup_{0\le t\le T} \|\boldsymbol{G}_N(t) - a(t)\boldsymbol{I}_n\|_{\mathrm{op}} \le &\sup_{0\le t\le T} \|\boldsymbol{G}_N(t) - a_N(t)\boldsymbol{I}_n\|_{\mathrm{op}} \\ &+ \sup_{0\le t\le T} |a_N(t) - a(t)|. \end{aligned} \tag{380}$$

The two terms tend to zero by Theorem C.16 and Proposition C.19, respectively. This proves equation 199. The same argument using equation 280 and equation 361 proves equation 200. □

Theorem C.12 is the precise implementation of the five-step roadmap stated at the beginning of the section. The dynamics first produces a data-dependent scalar approximation a_N, r_N; only afterward is the empirical spectral measure replaced by its deterministic limit ν_ϕ.

C.3.5 Consequences for population and train losses

We now return to Step 1 and substitute the scalarization obtained above into the exact loss identities.

Corollary C.21 (Population loss)**.** *Under the assumptions of Theorem C.12, for every fixed $T < \infty$,*

$$\sup_{0\le t\le T} \left| \psi_N\left(L_N(t) - \frac{1}{2}\right) - a(t)\left(\frac{a(t)}{2} - 1\right)\right| \xrightarrow{\mathbb{P}} 0. \tag{381}$$

Proof. From equation 218,

$$\psi_N\left(L_N - \frac{1}{2}\right) = -\frac{1}{n}\operatorname{Tr}\boldsymbol{G}_N + \frac{1}{2n}\operatorname{Tr}\boldsymbol{G}_N^2. \tag{382}$$

Equation (199) implies

$$\left|\frac{1}{n}\operatorname{Tr}\boldsymbol{G}_N - a\right| \le \|\boldsymbol{G}_N - a\boldsymbol{I}_n\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0 \tag{383}$$

uniformly on $0 \le t \le T$. Furthermore,

$$\|\boldsymbol{G}_N^2 - a^2\boldsymbol{I}_n\|_{\mathrm{op}} \le \|\boldsymbol{G}_N - a\boldsymbol{I}_n\|_{\mathrm{op}} \left(\|\boldsymbol{G}_N\|_{\mathrm{op}} + |a|\right), \tag{384}$$

so

$$\frac{1}{n}\operatorname{Tr}\boldsymbol{G}_N^2 \xrightarrow{\mathbb{P}} a^2 \tag{385}$$

uniformly on compact time intervals. Substitution into equation 382 gives equation 381. □

Corollary C.22 (Train loss)**.** *Under the assumptions of Theorem C.12, for every fixed $T < \infty$,*

$$\sup_{0\le t\le T}\left|\psi_N\left(\widehat{L}_N(t) - \frac{1}{2p}\operatorname{Tr}\boldsymbol{S}_N\right) - r(t)\left(\frac{a(t)}{2} - 1\right)\right| \xrightarrow{\mathbb{P}} 0. \tag{386}$$

If in addition

$$\frac{\operatorname{Tr}\boldsymbol{S}_N - p}{n} \xrightarrow{\mathbb{P}} 0, \tag{387}$$

then

$$\sup_{0\le t\le T}\left|\psi_N\left(\widehat{L}_N(t) - \frac{1}{2}\right) - r(t)\left(\frac{a(t)}{2} - 1\right)\right| \xrightarrow{\mathbb{P}} 0. \tag{388}$$

Proof. From equation 219,

$$\psi_N\left(\widehat{L}_N - \frac{1}{2p}\operatorname{Tr}\boldsymbol{S}_N\right) = -\frac{1}{n}\operatorname{Tr}\boldsymbol{H}_N + \frac{1}{2n}\operatorname{Tr}(\boldsymbol{G}_N\boldsymbol{H}_N). \tag{389}$$

Equation (200) gives

$$\frac{1}{n}\operatorname{Tr}\boldsymbol{H}_N \xrightarrow{\mathbb{P}} r. \tag{390}$$

Also,

$$\|\boldsymbol{G}_N\boldsymbol{H}_N - ar\boldsymbol{I}_n\|_{\mathrm{op}} \le \|\boldsymbol{G}_N - a\boldsymbol{I}_n\|_{\mathrm{op}}\|\boldsymbol{H}_N\|_{\mathrm{op}} + |a|\|\boldsymbol{H}_N - r\boldsymbol{I}_n\|_{\mathrm{op}}, \tag{391}$$

so

$$\frac{1}{n}\operatorname{Tr}(\boldsymbol{G}_N\boldsymbol{H}_N) \xrightarrow{\mathbb{P}} ar. \tag{392}$$

This proves equation 386. Finally, equation 387 is equivalent to

$$\psi_N\left(\frac{1}{2p}\operatorname{Tr}\boldsymbol{S}_N - \frac{1}{2}\right) \xrightarrow{\mathbb{P}} 0, \tag{393}$$

which yields equation 388. □

For the expected losses, isotropy fixes the empirical baseline:

$$\mathbb{E}\left[\frac{1}{p}\operatorname{Tr}\boldsymbol{S}_N\right] = 1. \tag{394}$$

Corollary C.23 (Expected losses, uniformly on compact time intervals)**.** *Under the assumptions of Theorem C.12 and the isotropy assumption equation 191, for every fixed $T < \infty$,*

$$\mathbb{E}\left[\sup_{0\le t\le T}\left|\psi_N\left(L_N(t) - \frac{1}{2}\right) - a(t)\left(\frac{a(t)}{2} - 1\right)\right|\right] \longrightarrow 0. \tag{395}$$

If, in addition, for some $\delta > 0$,

$$\sup_N \mathbb{E}\|\boldsymbol{S}_N\|_{\mathrm{op}}^{1+\delta} < \infty, \tag{396}$$

then

$$\mathbb{E}\left[\sup_{0\le t\le T}\left|\psi_N\left(\widehat{L}_N(t) - \frac{1}{2p}\operatorname{Tr}\boldsymbol{S}_N\right) - r(t)\left(\frac{a(t)}{2} - 1\right)\right|\right] \longrightarrow 0. \tag{397}$$

Consequently,

$$\sup_{0\le t\le T}\left|\psi_N\left(\mathbb{E}L_N(t)-\frac{1}{2}\right)-a(t)\left(\frac{a(t)}{2}-1\right)\right|\longrightarrow 0, \tag{398}$$

and, under equation 396,

$$\sup_{0\le t\le T}\left|\psi_N\left(\mathbb{E}\widehat{L}_N(t)-\frac{1}{2}\right)-r(t)\left(\frac{a(t)}{2}-1\right)\right|\longrightarrow 0. \tag{399}$$

Proof. Set

$$R_N:=\max\left\{1,\|\boldsymbol{U}_N(0)\|_{\mathrm{op}}^2\right\}. \tag{400}$$

Lemma C.2 gives, for all $t\ge 0$,

$$\|\boldsymbol{G}_N(t)\|_{\mathrm{op}}=\|\boldsymbol{U}_N(t)\|_{\mathrm{op}}^2\le R_N. \tag{401}$$

Moreover, the standing Gaussian initialization equation 36, the narrow-latent scaling equation 189, and the Gaussian singular-value bound equation 116 imply that, for every fixed $q<\infty$,

$$\sup_N \mathbb{E}R_N^q<\infty. \tag{402}$$

Indeed, $\boldsymbol{U}_N(0)=\sigma_N\boldsymbol{Z}_N$, while $p_N\sigma_N^2\to\kappa$ and $n_N/p_N\to 0$, so integration of the Gaussian operator-norm tail gives uniformly bounded moments of $\|\boldsymbol{U}_N(0)\|_{\mathrm{op}}$ of every fixed order.

For the population loss, the exact identity equation 382 and equation 401 give, uniformly in $t\ge 0$,

$$\left|\psi_N\left(L_N(t)-\frac{1}{2}\right)\right|\le\|\boldsymbol{G}_N(t)\|_{\mathrm{op}}+\frac{1}{2}\|\boldsymbol{G}_N(t)\|_{\mathrm{op}}^2\le R_N+\frac{1}{2}R_N^2. \tag{403}$$

The deterministic function $a(t)(a(t)/2-1)$ is bounded on every compact time interval. Hence equation 402 and equation 403 imply that the random variables

$$\sup_{0\le t\le T}\left|\psi_N\left(L_N(t)-\frac{1}{2}\right)-a(t)\left(\frac{a(t)}{2}-1\right)\right| \tag{404}$$

form a uniformly integrable family. By equation 381, they converge to zero in probability. Vitali's theorem therefore yields equation 395.

For the train loss, the exact identity equation 389 gives

$$\left|\psi_N\left(\widehat{L}_N(t)-\frac{1}{2p}\operatorname{Tr}\boldsymbol{S}_N\right)\right|\le\|\boldsymbol{H}_N(t)\|_{\mathrm{op}}\left(1+\frac{1}{2}\|\boldsymbol{G}_N(t)\|_{\mathrm{op}}\right). \tag{405}$$

Since

$$\|\boldsymbol{H}_N(t)\|_{\mathrm{op}}=\|\boldsymbol{U}_N(t)\boldsymbol{S}_N\boldsymbol{U}_N(t)^\top\|_{\mathrm{op}}\le\|\boldsymbol{S}_N\|_{\mathrm{op}}\|\boldsymbol{G}_N(t)\|_{\mathrm{op}}, \tag{406}$$

we obtain from equation 401

$$\left|\psi_N\left(\widehat{L}_N(t)-\frac{1}{2p}\operatorname{Tr}\boldsymbol{S}_N\right)\right|\le\|\boldsymbol{S}_N\|_{\mathrm{op}}R_N\left(1+\frac{1}{2}R_N\right). \tag{407}$$

By the standing initialization equation 36, $\boldsymbol{U}_N(0)$ is independent of $\boldsymbol{X}_N$, and hence R_N is independent of $\boldsymbol{S}_N$. Consequently, equation 396 and equation 402 imply

$$\begin{aligned}&\mathbb{E}\left[\|\boldsymbol{S}_N\|_{\mathrm{op}}R_N\left(1+\frac{1}{2}R_N\right)\right]^{1+\delta}\\&\quad=\mathbb{E}\|\boldsymbol{S}_N\|_{\mathrm{op}}^{1+\delta}\,\mathbb{E}\left[R_N\left(1+\frac{1}{2}R_N\right)\right]^{1+\delta}\le C,\end{aligned} \tag{408}$$

where $C<\infty$ is independent of N. Since $r(t)(a(t)/2-1)$ is bounded on $0\le t\le T$, it follows that the random variables

$$\sup_{0\le t\le T}\left|\psi_N\left(\widehat{L}_N(t)-\frac{1}{2p}\operatorname{Tr}\boldsymbol{S}_N\right)-r(t)\left(\frac{a(t)}{2}-1\right)\right| \tag{409}$$

are uniformly integrable. Their convergence to zero in probability is exactly equation 386. Vitali's theorem therefore gives equation 397.

Finally,

$$\begin{aligned}&\sup_{0\le t\le T}\left|\psi_N\left(\mathbb{E}L_N(t)-\frac{1}{2}\right)-a(t)\left(\frac{a(t)}{2}-1\right)\right|\\&\qquad\le\mathbb{E}\left[\sup_{0\le t\le T}\left|\psi_N\left(L_N(t)-\frac{1}{2}\right)-a(t)\left(\frac{a(t)}{2}-1\right)\right|\right],\end{aligned}\tag{410}$$

so equation 398 follows from equation 395. Similarly, equation 394 gives

$$\mathbb{E}\left[\frac{1}{2p}\operatorname{Tr}\boldsymbol{S}_N\right]=\frac{1}{2},\tag{411}$$

and therefore

$$\begin{aligned}&\sup_{0\le t\le T}\left|\psi_N\left(\mathbb{E}\widehat{L}_N(t)-\frac{1}{2}\right)-r(t)\left(\frac{a(t)}{2}-1\right)\right|\\&\qquad\le\mathbb{E}\left[\sup_{0\le t\le T}\left|\psi_N\left(\widehat{L}_N(t)-\frac{1}{2p}\operatorname{Tr}\boldsymbol{S}_N\right)-r(t)\left(\frac{a(t)}{2}-1\right)\right|\right].\end{aligned}\tag{412}$$

Eq. (399) now follows from equation 397. □

C.3.6 Large-time behavior of the limiting system

We now analyze the deterministic limiting system after taking the high-dimensional limit. The statements below concern the order of limits

$$N\longrightarrow\infty\quad\text{first},\qquad t\longrightarrow\infty\quad\text{second}.\tag{413}$$

No uniform-in-time approximation of the finite-dimensional trajectory is asserted.

For $\kappa\neq 1$, the limiting system has an exact first integral. Indeed,

$$\dot{a}=4a\,m_{\nu_\phi}(q)(1-a)\tag{414}$$

and

$$\dot{q}=2(2-a).\tag{415}$$

Eliminating time,

$$\frac{2-a}{a(1-a)}\,da=2m_{\nu_\phi}(q)\,dq.\tag{416}$$

Since

$$\frac{2-a}{a(1-a)}=\frac{2}{a}+\frac{1}{1-a},\tag{417}$$

integration gives

$$2\log a-\log|1-a|=2\log M_{\nu_\phi}(q)+C.\tag{418}$$

Using the initial condition,

$$\frac{a(t)^2}{|1-a(t)|}=\frac{\kappa^2}{|1-\kappa|}M_{\nu_\phi}(q(t))^2.\tag{419}$$

For $\kappa=1$, uniqueness gives

$$a(t)\equiv 1.\tag{420}$$

Alignment and possible null-space trapping. Let

$$w_0=\nu_\phi(\{0\}).\tag{421}$$

Because ν_ϕ is supported on the nonnegative half-line,

$$M_{\nu_\phi}(q)\longrightarrow w_0\qquad\text{as } q\to-\infty.\tag{422}$$

Proposition C.24 (Large-time alternatives)**.** *Assume $\nu_\phi \neq \delta_0$. If*

$$0 < \kappa \leq 2, \tag{423}$$

then

$$q(t) \longrightarrow +\infty, \qquad a(t) \longrightarrow 1. \tag{424}$$

If

$$\kappa > 2, \tag{425}$$

define

$$C_\kappa = \frac{\kappa^2}{\kappa - 1}. \tag{426}$$

If

$$C_\kappa w_0^2 < 4, \tag{427}$$

then

$$q(t) \longrightarrow +\infty, \qquad a(t) \longrightarrow 1. \tag{428}$$

If

$$C_\kappa w_0^2 > 4, \tag{429}$$

then

$$q(t) \longrightarrow -\infty, \qquad a(t) \longrightarrow a_* > 2, \tag{430}$$

where a_* is determined by

$$\frac{a_*^2}{a_* - 1} = C_\kappa w_0^2. \tag{431}$$

At the critical value

$$C_\kappa w_0^2 = 4, \tag{432}$$

one has

$$q(t) \longrightarrow -\infty, \qquad a(t) \longrightarrow 2. \tag{433}$$

Proof. Since $m_{\nu_\phi}(q) > 0$ for every finite q when $\nu_\phi \neq \delta_0$,

$$a < 1 \Longrightarrow \dot{a} > 0, \qquad a > 1 \Longrightarrow \dot{a} < 0. \tag{434}$$

If $0 < \kappa < 2$, then $a(t) < 2$ for all t, and hence

$$\dot{q}(t) > 0. \tag{435}$$

Thus $q(t) \geq q(0) = 0$. Since Eq. (326) gives $m'_{\nu_\phi}(q) \geq 0$ and $\nu_\phi \neq \delta_0$ implies

$$m_{\nu_\phi}(0) = \int \lambda \, d\nu_\phi(\lambda) > 0, \tag{436}$$

we have the uniform lower bound

$$m_{\nu_\phi}(q(t)) \geq m_{\nu_\phi}(0) > 0. \tag{437}$$

The monotonicity in Eq. (434) then forces $a(t) \to 1$. If $q(t)$ remained bounded above, this would give $\dot{q}(t) = 2(2 - a(t)) \to 2$, contradicting boundedness of an increasing q. Hence $q(t) \to +\infty$.

For $\kappa = 2$,

$$\dot{a}(0) = 4 \cdot 2 \cdot m_{\nu_\phi}(0)(1 - 2) = -8 m_{\nu_\phi}(0) < 0. \tag{438}$$

Therefore $a(t) < 2$ for all sufficiently small $t > 0$. Since $a(t) > 1$ implies $\dot{a}(t) < 0$, the trajectory then remains in $1 < a(t) < 2$, while $\dot{q}(t) > 0$. In particular $q(t) \geq 0$, and the same lower bound Eq. (437) and the preceding argument give $a(t) \to 1$ and $q(t) \to +\infty$.

Suppose now that $\kappa > 2$. While $a(t) > 2$,

$$\dot{q}(t) < 0. \tag{439}$$

On this branch, the first integral becomes

$$F(a(t)) = C_\kappa M_{\nu_\phi}(q(t))^2, \qquad F(a) = \frac{a^2}{a - 1}. \tag{440}$$

For $a \geq 2$,

$$F(a) \geq 4, \qquad F(2) = 4, \qquad F'(a) = \frac{a(a-2)}{(a-1)^2} > 0 \quad \text{for } a > 2. \tag{441}$$

As q decreases from 0 to $-\infty$, $M_{\nu_\phi}(q)$ decreases from 1 to w_0.

If equation 427 holds, continuity implies that $C_\kappa M_{\nu_\phi}(q)^2$ reaches the value 4 at a finite q. Let τ be the first corresponding time and set

$$q_* := q(\tau) > -\infty, \qquad a(\tau) = 2. \tag{442}$$

Because $m_{\nu_\phi}(q_*) > 0$,

$$\dot{a}(\tau) = -8 m_{\nu_\phi}(q_*) < 0. \tag{443}$$

Hence immediately after τ the trajectory enters $1 < a < 2$. On this region a is decreasing and cannot cross the invariant solution $a = 1$, while q is increasing. Therefore

$$q(t) \geq q_*, \qquad m_{\nu_\phi}(q(t)) \geq m_{\nu_\phi}(q_*) > 0, \qquad t \geq \tau, \tag{444}$$

where the second inequality again uses Eq. (326). Thus $a(t) \downarrow 1$. If $q(t)$ were bounded above, then $\dot{q}(t) \to 2$, a contradiction. Consequently $q(t) \to +\infty$ and $a(t) \to 1$. Notice that this argument does not require the crossing value q_* to equal zero.

If equation 429 holds, the right-hand side of equation 440 remains strictly larger than 4, so the trajectory never reaches $a = 2$. Hence q is strictly decreasing for all time. Suppose, for contradiction, that it were bounded below. Then there would exist a finite limit

$$q(t) \longrightarrow q_{\mathrm{f}} > -\infty. \tag{445}$$

By continuity in equation 440, $a(t)$ would then converge to the unique $a_\infty > 2$ satisfying

$$F(a_\infty) = C_\kappa M_{\nu_\phi}(q_{\mathrm{f}})^2 > 4. \tag{446}$$

It would follow that

$$\dot{q}(t) = 2(2 - a(t)) \longrightarrow 2(2 - a_\infty) < 0, \tag{447}$$

which is incompatible with convergence of $q(t)$ to a finite limit. Thus $q(t) \to -\infty$. Taking the limit in Eq. (440) gives Eq. (431).

At the critical value equation 432, the same argument requires one additional observation. For every finite q, the assumption $\nu_\phi \neq \delta_0$ implies

$$M_{\nu_\phi}(q) > w_0, \tag{448}$$

because the integral defining $M_{\nu_\phi}(q)$ has strictly positive contribution from $\lambda > 0$. Hence $C_\kappa M_{\nu_\phi}(q)^2 > 4$ at every finite q, so equation 440 gives $a(t) > 2$ and $q(t)$ is strictly decreasing for every finite time. If this decreasing trajectory converged to a finite q_{f}, then

$$C_\kappa M_{\nu_\phi}(q_{\mathrm{f}})^2 > C_\kappa w_0^2 = 4, \tag{449}$$

so again $a(t) \to a_\infty > 2$ and equation 447 gives a contradiction. Therefore $q(t) \to -\infty$ also in the critical case, and equation 440 together with $F(2) = 4$ yields $a(t) \to 2$. □

A regular lower spectral edge in the trapped phase. Assume the strict trapping condition equation 429. Then $w_0 > 0$, $q(t) \to -\infty$, and $a(t) \to a_* > 2$. It is useful to make the rate of this convergence explicit. Write

$$\nu_\phi = w_0 \delta_0 + \nu_{\phi_{\mathrm{c}}} \tag{450}$$

and suppose that the nonzero part has a strictly positive lower edge

$$\lambda_- := \inf \operatorname{supp} \nu_{\phi_{\mathrm{c}}} > 0 \tag{451}$$

with density

$$\frac{d\nu_{\phi_{\mathrm{c}}}}{d\lambda}(\lambda) = c_-(\lambda - \lambda_-)^{\alpha_-}\left(1 + O(\lambda - \lambda_-)\right), \qquad \alpha_- > -1, \qquad c_- > 0, \tag{452}$$

near λ_-. Set

$$\beta_- := \alpha_- + 1, \qquad A_- := c_- \Gamma(\alpha_- + 1). \tag{453}$$

For $s = -q \to +\infty$, Laplace's method at the lower edge gives

$$M_{\nu_\phi}(-s) = w_0 + A_- e^{-\lambda_- s} s^{-\beta_-} \left(1 + O(s^{-1})\right) \tag{454}$$

and

$$m_{\nu_\phi}(-s) = \frac{\lambda_- A_-}{w_0} e^{-\lambda_- s} s^{-\beta_-} \left(1 + O(s^{-1})\right). \tag{455}$$

Let

$$D_* := C_\kappa w_0^2 > 4. \tag{456}$$

The solution of Eq. (431) on the branch $a_* > 2$ is

$$a_* = \frac{D_* + \sqrt{D_*(D_* - 4)}}{2}. \tag{457}$$

Expanding the first integral Eq. (440) around a_*, and using

$$F'(a_*) = \frac{a_*(a_* - 2)}{(a_* - 1)^2}, \tag{458}$$

gives

$$a(t) - a_* \sim D_a^- e^{-\lambda_- s(t)} s(t)^{-\beta_-}, \qquad D_a^- := \frac{2a_*(a_* - 1)}{a_* - 2} \frac{A_-}{w_0}. \tag{459}$$

Indeed, $\dot{s}(t) = 2(a(t) - 2) \to 2(a_* - 2) > 0$, so $s(t) \geq (a_* - 2)t$ for all sufficiently large t; the error in Eq. (459) is therefore exponentially integrable. Define

$$v_* := 2(a_* - 2) \tag{460}$$

and the phase shift

$$q_- := \lim_{t \to \infty} \big(q(t) + v_* t\big) = -2 \int_0^\infty \big(a(s) - a_*\big)\, ds. \tag{461}$$

Then

$$q(t) = -v_* t + q_- + o(1) \tag{462}$$

and

$$a(t) - a_* \sim K_a^- t^{-\beta_-} e^{-\lambda_- v_* t}, \qquad K_a^- := D_a^- v_*^{-\beta_-} e^{\lambda_- q_-}. \tag{463}$$

Moreover,

$$r(t) \sim \frac{a_* \lambda_- A_-}{w_0} v_*^{-\beta_-} e^{\lambda_- q_-} t^{-\beta_-} e^{-\lambda_- v_* t}. \tag{464}$$

The exact self-consistency identity Eq. (370) also shows that $B(t)$ converges to

$$B_* := \frac{1}{2} \log\left(\frac{\kappa w_0}{a_*}\right), \tag{465}$$

with

$$B_* - B(t) \sim \frac{a_* A_-}{2w_0(a_* - 2)} v_*^{-\beta_-} e^{\lambda_- q_-} t^{-\beta_-} e^{-\lambda_- v_* t}. \tag{466}$$

For later comparison, define the scaled limiting loss corrections

$$F_{\mathrm{pop}}(t) := a(t) \left(\frac{a(t)}{2} - 1\right) \tag{467}$$

and

$$F_{\mathrm{tr}}(t) := r(t) \left(\frac{a(t)}{2} - 1\right). \tag{468}$$

In the trapped phase,

$$F_{\mathrm{pop}}(\infty) = a_* \left(\frac{a_*}{2} - 1\right), \qquad F_{\mathrm{tr}}(\infty) = 0, \tag{469}$$

and

$$F_{\mathrm{pop}}(t) - F_{\mathrm{pop}}(\infty) \sim (a_* - 1) K_a^- t^{-\beta_-} e^{-\lambda_- v_* t}, \tag{470}$$

while

$$F_{\mathrm{tr}}(t) \sim \frac{a_*(a_*-2)\lambda_- A_-}{2w_0} v_*^{-\beta_-} e^{\lambda_- q_-} t^{-\beta_-} e^{-\lambda_- v_* t}. \tag{471}$$

Thus the trapped branch is controlled by the lower nonzero spectral edge, in direct analogy with the upper-edge control of the alignment branch below. If the continuous spectrum touches zero rather than having $\lambda_- > 0$, the corresponding decay is polynomial and requires a separate lower-edge expansion.

In particular,

$$\nu_\phi(\{0\}) = 0 \quad \Longrightarrow \quad q(t) \longrightarrow +\infty, \qquad a(t) \longrightarrow 1 \tag{472}$$

for every $\kappa > 0$.

Critical boundary with a regular lower spectral edge. The boundary between alignment and strict trapping requires a separate expansion, since the derivative of the first-integral function $F(a) = a^2/(a-1)$ vanishes at $a = 2$. We retain the lower-edge assumptions above: write

$$\nu_\phi = w_0\delta_0 + \nu_{\phi c}, \qquad 0 < w_0 < 1, \qquad \lambda_- := \inf \operatorname{supp} \nu_{\phi c} > 0,$$

and suppose that, in a neighborhood of λ_-, the nonzero part is absolutely continuous with density

$$\frac{d\nu_{\phi c}}{d\lambda}(\lambda) = c_-(\lambda - \lambda_-)^{\alpha_-}\big(1 + O(\lambda - \lambda_-)\big), \qquad \alpha_- > -1, \quad c_- > 0.$$

As before, let

$$\beta_- := \alpha_- + 1, \qquad A_- := c_-\Gamma(\alpha_- + 1).$$

Proposition C.25 (Critical-boundary asymptotics)**.** *Assume the preceding lower-edge conditions and*

$$\kappa > 2, \qquad C_\kappa w_0^2 = 4, \qquad C_\kappa := \frac{\kappa^2}{\kappa - 1}. \tag{473}$$

Then the limiting scalar trajectory satisfies $a(t) > 2$ for every finite t, $a(t) \downarrow 2$, and $q(t) \to -\infty$. More precisely,

$$a(t) - 2 \sim \frac{1}{\lambda_- t}, \qquad m_{\nu_\phi}(q(t)) \sim \frac{1}{8\lambda_- t^2}, \qquad r(t) \sim \frac{1}{4\lambda_- t^2}. \tag{474}$$

The spectral-tilting clock has the expansion

$$q(t) = -\frac{2}{\lambda_-}\log t + \frac{\beta_-}{\lambda_-}\log\log t + q_{\mathrm{c}} + o(1), \qquad q_{\mathrm{c}} := \frac{1}{\lambda_-}\log\left[\frac{w_0}{8A_-\lambda_-^2}\left(\frac{2}{\lambda_-}\right)^{\beta_-}\right]. \tag{475}$$

For the scaled limiting loss corrections

$$F_{\mathrm{pop}}(t) = a(t)\left(\frac{a(t)}{2} - 1\right), \qquad F_{\mathrm{tr}}(t) = r(t)\left(\frac{a(t)}{2} - 1\right),$$

both limiting values are zero, and

$$F_{\mathrm{pop}}(t) \sim \frac{1}{\lambda_- t}, \qquad F_{\mathrm{tr}}(t) \sim \frac{1}{8\lambda_-^2 t^3}. \tag{476}$$

Finally, the cumulative quantity $B(t) = \int_0^t r(u)\,du$ obeys

$$B(t) \longrightarrow B_{\mathrm{c}} := \frac{1}{2}\log\left(\frac{\kappa w_0}{2}\right) = \frac{1}{4}\log(\kappa - 1), \qquad B_{\mathrm{c}} - B(t) \sim \frac{1}{4\lambda_- t}. \tag{477}$$

All statements concern the deterministic limiting system, with the high-dimensional limit taken before the large-time limit.

Proof. The qualitative limits and the strict inequality $a(t) > 2$ follow from the large-time alternatives proved above. Set

$$\delta(t) := a(t) - 2 > 0, \qquad s(t) := -q(t), \qquad \Delta(s) := M_{\nu_\phi}(-s) - w_0.$$

On the critical branch the first integral becomes

$$\frac{\delta(t)^2}{1+\delta(t)} = \frac{8\Delta(s(t))}{w_0} + \frac{4\Delta(s(t))^2}{w_0^2}, \tag{478}$$

because

$$\frac{(2+\delta)^2}{1+\delta} = 4 + \frac{\delta^2}{1+\delta}, \qquad C_\kappa M_{\nu_\phi}(-s)^2 = 4\left(1 + \frac{\Delta(s)}{w_0}\right)^2.$$

Laplace's method at the lower edge gives

$$\Delta(s) = A_- e^{-\lambda_- s} s^{-\beta_-}\big(1 + O(s^{-1})\big), \qquad m_{\nu_\phi}(-s) = \frac{\lambda_-}{w_0}\Delta(s)\big(1 + O(s^{-1})\big). \tag{479}$$

Combining equation 478 and equation 479 yields

$$\Delta(s(t)) \sim \frac{w_0}{8}\delta(t)^2, \qquad m_{\nu_\phi}(q(t)) \sim \frac{\lambda_-}{8}\delta(t)^2. \tag{480}$$

Consequently,

$$\dot\delta(t) = -4(2+\delta(t))(1+\delta(t))m_{\nu_\phi}(q(t)) = -\lambda_-\delta(t)^2\big(1 + o(1)\big).$$

Thus $(1/\delta(t))' \to \lambda_-$, which gives $\delta(t) \sim 1/(\lambda_- t)$. The other two equivalents in equation 474 follow from equation 480 and $r(t) = (2+\delta(t))m_{\nu_\phi}(q(t))$.

To refine the clock, set $d_{\mathrm{c}} := \sqrt{8A_-/w_0}$. Eqs. (478) and (479) give

$$\delta(t) = d_{\mathrm{c}} e^{-\lambda_- s(t)/2} s(t)^{-\beta_-/2}\big(1 + O(s(t)^{-1})\big).$$

Since $\dot s = 2\delta$, integration of the inverse equation gives

$$t = \frac{1}{\lambda_- d_{\mathrm{c}}} e^{\lambda_- s(t)/2} s(t)^{\beta_-/2}\big(1 + O(s(t)^{-1})\big). \tag{481}$$

Taking logarithms and inverting yields equation 475.

The losses satisfy the exact identities

$$F_{\mathrm{pop}}(t) = \delta(t) + \frac{1}{2}\delta(t)^2, \qquad F_{\mathrm{tr}}(t) = \frac{1}{2}r(t)\delta(t),$$

so equation 476 follows directly from equation 474. The self-consistency identity

$$B(t) = \frac{1}{2}\log\left(\frac{\kappa M_{\nu_\phi}(q(t))}{a(t)}\right)$$

identifies B_{c}, and $B_{\mathrm{c}} - B(t) = \int_t^\infty r(u)\,du$ gives the final equivalent in equation 477. □

Marchenko–Pastur specialization at the boundary. For the Marchenko–Pastur law the critical branch necessarily has $\phi > 1$ and

$$w_0 = 1 - \frac{1}{\phi}, \qquad \lambda_- = (\sqrt{\phi} - 1)^2, \qquad \beta_- = \frac{3}{2}.$$

Hence, when $\kappa > 2$ and $\frac{\kappa^2}{\kappa - 1}(1 - \phi^{-1})^2 = 4$,

$$F_{\mathrm{pop}}(t) \sim \frac{1}{(\sqrt{\phi} - 1)^2 t}, \qquad F_{\mathrm{tr}}(t) \sim \frac{1}{8(\sqrt{\phi} - 1)^4 t^3},$$

and

$$q(t) = -\frac{2}{\lambda_-}\log t + \frac{3}{2\lambda_-}\log\log t + q_{\mathrm{c}} + o(1).$$

The lower-edge exponent enters the logarithmic correction to the clock, but not the leading powers t^{-1} and t^{-3} of the loss tails. These equivalents assume a positive lower nonzero edge; they do not assert the same critical powers when the continuous spectrum reaches zero.

A regular upper spectral edge. Assume from now on that the trajectory lies in the alignment basin, so that

$$q(t) \longrightarrow +\infty, \qquad a(t) \longrightarrow 1. \tag{482}$$

Let

$$\lambda_+ = \sup \operatorname{supp} \nu_\phi \tag{483}$$

and assume that near λ_+, the measure ν_ϕ has a density of the form

$$\frac{d\nu_\phi}{d\lambda}(\lambda) = c_+(\lambda_+ - \lambda)^\alpha \left(1 + O(\lambda_+ - \lambda)\right), \qquad \alpha > -1, \qquad c_+ > 0. \tag{484}$$

Then Laplace's method gives

$$M_{\nu_\phi}(q) = A_{\nu_\phi} e^{\lambda_+ q} q^{-(\alpha+1)} \left(1 + O(q^{-1})\right), \tag{485}$$

where

$$A_{\nu_\phi} = c_+ \Gamma(\alpha + 1). \tag{486}$$

Indeed, after the change of variables

$$y = q(\lambda_+ - \lambda), \tag{487}$$

one obtains

$$e^{-\lambda_+ q} q^{\alpha+1} M_{\nu_\phi}(q) \longrightarrow c_+ \int_0^\infty e^{-y} y^\alpha \, dy = c_+ \Gamma(\alpha + 1). \tag{488}$$

Consequently,

$$m_{\nu_\phi}(q) = \lambda_+ - \frac{\alpha + 1}{q} + O(q^{-2}). \tag{489}$$

For $\kappa \neq 1$, the first integral gives

$$|1 - a(t)| = \frac{|1 - \kappa|}{\kappa^2} \frac{a(t)^2}{M_{\nu_\phi}(q(t))^2}. \tag{490}$$

Thus

$$|1 - a(t)| = O\left(q(t)^{2(\alpha+1)} e^{-2\lambda_+ q(t)}\right). \tag{491}$$

This error is integrable once q is sufficiently large. Since

$$\dot{q}(t) = 2 - 2(a(t) - 1), \tag{492}$$

the phase shift exists and has the exact representation

$$q_\infty := \lim_{t \to \infty} \big(q(t) - 2t\big) = -2 \int_0^\infty \big(a(s) - 1\big) \, ds. \tag{493}$$

Thus

$$q(t) = 2t + q_\infty + o(1). \tag{494}$$

It follows that, for $\kappa \neq 1$,

$$|1 - a(t)| \sim K_a t^{2(\alpha+1)} e^{-4\lambda_+ t}, \qquad K_a := \frac{|1 - \kappa|}{\kappa^2} \frac{2^{2(\alpha+1)}}{A_{\nu_\phi}^2} e^{-2\lambda_+ q_\infty}. \tag{495}$$

For $\kappa = 1$,

$$a(t) \equiv 1. \tag{496}$$

Since

$$r(t) = a(t) m_{\nu_\phi}(q(t)), \tag{497}$$

Equations (489) and (494) give

$$r(t) = \lambda_+ - \frac{\alpha + 1}{2t} + O(t^{-2}) + O\left(t^{2(\alpha+1)} e^{-4\lambda_+ t}\right). \tag{498}$$

The exact self-consistency relation Eq. (370) gives

$$B(t) = \frac{1}{2}\left[\log \kappa + \log M_{\nu_\phi}(q(t)) - \log a(t)\right]. \tag{499}$$

Combining this identity with Eqs. (485) and (494) yields

$$B(t) = \lambda_+ t - \frac{\alpha+1}{2}\log t + B_\infty + O(t^{-1}), \qquad B_\infty := \frac{\lambda_+ q_\infty}{2} + \frac{1}{2}\log\left(\frac{\kappa A_{\nu_\phi}}{2^{\alpha+1}}\right). \tag{500}$$

Equivalently, for $\kappa \neq 1$, the two exact asymptotic constants satisfy

$$K_a = |1-\kappa| e^{-4B_\infty}. \tag{501}$$

For the scaled loss corrections defined in Eqs. (467) and (468), since

$$F_{\mathrm{pop}}(t) = -\frac{1}{2} + \frac{1}{2}(a(t)-1)^2, \tag{502}$$

we obtain

$$F_{\mathrm{pop}}(t) - F_{\mathrm{pop}}(\infty) \sim \frac{1}{2} K_a^2 t^{4(\alpha+1)} e^{-8\lambda_+ t}. \tag{503}$$

Similarly,

$$F_{\mathrm{tr}}(t) - F_{\mathrm{tr}}(\infty) = \frac{\lambda_+ - r(t)}{2} + \frac{r(t)}{2}(a(t)-1), \tag{504}$$

and hence

$$F_{\mathrm{tr}}(t) - F_{\mathrm{tr}}(\infty) = \frac{\alpha+1}{4t} + O(t^{-2}) + O\left(t^{2(\alpha+1)} e^{-4\lambda_+ t}\right). \tag{505}$$

Thus the large-time train-loss tail is controlled by the local spectral density near the upper edge. The Marchenko–Pastur coefficient is a special case of this general edge calculation.

Marchenko–Pastur specialization. Suppose now that the limiting spectral measure is the Marchenko–Pastur law with aspect ratio

$$\frac{p}{m} \longrightarrow \phi \in (0,\infty). \tag{506}$$

Its law is

$$d\mu_\phi(\lambda) = \left(1 - \frac{1}{\phi}\right)_+ \delta_0(d\lambda) + \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\phi\lambda} \mathbf{1}_{[\lambda_-,\lambda_+]}(\lambda)\, d\lambda, \tag{507}$$

where

$$\lambda_\pm = (1 \pm \sqrt{\phi})^2. \tag{508}$$

At the upper edge,

$$\frac{d\mu_\phi}{d\lambda}(\lambda) = \frac{1}{\pi\phi^{3/4}\lambda_+}(\lambda_+ - \lambda)^{1/2}\left(1 + O(\lambda_+ - \lambda)\right). \tag{509}$$

Thus

$$\alpha = \frac{1}{2} \tag{510}$$

and

$$A_\phi = \frac{1}{2\sqrt{\pi}\,\phi^{3/4}\lambda_+}. \tag{511}$$

Therefore

$$M_\phi(q) = A_\phi e^{\lambda_+ q} q^{-3/2}\left(1 + O(q^{-1})\right) \tag{512}$$

and

$$m_\phi(q) = \lambda_+ - \frac{3}{2q} + O(q^{-2}). \tag{513}$$

Within the alignment basin,

$$q(t) = 2t + q_\infty + o(1), \tag{514}$$

$$r(t) = \lambda_+ - \frac{3}{4t} + O(t^{-2}) + O(t^3 e^{-4\lambda_+ t}), \tag{515}$$

and, for $\kappa \neq 1$,

$$|1 - a(t)| \sim K_a t^3 e^{-4\lambda_+ t}. \tag{516}$$

Moreover,

$$B(t) = \lambda_+ t - \frac{3}{4}\log t + B_\infty + O(t^{-1}). \tag{517}$$

The limiting scaled losses satisfy

$$F_{\mathrm{pop}}(t) - F_{\mathrm{pop}}(\infty) \sim \frac{1}{2}K_a^2 t^6 e^{-8\lambda_+ t} \tag{518}$$

and

$$F_{\mathrm{tr}}(t) - F_{\mathrm{tr}}(\infty) = \frac{3}{8t} + O(t^{-2}) + O(t^3 e^{-4\lambda_+ t}). \tag{519}$$

When $\phi > 1$, the Marchenko–Pastur law has a zero atom of mass

$$w_0 = 1 - \frac{1}{\phi}. \tag{520}$$

Thus, for $\kappa > 2$, the trajectory lies in the alignment basin precisely when

$$\frac{\kappa^2}{\kappa - 1}\left(1 - \frac{1}{\phi}\right)^2 < 4. \tag{521}$$

If the reverse strict inequality holds, the limiting scalar dynamics is trapped in the null space as in Proposition C.24. In that case the nonzero Marchenko–Pastur spectrum has lower edge $\lambda_- = (\sqrt{\phi} - 1)^2 > 0$ and square-root exponent $\alpha_- = 1/2$. Therefore Eqs. (462) and (463) give

$$q(t) = -2(a_* - 2)t + q_- + o(1), \qquad a(t) - a_* \asymp t^{-3/2} e^{-2\lambda_-(a_* - 2)t}. \tag{522}$$

C.3.7 Landscape interpretation of null-space trapping

The large-time analysis above shows that, in the strict trapped phase, the limiting system satisfies

$$a(t) \longrightarrow a_* > 2, \qquad r(t) \longrightarrow 0,$$

while at the critical boundary $a(t) \to 2$. We now relate these values to the finite-dimensional landscape of the empirical loss. The statement below is deterministic and does not involve a large-dimensional limit.

Proposition C.26 (Null-space critical points of the empirical loss)**.** *Let $\boldsymbol{S} \in \mathbb{R}^{p\times p}$ be symmetric positive semidefinite and nonzero, and consider*

$$\widehat{L}(\boldsymbol{U}) = \frac{1}{2p}\operatorname{Tr}\left[(\boldsymbol{I}_p - \boldsymbol{U}^\top \boldsymbol{U})\boldsymbol{S}(\boldsymbol{I}_p - \boldsymbol{U}^\top \boldsymbol{U})\right], \qquad \boldsymbol{U} \in \mathbb{R}^{n\times p}.$$

Suppose

$$\boldsymbol{U}\boldsymbol{S} = 0, \tag{523}$$

and write

$$\boldsymbol{G} := \boldsymbol{U}\boldsymbol{U}^\top.$$

Then $\boldsymbol{U}$ is a critical point of $\widehat{L}$. Moreover:

1. *If*
$$\lambda_{\min}(\boldsymbol{G}) > 2,$$
then $\boldsymbol{U}$ is a nonglobal local minimum of $\widehat{L}$.

2. *If*
$$\lambda_{\min}(\boldsymbol{G}) < 2,$$
then $\boldsymbol{U}$ is a strict saddle: the Hessian of $\widehat{L}$ has a negative direction at $\boldsymbol{U}$.

3. *If*
$$\lambda_{\min}(\boldsymbol{G}) = 2,$$
then $\boldsymbol{U}$ is not a local minimum. In this case the negative variation may first appear beyond quadratic order.

Consequently, among the critical points satisfying equation 523, the local minima are precisely those for which

$$\boldsymbol{U}\boldsymbol{U}^\top \succ 2\boldsymbol{I}_n. \tag{524}$$

Proof. Let

$$\boldsymbol{G} = \boldsymbol{U}\boldsymbol{U}^\top, \qquad \boldsymbol{H} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^\top.$$

By the exact train-loss identity equation 219,

$$\begin{aligned} \widehat{L}(\boldsymbol{U}) &= \frac{1}{2p}\operatorname{Tr}\boldsymbol{S} - \frac{1}{p}\operatorname{Tr}\boldsymbol{H} + \frac{1}{2p}\operatorname{Tr}(\boldsymbol{G}\boldsymbol{H}) \\ &= \frac{1}{2p}\operatorname{Tr}\boldsymbol{S} + \frac{1}{2p}\operatorname{Tr}[(\boldsymbol{G} - 2\boldsymbol{I}_n)\boldsymbol{H}]\,. \end{aligned} \tag{525}$$

Under equation 523, one has $\boldsymbol{H} = 0$, and therefore

$$\widehat{L}(\boldsymbol{U}) = \frac{1}{2p}\operatorname{Tr}\boldsymbol{S}. \tag{526}$$

Moreover, $\boldsymbol{U}$ is critical. Indeed, the common gradient identity equation 53 gives

$$\frac{\partial \widehat{L}}{\partial \boldsymbol{U}} = -\frac{1}{p}\boldsymbol{U}\left[\boldsymbol{S}(\boldsymbol{I}_p - \boldsymbol{U}^\top\boldsymbol{U}) + (\boldsymbol{I}_p - \boldsymbol{U}^\top\boldsymbol{U})\boldsymbol{S}\right],$$

and both terms vanish when $\boldsymbol{U}\boldsymbol{S} = 0$.

Case 1: $\lambda_{\min}(\boldsymbol{G}) > 2$**.** By continuity, there exists a neighborhood of $\boldsymbol{U}$ in which

$$\boldsymbol{U}'\boldsymbol{U}'^\top - 2\boldsymbol{I}_n \succ 0.$$

For every $\boldsymbol{U}'$,

$$\boldsymbol{U}'\boldsymbol{S}\boldsymbol{U}'^\top \succeq 0.$$

Hence equation 525 gives

$$\begin{aligned} \widehat{L}(\boldsymbol{U}') - \widehat{L}(\boldsymbol{U}) &= \frac{1}{2p}\operatorname{Tr}\big[(\boldsymbol{U}'\boldsymbol{U}'^\top - 2\boldsymbol{I}_n)\boldsymbol{U}'\boldsymbol{S}\boldsymbol{U}'^\top\big] \\ &\geq 0 \end{aligned} \tag{527}$$

for all sufficiently close $\boldsymbol{U}'$. Thus $\boldsymbol{U}$ is a local minimum.

It is generally non-strict: perturbations that remain inside $\ker \boldsymbol{S}$ may preserve $\boldsymbol{U}'\boldsymbol{S} = 0$ and therefore leave the loss equal to $\operatorname{Tr}\boldsymbol{S}/(2p)$.

It is also nonglobal. Since $\boldsymbol{S} \neq 0$, choose a unit eigenvector $\boldsymbol{x}$ such that

$$\boldsymbol{S}\boldsymbol{x} = \lambda\boldsymbol{x}, \qquad \lambda > 0.$$

Let $\widetilde{\boldsymbol{U}}$ have first row $\boldsymbol{x}^\top$ and all remaining rows equal to zero. Then

$$\widetilde{\boldsymbol{U}}^\top\widetilde{\boldsymbol{U}} = \boldsymbol{x}\boldsymbol{x}^\top,$$

and a direct substitution gives

$$\widehat{L}(\widetilde{\boldsymbol{U}}) = \frac{\operatorname{Tr}\boldsymbol{S} - \lambda}{2p} < \frac{\operatorname{Tr}\boldsymbol{S}}{2p} = \widehat{L}(\boldsymbol{U}).$$

Case 2: $\lambda_{\min}(\boldsymbol{G}) < 2$**.** Let $\boldsymbol{v} \in \mathbb{R}^n$ be a unit eigenvector of $\boldsymbol{G}$ with

$$\boldsymbol{G}\boldsymbol{v} = g\boldsymbol{v}, \qquad g < 2.$$

Choose a unit eigenvector $\boldsymbol{x}$ of $\boldsymbol{S}$ with

$$\boldsymbol{S}\boldsymbol{x} = \lambda\boldsymbol{x}, \qquad \lambda > 0.$$

Since $\boldsymbol{U}\boldsymbol{S} = 0$, we have $\boldsymbol{U}\boldsymbol{x} = 0$. Consider

$$\boldsymbol{U}_\epsilon = \boldsymbol{U} + \epsilon\boldsymbol{v}\boldsymbol{x}^\top.$$

Then

$$\boldsymbol{U}_\epsilon \boldsymbol{U}_\epsilon^\top = \boldsymbol{G} + \epsilon^2 \boldsymbol{v}\boldsymbol{v}^\top \tag{528}$$

and

$$\boldsymbol{U}_\epsilon \boldsymbol{S} \boldsymbol{U}_\epsilon^\top = \lambda \epsilon^2 \boldsymbol{v}\boldsymbol{v}^\top. \tag{529}$$

Substitution into equation 525 yields

$$\widehat{L}(\boldsymbol{U}_\epsilon) - \widehat{L}(\boldsymbol{U}) = \frac{\lambda \epsilon^2}{2p}\left(g - 2 + \epsilon^2\right). \tag{530}$$

This is negative for all sufficiently small nonzero ϵ, and its leading term is

$$\frac{\lambda(g-2)}{2p}\epsilon^2 < 0.$$

Thus the Hessian has a negative direction, and $\boldsymbol{U}$ is a strict saddle.

Case 3: $\lambda_{\min}(\boldsymbol{G}) = 2$**.** Let $\boldsymbol{v}$ be a unit vector satisfying

$$\boldsymbol{G}\boldsymbol{v} = 2\boldsymbol{v},$$

and define

$$\boldsymbol{w} := \boldsymbol{U}^\top \boldsymbol{v}.$$

Then

$$\|\boldsymbol{w}\|^2 = \boldsymbol{v}^\top \boldsymbol{G}\boldsymbol{v} = 2.$$

Also, since $\boldsymbol{U}\boldsymbol{S} = 0$ and $\boldsymbol{S}$ is symmetric,

$$\boldsymbol{S}\boldsymbol{w} = \boldsymbol{S}\boldsymbol{U}^\top \boldsymbol{v} = 0.$$

Choose again a unit eigenvector $\boldsymbol{x}$ of $\boldsymbol{S}$ with

$$\boldsymbol{S}\boldsymbol{x} = \lambda \boldsymbol{x}, \qquad \lambda > 0.$$

Then $\boldsymbol{w}^\top \boldsymbol{x} = 0$ and $\boldsymbol{U}\boldsymbol{x} = 0$.

For $\epsilon \neq 0$, set

$$\alpha_\epsilon := \frac{\epsilon^2}{2}, \qquad \boldsymbol{z}_\epsilon := -\alpha_\epsilon \boldsymbol{w} + \epsilon \boldsymbol{x},$$

and consider the perturbation

$$\boldsymbol{U}_\epsilon := \boldsymbol{U} + \boldsymbol{v}\boldsymbol{z}_\epsilon^\top.$$

Since

$$\boldsymbol{U}\boldsymbol{w} = \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{v} = 2\boldsymbol{v}, \qquad \boldsymbol{U}\boldsymbol{x} = 0,$$

we obtain

$$\boldsymbol{U}\boldsymbol{z}_\epsilon = -2\alpha_\epsilon \boldsymbol{v}.$$

Therefore

$$\boldsymbol{U}_\epsilon \boldsymbol{U}_\epsilon^\top = \boldsymbol{G} + \left(-4\alpha_\epsilon + \|\boldsymbol{z}_\epsilon\|^2\right)\boldsymbol{v}\boldsymbol{v}^\top.$$

Using

$$\|\boldsymbol{z}_\epsilon\|^2 = 2\alpha_\epsilon^2 + \epsilon^2,$$

we get

$$\boldsymbol{v}^\top\left(\boldsymbol{U}_\epsilon \boldsymbol{U}_\epsilon^\top - 2\boldsymbol{I}_n\right)\boldsymbol{v} = -\epsilon^2 + \frac{\epsilon^4}{2}. \tag{531}$$

On the other hand, the component proportional to $\boldsymbol{w}$ lies in $\ker \boldsymbol{S}$, so

$$\boldsymbol{U}_\epsilon \boldsymbol{S} \boldsymbol{U}_\epsilon^\top = \lambda \epsilon^2 \boldsymbol{v}\boldsymbol{v}^\top. \tag{532}$$

Thus

$$\begin{aligned}\widehat{L}(\boldsymbol{U}_\epsilon) - \widehat{L}(\boldsymbol{U}) &= \frac{\lambda \epsilon^2}{2p}\left(-\epsilon^2 + \frac{\epsilon^4}{2}\right) \\ &< 0\end{aligned} \tag{533}$$

for every sufficiently small nonzero ϵ. Hence $\boldsymbol{U}$ is not a local minimum.

The three cases establish equation 524. □

Relation to the trapped limiting dynamics. Proposition C.24 shows that under the strict trapping condition equation 429,

$$a(t) \longrightarrow a_* > 2, \qquad q(t) \longrightarrow -\infty,$$

while equation 464 gives

$$r(t) \longrightarrow 0.$$

Together with Theorem C.12, which yields

$$\boldsymbol{G}_N(t) \approx a(t)\boldsymbol{I}_n, \qquad \boldsymbol{H}_N(t) \approx r(t)\boldsymbol{I}_n$$

on every fixed time interval, this identifies the strict trapped large-time state of the limiting system with the local-minimum region

$$\boldsymbol{U}\boldsymbol{S} = 0, \qquad \boldsymbol{U}\boldsymbol{U}^\top \succ 2\boldsymbol{I}_n$$

of the finite-dimensional tied empirical loss.

At the critical boundary equation 432, instead, $a(t) \to 2$. The corresponding null-space configuration lies on the degenerate boundary

$$\lambda_{\min}(\boldsymbol{U}\boldsymbol{U}^\top) = 2$$

and, by Proposition C.26, is not a local minimum.

We stress that this is a landscape interpretation of the large-time limit of the deterministic limiting system. The finite-dimensional convergence theorem is uniform only on fixed time intervals and does not assert that a finite-N trajectory converges as $t \to \infty$ to an exact null-space critical point.

C.3.8 Role of Gaussianity and brief extension to anisotropic data

The preceding proof separates three assumptions that are sometimes conflated.

First, Gaussianity of the *training data* is not used in the quenched approximation Theorem C.16. Conditional on $\boldsymbol{S}_N$, the training data appear only through the deterministic matrix $\boldsymbol{S}_N$. The Gaussian concentration argument is applied to the initialization $\boldsymbol{Z}$, not to the samples $\boldsymbol{x}_i$.

Second, isotropy of the data is needed only to reduce the population loss to equation 218. Neither the training flow nor the scalarization of $\boldsymbol{G}$ and $\boldsymbol{H}$ requires population isotropy.

Third, isotropy alone does not imply a particular limiting empirical spectral measure. The passage from a_N, r_N to a, r uses the standing spectral assumptions equation 41–equation 42. The limiting spectral measure may be Marchenko–Pastur, but it need not be.

We briefly describe the anisotropic extension. Suppose

$$\boldsymbol{\Sigma}_p = \mathbb{E}\boldsymbol{x}\boldsymbol{x}^\top \neq \boldsymbol{I}_p \tag{534}$$

and assume

$$\sup_p \|\boldsymbol{\Sigma}_p\|_{\mathrm{op}} < \infty. \tag{535}$$

The empirical training flow remains the common flow equation 49 with $\gamma = 1$, since it depends on the observed data only through $\boldsymbol{S}_N$. Thus the conclusions for $\boldsymbol{G}_N$, $\boldsymbol{H}_N$, a, and r are unchanged whenever the same assumptions on $\boldsymbol{S}_N$ hold.

The population loss becomes

$$L_\Sigma(\boldsymbol{U}) = \frac{1}{2p} \operatorname{Tr}\left[(\boldsymbol{I}_p - \boldsymbol{U}^\top\boldsymbol{U})\boldsymbol{\Sigma}_p(\boldsymbol{I}_p - \boldsymbol{U}^\top\boldsymbol{U})\right]. \tag{536}$$

Introduce

$$\boldsymbol{K}_N(t) = \boldsymbol{U}_N(t)\boldsymbol{\Sigma}_p\boldsymbol{U}_N(t)^\top. \tag{537}$$

For the proxy trajectory,

$$\overline{\boldsymbol{K}}_N(t) = \frac{1}{p}\boldsymbol{Z}\boldsymbol{C}_{\Sigma,N}(t)\boldsymbol{Z}^\top, \tag{538}$$

where

$$\boldsymbol{C}_{\Sigma,N}(t) = \kappa_N e^{-2B_N(t)} e^{q_N(t)\boldsymbol{S}_N/2}\boldsymbol{\Sigma}_p e^{q_N(t)\boldsymbol{S}_N/2}. \tag{539}$$

Its normalized trace is

$$\frac{1}{p}\operatorname{Tr}\boldsymbol{C}_{\Sigma,N}(t)=\kappa_N e^{-2B_N(t)}W_N(q_N(t)), \tag{540}$$

where

$$W_N(q)=\frac{1}{p}\operatorname{Tr}\left(\boldsymbol{\Sigma}_p e^{q\boldsymbol{S}_N}\right). \tag{541}$$

If, uniformly on compact q-intervals,

$$W_N(q)\xrightarrow{\mathbb{P}}W(q), \tag{542}$$

then the same Gaussian Gram concentration argument gives

$$\boldsymbol{K}_N(t)=b(t)\boldsymbol{I}_n+o_{\mathbb{P}}(1)\quad\text{in operator norm}, \tag{543}$$

where

$$b(t)=\kappa e^{-2B(t)}W(q(t)). \tag{544}$$

Expanding equation 536,

$$L_{\Sigma,N}=\frac{1}{2p}\operatorname{Tr}\boldsymbol{\Sigma}_p-\frac{1}{p}\operatorname{Tr}\boldsymbol{K}_N+\frac{1}{2p}\operatorname{Tr}(\boldsymbol{G}_N\boldsymbol{K}_N), \tag{545}$$

so that

$$\psi_N\left(L_{\Sigma,N}(t)-\frac{1}{2p}\operatorname{Tr}\boldsymbol{\Sigma}_p\right)\xrightarrow{\mathbb{P}}b(t)\left(\frac{a(t)}{2}-1\right). \tag{546}$$

In the isotropic case,

$$\boldsymbol{\Sigma}_p=\boldsymbol{I}_p\quad\Longrightarrow\quad W(q)=M_{\nu_\phi}(q)\quad\Longrightarrow\quad b(t)=a(t), \tag{547}$$

recovering the previous population formula.

For strongly anisotropic covariance sequences, bounded operator norm may fail and the relevant dimension controlling Gram concentration may become smaller than p. A useful effective-rank quantity is

$$r_{\text{eff}}(\boldsymbol{C})=\frac{\operatorname{Tr}\boldsymbol{C}}{\|\boldsymbol{C}\|_{\text{op}}}. \tag{548}$$

A natural strengthened narrow-latent condition is then

$$\frac{n}{r_{\text{eff}}\left(e^{q\boldsymbol{S}_N/2}\boldsymbol{\Sigma}_p e^{q\boldsymbol{S}_N/2}\right)}\longrightarrow 0 \tag{549}$$

uniformly on the finite time interval under consideration. Thus ordinary bounded anisotropy changes primarily the population observable, whereas sufficiently strong anisotropy can alter the effective dimension controlling row decoupling itself.

C.4 Large-data (Section 4.2)

We now specialize the common setup of Appendix C.1 to the large-data corner. The common empirical covariance $\boldsymbol{S}_p$, Gram state $\boldsymbol{A}_p$, error $\boldsymbol{E}_p$, losses, and gradient flow will not be redefined here. We take $\eta=p$, so $\gamma=1$, and assume

$$\phi_p:=\frac{p}{m}\longrightarrow 0,\qquad \psi_p:=\frac{p}{n}\longrightarrow\psi\in(0,\infty),\qquad \rho_p:=n\sigma_p^2\longrightarrow\rho\in(0,\infty). \tag{550}$$

The training samples are independent copies of a centered isotropic vector,

$$\mathbb{E}\boldsymbol{x}=0,\qquad \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]=\boldsymbol{I}_p, \tag{551}$$

The data hypothesis is the common near-isotropy condition equation 45, now applied to $\boldsymbol{S}_p$. For convenience set

$$\varepsilon_p:=\|\boldsymbol{S}_p-\boldsymbol{I}_p\|_{\text{op}}. \tag{552}$$

Thus $\varepsilon_p\to 0$ in probability; in particular the standing spectral assumptions of Appendix C.1 hold automatically with $\nu_\phi=\delta_1$. The covariance concentration is the data-side input that replaces the

empirical trajectory by the $\boldsymbol{S} = \boldsymbol{I}_p$ proxy and identifies both train and population losses with the same logistic limiting curve. We write $L_p(t) := L(\boldsymbol{U}_p(t))$ and $\widehat{L}_p(t) := \widehat{L}(\boldsymbol{U}_p(t))$.

For $t \geq 0$ and $s \geq 0$, define the scalar logistic flow map

$$\Theta_t(s) := \frac{s}{s + (1-s)e^{-4t}} = \frac{se^{4t}}{1 + s(e^{4t} - 1)}. \tag{553}$$

For a positive semidefinite matrix $\boldsymbol{B}$, $\Theta_t(\boldsymbol{B})$ denotes the corresponding functional calculus.

We also recall the Marchenko–Pastur distribution with aspect ratio ψ:

$$\mu_\psi(d\lambda) = \left(1 - \frac{1}{\psi}\right)_+ \delta_0(d\lambda) + \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\psi\lambda} \mathbf{1}_{\{\lambda \in [\lambda_-, \lambda_+]\}}\, d\lambda, \tag{554}$$

where

$$\lambda_- = (1 - \sqrt{\psi})^2, \qquad \lambda_+ = (1 + \sqrt{\psi})^2. \tag{555}$$

Define the deterministic limiting loss

$$\mathcal{L}_{\psi,\rho}(t) := \frac{1}{2} \int \big(\Theta_t(\rho\lambda) - 1\big)^2 \, \mu_\psi(d\lambda). \tag{556}$$

Main theorem.

Theorem C.27 (Large-data limit)**.** *Assume equation 550, equation 551, and the common near-isotropy condition equation 45. Let $\boldsymbol{U}_p(t)$ be the gradient-flow trajectory for the empirical loss equation 38 with learning rate $\eta = p$, and let $\boldsymbol{A}_p(t) = \boldsymbol{U}_p(t)^\top \boldsymbol{U}_p(t)$. Then, for every $T < \infty$,*

$$\sup_{0 \leq t \leq T} \|\boldsymbol{A}_p(t) - \Theta_t(\boldsymbol{A}_p(0))\|_{\mathrm{op}} \xrightarrow[p\to\infty]{\mathbb{P}} 0. \tag{557}$$

Moreover,

$$\sup_{0 \leq t \leq T} |L_p(t) - \mathcal{L}_{\psi,\rho}(t)| \xrightarrow[p\to\infty]{\mathbb{P}} 0, \tag{558}$$

and

$$\sup_{0 \leq t \leq T} \left|\widehat{L}_p(t) - \mathcal{L}_{\psi,\rho}(t)\right| \xrightarrow[p\to\infty]{\mathbb{P}} 0. \tag{559}$$

If, in addition,

$$\mathbb{E}\|\boldsymbol{S} - \boldsymbol{I}_p\|_{\mathrm{op}} \longrightarrow 0 \tag{560}$$

and the family $\{\|\boldsymbol{S}\|_{\mathrm{op}}\}_p$ has a uniformly bounded moment of some order strictly larger than one, then

$$\mathbb{E} \sup_{0 \leq t \leq T} |L_p(t) - \mathcal{L}_{\psi,\rho}(t)| \longrightarrow 0 \tag{561}$$

and

$$\mathbb{E} \sup_{0 \leq t \leq T} \left|\widehat{L}_p(t) - \mathcal{L}_{\psi,\rho}(t)\right| \longrightarrow 0. \tag{562}$$

In particular, the expected train and population losses converge uniformly on every compact time interval to the same deterministic curve equation 556.

Proof plan. The proof is a data-side stability argument. First, the common Gram-flow results reduce the empirical trajectory to the proxy with $\boldsymbol{S} = \boldsymbol{I}_p$, with an error controlled by $\|\boldsymbol{S} - \boldsymbol{I}_p\|_{\mathrm{op}}$. Second, the assumed covariance concentration (or the sub-Gaussian sufficient condition proved below) makes this perturbation vanish. Third, the proxy is the logistic matrix flow of Lemma C.7. Finally, the Marchenko–Pastur law for the Gaussian initialization identifies the limiting proxy loss; Lemma C.6 transfers this limit to both train and population losses. The final uniform-integrability argument gives the stated convergence of expectations.

Exact finite-dimensional dynamics and deterministic stability. With $\eta = p$, Proposition C.1 and Lemma C.2 give the exact $\boldsymbol{U}$- and $\boldsymbol{A}$-flows and the dimension-free global bound. For a positive semidefinite covariance $\boldsymbol{S}$, write $\boldsymbol{A}_{\boldsymbol{S}}(t)$ for the solution of the common Gram ODE with initial state $\boldsymbol{A}_0$, and write $\boldsymbol{A}_{\boldsymbol{I}}(t)$ for the corresponding solution with covariance $\boldsymbol{I}_p$. Proposition C.5 gives equation 86 with the constant equation 87 whenever $\|\boldsymbol{S} - \boldsymbol{I}_p\|_{\mathrm{op}} \leq 1$. The population- and train-loss transfers are exactly equation 95–equation 96 from Lemma C.6.

A sufficient condition for empirical covariance concentration. The deterministic argument above only needs equation 45. We now verify this assumption for a broad class of non-Gaussian data.

For a real random variable Y, define the sub-Gaussian and sub-exponential Orlicz norms by

$$\|Y\|_{\psi_2} := \inf\left\{s > 0 : \mathbb{E}\exp(Y^2/s^2) \le 2\right\}, \qquad \|Y\|_{\psi_1} := \inf\left\{s > 0 : \mathbb{E}\exp(|Y|/s) \le 2\right\}. \tag{563}$$

Assume that the isotropic data vectors satisfy the uniform sub-Gaussian condition

$$\sup_{p} \sup_{\|\boldsymbol{v}\|_2=1} \|\langle \boldsymbol{x}^{(p)}, \boldsymbol{v}\rangle\|_{\psi_2} \le K \tag{564}$$

for a constant $K < \infty$ independent of dimension. No independence between the coordinates of $\boldsymbol{x}^{(p)}$ is required.

Proposition C.28 (Covariance concentration)**.** *Under equation 551 and equation 564, there exist constants $C_K, c_K > 0$ depending only on K such that, for every $u \ge 0$,*

$$\mathbb{P}\left(\|\boldsymbol{S} - \boldsymbol{I}_p\|_{\mathrm{op}} > C_K\left[\sqrt{\frac{p+u}{m}} + \frac{p+u}{m}\right]\right) \le 2e^{-c_K u}. \tag{565}$$

In particular, if $p/m \to 0$, then

$$\|\boldsymbol{S} - \boldsymbol{I}_p\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{566}$$

Furthermore, for every fixed $q \in [1, \infty)$,

$$\left(\mathbb{E}\|\boldsymbol{S} - \boldsymbol{I}_p\|_{\mathrm{op}}^q\right)^{1/q} \le C_{K,q}\left(\sqrt{\frac{p}{m}} + \frac{p}{m}\right), \tag{567}$$

for all sufficiently large p. In particular, equation 560 holds.

Proof. Fix $\boldsymbol{v} \in \mathbb{S}^{p-1}$. By isotropy,

$$\boldsymbol{v}^\top(\boldsymbol{S} - \boldsymbol{I}_p)\boldsymbol{v} = \frac{1}{m}\sum_{j=1}^{m}\left(\langle \boldsymbol{x}_j, \boldsymbol{v}\rangle^2 - 1\right). \tag{568}$$

For

$$Y_j(\boldsymbol{v}) := \langle \boldsymbol{x}_j, \boldsymbol{v}\rangle^2 - 1, \tag{569}$$

we have $\mathbb{E}Y_j(\boldsymbol{v}) = 0$. The elementary implication "sub-Gaussian squared is sub-exponential" gives

$$\|Y_j(\boldsymbol{v})\|_{\psi_1} \le CK^2. \tag{570}$$

For completeness, before centering one has directly from equation 563

$$\|\langle \boldsymbol{x}_j, \boldsymbol{v}\rangle^2\|_{\psi_1} = \|\langle \boldsymbol{x}_j, \boldsymbol{v}\rangle\|_{\psi_2}^2 \le K^2, \tag{571}$$

and centering a sub-exponential random variable changes its ψ_1 norm by at most a universal multiplicative constant, yielding equation 570.

The scalar Bernstein inequality for independent centered sub-exponential variables therefore gives universal constants $c, C > 0$ such that

$$\mathbb{P}\left(|\boldsymbol{v}^\top(\boldsymbol{S} - \boldsymbol{I}_p)\boldsymbol{v}| > t\right) \le 2\exp\left[-cm\min\left\{\frac{t^2}{K^4}, \frac{t}{K^2}\right\}\right]. \tag{572}$$

Let $\mathcal{N}$ be a $1/4$-net of the Euclidean unit sphere. A volumetric argument gives

$$|\mathcal{N}| \le 9^p. \tag{573}$$

For every symmetric $p \times p$ matrix $\boldsymbol{M}$,

$$\|\boldsymbol{M}\|_{\mathrm{op}} \le 2\max_{\boldsymbol{v}\in\mathcal{N}} |\boldsymbol{v}^\top \boldsymbol{M}\boldsymbol{v}|. \tag{574}$$

Indeed, choose a unit vector $\boldsymbol{x}$ for which $|\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x}| = \|\boldsymbol{M}\|_{\mathrm{op}}$, and choose $\boldsymbol{v} \in \mathcal{N}$ with $\|\boldsymbol{x} - \boldsymbol{v}\|_2 \le 1/4$. Then

$$|\boldsymbol{x}^\top \boldsymbol{M}\boldsymbol{x} - \boldsymbol{v}^\top \boldsymbol{M}\boldsymbol{v}| \le |(\boldsymbol{x} - \boldsymbol{v})^\top \boldsymbol{M}\boldsymbol{x}| + |\boldsymbol{v}^\top \boldsymbol{M}(\boldsymbol{x} - \boldsymbol{v})| \le \frac{1}{2}\|\boldsymbol{M}\|_{\mathrm{op}}, \tag{575}$$

which implies equation 574.

Combining equation 572, equation 573, and a union bound gives

$$\mathbb{P}\left(\|\boldsymbol{S}-\boldsymbol{I}_p\|_{\mathrm{op}}>2t\right)\le 2\exp\left[p\log 9-cm\min\left\{\frac{t^2}{K^4},\frac{t}{K^2}\right\}\right]. \tag{576}$$

Taking

$$t=C_K\left[\sqrt{\frac{p+u}{m}}+\frac{p+u}{m}\right] \tag{577}$$

with C_K sufficiently large proves equation 565 after a harmless adjustment of constants.

When $p/m\to 0$, the right-hand side in brackets in equation 577 converges to zero for each fixed u, which proves equation 566. Finally, integrating the tail bound equation 565 yields, for every fixed $q\ge 1$,

$$\mathbb{E}\|\boldsymbol{S}-\boldsymbol{I}_p\|_{\mathrm{op}}^q\le C_{K,q}\left[\left(\frac{p}{m}\right)^{q/2}+\left(\frac{p}{m}\right)^q\right], \tag{578}$$

which is equivalent, after changing the constant, to equation 567. □

Remark C.29 (Isotropy alone does not imply the large-data reduction)**.** The concentration assumption equation 45 cannot be deduced from isotropy and $p/m\to 0$ alone. For example, let J be uniform on $\{1,\dots,p\}$, let ϵ be an independent Rademacher variable, and set

$$\boldsymbol{x}=\sqrt{p}\,\epsilon e_J. \tag{579}$$

Then

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]=\boldsymbol{I}_p. \tag{580}$$

Take

$$m=p\log\log p. \tag{581}$$

Then $p/m\to 0$, but if N_0 is the number of coordinates that are never sampled, one has

$$\mathbb{E}N_0=p\left(1-\frac{1}{p}\right)^m\sim\frac{p}{\log p}\longrightarrow\infty. \tag{582}$$

Moreover, if I_j is the indicator that coordinate j is unsampled, then for $j\neq k$,

$$\mathrm{Cov}(I_j,I_k)=\left(1-\frac{2}{p}\right)^m-\left(1-\frac{1}{p}\right)^{2m}\le 0. \tag{583}$$

Consequently,

$$\mathrm{Var}(N_0)\le\mathbb{E}N_0,\qquad\mathbb{P}(N_0=0)\le\frac{1}{\mathbb{E}N_0}\longrightarrow 0. \tag{584}$$

With probability tending to one there is therefore a coordinate $\boldsymbol{e}_j$ for which $Se_j=0$, and on that event

$$\|\boldsymbol{S}-\boldsymbol{I}_p\|_{\mathrm{op}}\ge 1. \tag{585}$$

Thus some concentration or regularity assumption beyond mere second-moment isotropy is indispensable.

Exact solution of the population-covariance proxy. Lemma C.7, specialized to $\boldsymbol{S}=\boldsymbol{I}_p$, gives

$$\boldsymbol{A}_{\boldsymbol{I}}(t)=\Theta_t(\boldsymbol{A}_0), \tag{586}$$

where Θ_t is defined in equation 553. Thus the dynamical part of the large-data approximation is already deterministic once $\|\boldsymbol{S}-\boldsymbol{I}_p\|_{\mathrm{op}}\to 0$.

The initialization spectral limit. Under the standing initialization Eq. (36),

$$\boldsymbol{A}_p(0) = \boldsymbol{U}(0)^\top \boldsymbol{U}(0) = \rho_p \boldsymbol{W}_p, \qquad \boldsymbol{W}_p := \frac{1}{n} Z^\top \boldsymbol{Z}. \tag{587}$$

We use the following standard random-matrix facts. If $p/n \to \psi \in (0, \infty)$ and $\boldsymbol{Z}$ has independent standard Gaussian entries, then almost surely

$$\mu_{\boldsymbol{W}_p} := \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(\boldsymbol{W}_p)} \Longrightarrow \mu_\psi, \tag{588}$$

and

$$\lambda_{\max}(\boldsymbol{W}_p) \longrightarrow (1 + \sqrt{\psi})^2. \tag{589}$$

The first statement is the Marchenko–Pastur theorem, and the second is the corresponding extreme-eigenvalue theorem. In particular,

$$\|\boldsymbol{A}_p(0)\|_{\mathrm{op}} = O_{\mathbb{P}}(1). \tag{590}$$

The Gaussian operator-norm tail bound

$$\mathbb{P}\left(\|\boldsymbol{Z}\|_{\mathrm{op}} > \sqrt{n} + \sqrt{p} + s\right) \le 2e^{-s^2/2} \tag{591}$$

also implies that, for every fixed $q < \infty$,

$$\sup_p \mathbb{E}\|\boldsymbol{A}_p(0)\|_{\mathrm{op}}^q < \infty. \tag{592}$$

We next turn the pointwise Marchenko–Pastur convergence into uniform convergence of the loss curve.

Lemma C.30 (Uniform spectral convergence of the proxy loss)**.** *Let*

$$L_p^I(t) := \frac{1}{2p} \operatorname{Tr}\big(\boldsymbol{I}_p - \Theta_t(\boldsymbol{A}_p(0))\big)^2. \tag{593}$$

Then, for every $T < \infty$,

$$\sup_{0 \le t \le T} |L_p^I(t) - \mathcal{L}_{\psi,\rho}(t)| \longrightarrow 0 \tag{594}$$

almost surely along the Gaussian initialization sequence.

Proof. Using equation 587,

$$L_p^I(t) = \frac{1}{p} \sum_{i=1}^{p} g_p(t, \lambda_i(\boldsymbol{W}_p)), \tag{595}$$

where

$$g_p(t, \lambda) := \frac{1}{2}\big[\Theta_t(\rho_p \lambda) - 1\big]^2. \tag{596}$$

Similarly, define

$$g(t, \lambda) := \frac{1}{2}\big[\Theta_t(\rho \lambda) - 1\big]^2. \tag{597}$$

By equation 589, almost surely there is a finite M such that, for all sufficiently large p,

$$\operatorname{spec}(\boldsymbol{W}_p) \subset [0, M]. \tag{598}$$

The map

$$(t, \lambda, r) \longmapsto \frac{1}{2}[\Theta_t(r\lambda) - 1]^2 \tag{599}$$

is continuous on every compact subset of $[0, \infty)^3$. Since $\rho_p \to \rho$, it follows that

$$\sup_{\substack{0 \le t \le T \\ 0 \le \lambda \le M}} |g_p(t, \lambda) - g(t, \lambda)| \longrightarrow 0. \tag{600}$$

It remains to prove

$$\sup_{0 \le t \le T} \left| \int g(t, \lambda)\, \mu_{\boldsymbol{W}_p}(d\lambda) - \int g(t, \lambda)\, \mu_\psi(d\lambda) \right| \longrightarrow 0. \tag{601}$$

The function g is uniformly continuous on $[0,T]\times[0,M]$. Therefore, for every $\delta>0$, there exist times $t_1,\dots,t_N\in[0,T]$ such that for every $t\in[0,T]$ there is an index j with

$$\sup_{0\le\lambda\le M}|g(t,\lambda)-g(t_j,\lambda)|\le\delta. \tag{602}$$

For each fixed j, weak convergence equation 588 gives

$$\int g(t_j,\lambda)\,\mu_{\boldsymbol{W}_p}(d\lambda)\longrightarrow\int g(t_j,\lambda)\,\mu_{\psi}(d\lambda). \tag{603}$$

Since the collection of t_j is finite, equation 602 and equation 603 imply

$$\limsup_{p\to\infty}\sup_{0\le t\le T}\left|\int g(t,\lambda)\,\mu_{\boldsymbol{W}_p}(d\lambda)-\int g(t,\lambda)\,\mu_{\psi}(d\lambda)\right|\le 2\delta. \tag{604}$$

Letting $\delta\downarrow 0$ proves equation 601. Together with equation 600, this proves equation 594. □

Proof of Theorem C.27. Let

$$R_p:=\max\{1,\|\boldsymbol{A}_p(0)\|_{\mathrm{op}}\}. \tag{605}$$

By equation 590, the sequence $(R_p)_p$ is tight. Fix $\delta>0$. We can therefore choose a deterministic $R<\infty$ such that

$$\limsup_{p\to\infty}\mathbb{P}(R_p>R)\le\delta. \tag{606}$$

On the event

$$\{R_p\le R\}\cap\{\varepsilon_p\le 1\}, \tag{607}$$

Proposition C.5 and Lemma C.7 give

$$\sup_{0\le t\le T}\|\boldsymbol{A}_p(t)-\Theta_t(\boldsymbol{A}_p(0))\|_{\mathrm{op}}\le K_{T,R}\varepsilon_p. \tag{608}$$

Since $\varepsilon_p\to 0$ in probability, while δ is arbitrary, equation 608 proves equation 557.

Similarly, Lemma C.6 gives on the same event

$$\sup_{0\le t\le T}|L_p(t)-L_p^I(t)|\le(1+R)K_{T,R}\varepsilon_p. \tag{609}$$

Combining this with Lemma C.30 proves equation 558. Also,

$$\sup_{0\le t\le T}|\widehat{L}_p(t)-L_p(t)|\le\frac{1}{2}(1+R_p)^2\varepsilon_p, \tag{610}$$

and therefore equation 559 follows.

It remains to justify the L^1 assertions. Lemma C.2 gives

$$\sup_{t\ge 0}L_p(t)\le\frac{1}{2}(1+R_p)^2. \tag{611}$$

Likewise,

$$\sup_{t\ge 0}\widehat{L}_p(t)\le\frac{1}{2}\|\boldsymbol{S}\|_{\mathrm{op}}(1+R_p)^2. \tag{612}$$

By equation 592, $(R_p)_p$ has uniformly bounded moments of every fixed order. Under the additional assumption in Theorem C.27, the right-hand sides of equation 611 and equation 612 are uniformly integrable. Therefore the convergence in probability of the suprema in equation 558 and equation 559 upgrades to L^1 convergence, proving equation 561 and equation 562.

In particular, under the sub-Gaussian hypothesis equation 564, Proposition C.28 verifies all the required data-side assumptions.

Closed-form evaluation of the limiting loss. We now evaluate equation 556. Define

$$q_t := \rho(1 - e^{4t}) \le 0. \tag{613}$$

From equation 553,

$$\Theta_t(\rho\lambda) = \frac{\rho\lambda e^{4t}}{1 - q_t\lambda}. \tag{614}$$

Hence

$$\begin{aligned}\big(\Theta_t(\rho\lambda) - 1\big)^2 = 1 - 2\rho e^{4t}\frac{\lambda}{1 - q_t\lambda} \\ + \rho^2 e^{8t}\frac{\lambda^2}{(1 - q_t\lambda)^2}.\end{aligned} \tag{615}$$

Introduce

$$H_\psi(q) := \int \frac{\lambda}{1 - q\lambda}\,\mu_\psi(d\lambda), \qquad q \le 0. \tag{616}$$

Since the Marchenko–Pastur distribution has compact support on its nonzero part and $1 - q\lambda \ge 1$ for $q \le 0$, differentiation under the integral gives

$$H'_\psi(q) = \int \frac{\lambda^2}{(1 - q\lambda)^2}\,\mu_\psi(d\lambda). \tag{617}$$

Substitution into equation 556 yields

$$\boxed{\mathcal{L}_{\psi,\rho}(t) = \frac{1}{2} - \rho e^{4t} H_\psi(q_t) + \frac{\rho^2 e^{8t}}{2} H'_\psi(q_t), \qquad q_t = \rho(1 - e^{4t}).} \tag{618}$$

For completeness, we also derive the explicit expression for H_ψ. Let

$$M_\psi(q) := \int \frac{1}{1 - q\lambda}\,\mu_\psi(d\lambda). \tag{619}$$

The Marchenko–Pastur Stieltjes-transform equation, expressed in the variable q, is

$$\psi q M_\psi(q)^2 - \big[1 + (\psi - 1)q\big] M_\psi(q) + 1 = 0. \tag{620}$$

The solution satisfying $M_\psi(0) = 1$ is

$$M_\psi(q) = \frac{1 + (\psi - 1)q - \sqrt{1 - 2(1 + \psi)q + (1 - \psi)^2 q^2}}{2\psi q}, \tag{621}$$

with the value at $q = 0$ understood by continuity. Since

$$M_\psi(q) = 1 + qH_\psi(q), \tag{622}$$

we obtain

$$\boxed{H_\psi(q) = \frac{1 - (1 + \psi)q - \sqrt{1 - 2(1 + \psi)q + (1 - \psi)^2 q^2}}{2\psi q^2}.} \tag{623}$$

The branch of the square root is the one for which H_ψ is regular at the origin. Expanding equation 623 at $q = 0$ gives

$$H_\psi(0) = 1, \qquad H'_\psi(0) = 1 + \psi. \tag{624}$$

Equations (618) and (623) reproduce the closed large-data formula in the main text.

C.4.1 Large-time behavior

We emphasize that the results in this subsection concern the large-time behavior of the already established proportional-limit curve $\mathcal{L}_{\psi,\rho}(t)$. Thus the order of limits is

$$\lim_{t\to\infty} \lim_{\substack{p,n,m\to\infty \\ p/n\to\psi,\ p/m\to 0}} L_p(t), \tag{625}$$

and the compact-time estimates proved above do not by themselves justify interchanging these two limits.

For every $s > 0$,

$$\lim_{t\to\infty} \Theta_t(s) = 1, \qquad \Theta_t(0) = 0. \tag{626}$$

The Marchenko–Pastur measure has an atom at zero of mass $(1 - \psi^{-1})_+$. Dominated convergence in equation 556 therefore yields

$$\boxed{\mathcal{L}_{\psi,\rho}(\infty) = \frac{1}{2}\left(1 - \frac{1}{\psi}\right)_+.} \tag{627}$$

The same conclusion can also be read directly from the finite dimensional population-covariance flow. If $a_i(0) > 0$, then equation 103 implies

$$a_i(t) - 1 = \frac{(a_i(0) - 1)e^{-4t}}{a_i(0) + (1 - a_i(0))e^{-4t}}. \tag{628}$$

If the positive initial eigenvalues are fixed and bounded away from zero, then

$$\begin{aligned} L_p^I(t) - L_p^I(\infty) = \frac{e^{-8t}}{2p} \sum_{i:\, a_i(0)>0} \left(\frac{1 - a_i(0)}{a_i(0)}\right)^2 \\ + O(e^{-12t}). \end{aligned} \tag{629}$$

The prefactor can, however, become singular when the smallest positive initial eigenvalues tend to zero. This is precisely what happens at the critical proportional aspect ratio $\psi = 1$.

Noncritical aspect ratio. Assume $\psi \neq 1$ and put

$$\epsilon := e^{-4t}. \tag{630}$$

For every $s > 0$,

$$1 - \Theta_t(s) = \frac{(1 - s)\epsilon}{s + (1 - s)\epsilon}. \tag{631}$$

For $\psi \neq 1$, the positive part of the Marchenko–Pastur support is separated from zero:

$$\inf\{\lambda > 0 : \lambda \in \operatorname{supp} \mu_\psi\} = (1 - \sqrt{\psi})^2 > 0. \tag{632}$$

Consequently, uniformly over the positive support,

$$\big(1 - \Theta_t(\rho\lambda)\big)^2 = \epsilon^2 \left(\frac{1 - \rho\lambda}{\rho\lambda}\right)^2 + O(\epsilon^3). \tag{633}$$

The zero atom, when $\psi > 1$, contributes exactly $\mathcal{L}_{\psi,\rho}(\infty)$. Therefore

$$\boxed{\mathcal{L}_{\psi,\rho}(t) - \mathcal{L}_{\psi,\rho}(\infty) = C_{\psi,\rho} e^{-8t} + O(e^{-12t}), \qquad \psi \neq 1,} \tag{634}$$

where

$$C_{\psi,\rho} := \frac{1}{2} \int_{\lambda>0} \left(\frac{1 - \rho\lambda}{\rho\lambda}\right)^2 \mu_\psi(d\lambda). \tag{635}$$

Critical aspect ratio. We now assume $\psi = 1$. In this case

$$\mu_1(d\lambda) = \frac{1}{2\pi} \sqrt{\frac{4 - \lambda}{\lambda}} \mathbf{1}_{\{0<\lambda<4\}}\, d\lambda. \tag{636}$$

Since there is no atom at zero,

$$\mathcal{L}_{1,\rho}(\infty) = 0. \tag{637}$$

Using equation 631 with $s = \rho\lambda$ gives

$$\mathcal{L}_{1,\rho}(t) = \frac{1}{4\pi} \int_0^4 \left[\frac{\epsilon(1 - \rho\lambda)}{\rho\lambda + (1 - \rho\lambda)\epsilon}\right]^2 \sqrt{\frac{4 - \lambda}{\lambda}}\, d\lambda. \tag{638}$$

Choose a fixed

$$0 < \delta < \min\left\{4, \frac{1}{2\rho}\right\}. \tag{639}$$

On $\lambda \in [\delta, 4]$, for all sufficiently small ϵ, the denominator in equation 638 is bounded below by a positive constant depending only on δ and ρ. Therefore

$$\int_\delta^4 \left[\frac{\epsilon(1-\rho\lambda)}{\rho\lambda + (1-\rho\lambda)\epsilon}\right]^2 \sqrt{\frac{4-\lambda}{\lambda}}\, d\lambda = O(\epsilon^2). \tag{640}$$

Thus the leading contribution comes from λ of order ϵ. In the integral over $(0, \delta)$, set

$$\lambda = \frac{\epsilon y}{\rho}. \tag{641}$$

Then

$$\frac{\epsilon(1-\rho\lambda)}{\rho\lambda + (1-\rho\lambda)\epsilon} = \frac{1-\epsilon y}{1+y-\epsilon y}, \tag{642}$$

and a direct substitution yields

$$\begin{aligned}\frac{1}{\sqrt{\epsilon}}\mathcal{L}_{1,\rho}^{(0,\delta)}(t) = \frac{1}{4\pi\sqrt{\rho}} \int_0^{\rho\delta/\epsilon} \left(\frac{1-\epsilon y}{1+y-\epsilon y}\right)^2 \\ \times \sqrt{4 - \frac{\epsilon y}{\rho}}\, y^{-1/2}\, dy.\end{aligned} \tag{643}$$

For every fixed $y > 0$, the integrand in equation 643 converges to

$$\frac{2y^{-1/2}}{(1+y)^2}. \tag{644}$$

Moreover, because $\rho\lambda \le \rho\delta < 1/2$ on the region under consideration, for sufficiently small ϵ the integrand in equation 643 is bounded by

$$C\frac{y^{-1/2}}{(1+y)^2}, \tag{645}$$

where the right-hand side is integrable on $(0, \infty)$. Dominated convergence therefore gives

$$\lim_{t\to\infty} \frac{\mathcal{L}_{1,\rho}(t)}{\sqrt{\epsilon}} = \frac{1}{2\pi\sqrt{\rho}} \int_0^\infty \frac{y^{-1/2}}{(1+y)^2}\, dy. \tag{646}$$

The remaining integral is a beta integral:

$$\int_0^\infty \frac{y^{-1/2}}{(1+y)^2}\, dy = B\left(\frac{1}{2}, \frac{3}{2}\right) = \frac{\pi}{2}. \tag{647}$$

Since $\sqrt{\epsilon} = e^{-2t}$, we conclude that

$$\boxed{\mathcal{L}_{1,\rho}(t) \sim \frac{e^{-2t}}{4\sqrt{\rho}}, \qquad t \to \infty.} \tag{648}$$

Thus the critical proportional limit has a slower e^{-2t} decay, even though at any fixed finite dimension the idealized population-covariance flow with full-rank initialization eventually has the e^{-8t} loss decay in equation 629. The discrepancy is caused by the hard edge of the Marchenko–Pastur spectrum and is an explicit noncommutation of the large-dimension and large-time limits.

Possible extension to anisotropic data. The deterministic part of the argument extends directly to non-isotropic data. Suppose instead that

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{\Sigma}_p, \qquad \boldsymbol{\Sigma}_p \succeq 0, \qquad \sup_p \|\boldsymbol{\Sigma}_p\|_{\mathrm{op}} < \infty, \tag{649}$$

and assume the empirical covariance satisfies

$$\|\boldsymbol{S} - \boldsymbol{\Sigma}_p\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{650}$$

The exact population loss then becomes

$$L_{\boldsymbol{\Sigma}_p}(\boldsymbol{A}) = \frac{1}{2p}\operatorname{Tr}\big[\boldsymbol{\Sigma}_p(\boldsymbol{I}_p - \boldsymbol{A})^2\big]. \tag{651}$$

Repeating Proposition C.5, with $\boldsymbol{I}_p$ replaced by $\boldsymbol{\Sigma}_p$, shows that on every fixed time interval the empirical trajectory is close to the deterministic population-covariance flow

$$\dot{\boldsymbol{A}}_\Sigma = 2(\boldsymbol{\Sigma}_p\boldsymbol{A}_\Sigma + \boldsymbol{A}_\Sigma\boldsymbol{\Sigma}_p) - \boldsymbol{\Sigma}_p\boldsymbol{A}_\Sigma^2 - \boldsymbol{A}_\Sigma^2\boldsymbol{\Sigma}_p - 2\boldsymbol{A}_\Sigma\boldsymbol{\Sigma}_p\boldsymbol{A}_\Sigma. \tag{652}$$

Indeed, on bounded operator-norm sets the same calculation gives

$$\sup_{0\le t\le T}\|\boldsymbol{A}_{\boldsymbol{S}}(t) - \boldsymbol{A}_\Sigma(t)\|_{\mathrm{op}} \le C_{T,R,\sup_p\|\boldsymbol{\Sigma}_p\|_{\mathrm{op}}}\|\boldsymbol{S} - \boldsymbol{\Sigma}_p\|_{\mathrm{op}}. \tag{653}$$

Thus the large-data empirical-to-population reduction remains straightforward.

What is special about isotropy is not the concentration step but the subsequent solvability. For $\boldsymbol{\Sigma}_p = \boldsymbol{I}_p$, equation 652 collapses to the scalar spectral logistic flow equation 586. For a general covariance, $\boldsymbol{A}_\Sigma(t)$ need not commute with $\boldsymbol{\Sigma}_p$, and the eigenvectors of $\boldsymbol{A}_\Sigma(t)$ may rotate.

If the initial Gram matrix happens to commute with the covariance,

$$[\boldsymbol{A}_0, \boldsymbol{\Sigma}_p] = 0, \tag{654}$$

then the two matrices can be diagonalized simultaneously:

$$\boldsymbol{\Sigma}_p = \operatorname{diag}(\lambda_1, \dots, \lambda_p), \qquad \boldsymbol{A}_0 = \operatorname{diag}(a_1(0), \dots, a_p(0)). \tag{655}$$

Equation (652) then reduces to

$$\dot{a}_i(t) = 4\lambda_i a_i(t)(1 - a_i(t)), \tag{656}$$

with solution

$$a_i(t) = \Theta_{\lambda_i t}(a_i(0)). \tag{657}$$

For generic Gaussian initialization in the proportional regime $p/n \to \psi > 0$, however, $\boldsymbol{A}_0$ does not commute with a deterministic non-scalar $\boldsymbol{\Sigma}_p$. The large-dimensional limit is then genuinely noncommutative. One possible route is to combine the deterministic stability estimate equation 653 with asymptotic freeness of the orthogonally invariant Wishart initialization and $\boldsymbol{\Sigma}_p$, leading to an operator-valued version of equation 652. Establishing and analyzing that limit requires additional arguments and lies beyond the isotropic large-data result proved here.

C.5 Small-data (Section 4.4)

The small-data proof also starts from the common identities of Appendix C.1, but the natural data normalization is the one in equation 47. We therefore do not repeat the definitions of $\boldsymbol{S}_N$, the error matrix, the losses, or the full gradient flow. The regime-specific dimensions satisfy

$$p = p_N \longrightarrow \infty, \qquad n = n_N \longrightarrow \infty, \qquad m = m_N \longrightarrow \infty, \tag{658}$$

and

$$\phi_N := \frac{p}{m} \longrightarrow \infty, \qquad \psi_N := \frac{p}{n} \longrightarrow \psi \in (0,\infty), \qquad \rho_N := n\sigma_N^2 \longrightarrow \rho \in (0,\infty). \tag{659}$$

In particular,

$$\frac{m}{n} = \frac{\psi_N}{\phi_N} \longrightarrow 0. \tag{660}$$

The training matrix may be random or deterministic. We assume the common active-space near-isometry equation 48. In particular the bounded-spectrum and spectral-limit requirements of Appendix C.1 hold for the normalized operator $\widetilde{\boldsymbol{S}}_N = \phi_N^{-1}\boldsymbol{S}_N$; no Gaussianity of the training samples is required.

We use the small-data learning rate $\eta = m$, so

$$\gamma_N = \frac{\eta}{p} = \frac{1}{\phi_N}. \tag{661}$$

Thus the full dynamics are exactly the common flow equation 49 with $\gamma = \phi_N^{-1}$. For the population loss we assume only second-moment isotropy of the test distribution,

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{I}_p, \tag{662}$$

which, by equation 38, gives the common isotropic population loss equation 39. The train-loss identity is likewise already contained in equation 38.

Main theorem. Let

$$\mathrm{Nar}(k,j) := \frac{1}{k}\binom{k}{j}\binom{k}{j-1}, \qquad 1 \le j \le k, \tag{663}$$

denote the Narayana numbers.

Theorem C.31 (Small-data finite-time limit)**.** *Assume equation 658, equation 659, and the common active-space near-isometry condition equation 48, and fix $T < \infty$. There is a unique finite-time solution*

$$(a(t), \mu_1(t), \mu_2(t), \dots), \qquad 0 \le t \le T, \tag{664}$$

in the class of moment sequences with uniformly bounded support, satisfying

$$\dot{a} = 4a(1-a) - 2a\mu_1, \qquad a(0) = \rho, \tag{665}$$

and, for every $k \ge 1$,

$$\dot{\mu}_k = -2\mu_{k+1} - 2a\sum_{j=0}^{k-1}\mu_j\mu_{k-j} + 2\mu_1\mu_k, \qquad \mu_0 \equiv 1, \tag{666}$$

with initial conditions

$$\mu_k(0) = \rho^k \sum_{j=1}^{k}\mathrm{Nar}(k,j)\psi^j, \qquad k \ge 1. \tag{667}$$

More precisely, after rotating coordinates so that the training span is the first m coordinate directions and writing

$$\boldsymbol{U}_N(t) = [\boldsymbol{A}_N(t) \quad \boldsymbol{B}_N(t)], \qquad \boldsymbol{M}_N(t) := \boldsymbol{B}_N(t)\boldsymbol{B}_N(t)^\top, \tag{668}$$

one has, for every fixed $k \ge 0$,

$$\sup_{0\le t\le T}\left\|\boldsymbol{A}_N(t)^\top \boldsymbol{M}_N(t)^k \boldsymbol{A}_N(t) - a(t)\mu_k(t)\boldsymbol{I}_m\right\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0, \tag{669}$$

where $\mu_0 \equiv 1$.

The train loss converges uniformly on compact time intervals according to

$$\sup_{0\le t\le T}\left|\widehat{L}_N(t) - \left\{\frac{1}{2}(1-a(t))^2 + \frac{1}{2}a(t)\mu_1(t)\right\}\right| \xrightarrow{\mathbb{P}} 0. \tag{670}$$

The population loss has the constant leading-order limit

$$L_\infty^{(0)} := \frac{1}{2}\left[1 - 2\rho + (1+\psi)\rho^2\right], \tag{671}$$

in the sense that

$$\sup_{0\le t\le T}\left|L_N(t) - L_\infty^{(0)}\right| \xrightarrow{\mathbb{P}} 0. \tag{672}$$

Moreover, the first nontrivial population correction is obtained without any further assumption on the data beyond equation 48:

$$\sup_{0\le t\le T}\left|\phi_N\big(L_N(t) - L_N(0)\big) - C(t)\right| \xrightarrow{\mathbb{P}} 0, \tag{673}$$

where

$$C(t) = \frac{1}{2}(1-a(t))^2 + a(t)\mu_1(t) - \frac{1}{2}(1-\rho)^2 - \psi\rho^2 + 2\int_0^t a(s)\big[\mu_1(s) - \mu_2(s)\big]\,ds. \tag{674}$$

In particular, $C(0) = 0$.

The centered formulation equation 673 is convenient because it does not require a rate of convergence of ρ_N or ψ_N. If these ratios are fixed to greater accuracy than $1/\phi_N$, it is equivalent to the usual absolute $1/\phi_N$ expansion.

Proof plan. The proof is divided into the following steps.

1. We first replace the data by exactly orthogonalized data. In this case the active and inactive weight blocks satisfy an exact reduced matrix ODE.
2. For the orthogonalized system we introduce $\boldsymbol{Q}_{k,N} = \boldsymbol{A}_N^\top \boldsymbol{M}_N^k \boldsymbol{A}_N$. These matrices satisfy an exact closed infinite hierarchy. The submanifold $\boldsymbol{Q}_{k,N} = c_k \boldsymbol{I}_m$ is exactly invariant.
3. We show that the random Gaussian initialization is asymptotically scalar in the strong sense $\boldsymbol{Q}_{k,N}(0) \simeq c_k(0)\boldsymbol{I}_m$. The condition $m/n \to 0$ is precisely what turns quadratic-form concentration into operator-norm concentration on the m-dimensional active block.
4. We place the matrix hierarchy in a weighted Banach algebra and use a dimension-independent Gronwall estimate. This propagates initial scalarity uniformly for $0 \le t \le T$ and rigorously produces the scalar moment hierarchy equation 665–equation 666.
5. The train-loss limit and the $1/\phi_N$ population correction are then direct consequences of exact trace identities.
6. Finally, for a general training matrix satisfying equation 48, we use the polar decomposition of $\boldsymbol{X}_N$. The normalized Gram matrix $\boldsymbol{\Gamma}_N$ appears as a perturbation of the orthogonalized block equations. A finite-time ODE stability estimate transfers all of the preceding limits to the original data.

C.5.1 Exact orthogonal-data reduction

We first consider the reference model

$$\boldsymbol{X}_N = \sqrt{p}\,\boldsymbol{Q}_N, \qquad \boldsymbol{Q}_N^\top \boldsymbol{Q}_N = \boldsymbol{I}_m. \tag{675}$$

After a fixed orthogonal change of coordinates we may assume that

$$\boldsymbol{Q}_N = \begin{pmatrix} \boldsymbol{I}_m \\ 0 \end{pmatrix}. \tag{676}$$

The right-orthogonal invariance of the standing initialization in Eq. (36) implies that this rotation does not change its law.

For equation 675, the empirical covariance is

$$\boldsymbol{S}_N = \phi_N \boldsymbol{P}_m, \qquad \boldsymbol{P}_m = \begin{pmatrix} \boldsymbol{I}_m & 0 \\ 0 & 0 \end{pmatrix}. \tag{677}$$

Thus the common flow equation 49, with $\gamma = \phi_N^{-1}$, reduces exactly to

$$\dot{\boldsymbol{U}}_N = \boldsymbol{U}_N(\boldsymbol{P}_m \boldsymbol{E}_N + \boldsymbol{E}_N \boldsymbol{P}_m). \tag{678}$$

With the decomposition equation 668,

$$\boldsymbol{E}_N = \begin{pmatrix} \boldsymbol{I}_m - \boldsymbol{A}_N^\top \boldsymbol{A}_N & -\boldsymbol{A}_N^\top \boldsymbol{B}_N \\ -\boldsymbol{B}_N^\top \boldsymbol{A}_N & \boldsymbol{I}_{p-m} - \boldsymbol{B}_N^\top \boldsymbol{B}_N \end{pmatrix}. \tag{679}$$

Multiplying the blocks in equation 678 gives

$$\dot{\boldsymbol{A}}_N = 2\boldsymbol{A}_N(\boldsymbol{I}_m - \boldsymbol{A}_N^\top \boldsymbol{A}_N) - \boldsymbol{M}_N \boldsymbol{A}_N, \tag{680}$$

and

$$\dot{\boldsymbol{B}}_N = -\boldsymbol{A}_N \boldsymbol{A}_N^\top \boldsymbol{B}_N. \tag{681}$$

Consequently,

$$\dot{\boldsymbol{M}}_N = -\boldsymbol{A}_N \boldsymbol{A}_N^\top \boldsymbol{M}_N - \boldsymbol{M}_N \boldsymbol{A}_N \boldsymbol{A}_N^\top. \tag{682}$$

Equations (680) to (682) are exact.

C.5.2 THE EXACT MATRIX MOMENT HIERARCHY

For every integer $k \geq 0$, define

$$\boldsymbol{Q}_{k,N}(t) := \boldsymbol{A}_N(t)^\top \boldsymbol{M}_N(t)^k \boldsymbol{A}_N(t). \tag{683}$$

In particular,

$$\boldsymbol{Q}_{0,N} = \boldsymbol{A}_N^\top \boldsymbol{A}_N. \tag{684}$$

We write the resulting matrix sequence as

$$\mathbf{Q}_N(t) := \big(\boldsymbol{Q}_{0,N}(t), \boldsymbol{Q}_{1,N}(t), \ldots\big). \tag{685}$$

Lemma C.32 (Exact matrix hierarchy)**.** *For the orthogonal-data dynamics equation 680–equation 682,*

$$\begin{aligned}\dot{\boldsymbol{Q}}_{k,N} = 4\boldsymbol{Q}_{k,N} - 2\boldsymbol{Q}_{0,N}\boldsymbol{Q}_{k,N} - 2\boldsymbol{Q}_{k,N}\boldsymbol{Q}_{0,N} - 2\boldsymbol{Q}_{k+1,N} \\ - \sum_{j=0}^{k-1} \left(\boldsymbol{Q}_{j,N}\boldsymbol{Q}_{k-j,N} + \boldsymbol{Q}_{j+1,N}\boldsymbol{Q}_{k-1-j,N}\right), \qquad k \geq 0,\end{aligned} \tag{686}$$

where the sum is empty when $k = 0$.

Proof. From equation 680,

$$\dot{\boldsymbol{A}}_N = 2\boldsymbol{A}_N - 2\boldsymbol{A}_N\boldsymbol{Q}_{0,N} - \boldsymbol{M}_N\boldsymbol{A}_N. \tag{687}$$

Hence

$$\dot{\boldsymbol{A}}_N^\top = 2\boldsymbol{A}_N^\top - 2\boldsymbol{Q}_{0,N}\boldsymbol{A}_N^\top - \boldsymbol{A}_N^\top\boldsymbol{M}_N. \tag{688}$$

Differentiating equation 683, the two terms in which one differentiates $\boldsymbol{A}_N$ give

$$\begin{aligned}&\dot{\boldsymbol{A}}_N^\top \boldsymbol{M}_N^k \boldsymbol{A}_N + \boldsymbol{A}_N^\top \boldsymbol{M}_N^k \dot{\boldsymbol{A}}_N \\ &\quad = 4\boldsymbol{Q}_{k,N} - 2\boldsymbol{Q}_{0,N}\boldsymbol{Q}_{k,N} - 2\boldsymbol{Q}_{k,N}\boldsymbol{Q}_{0,N} - 2\boldsymbol{Q}_{k+1,N}.\end{aligned} \tag{689}$$

For the remaining contribution,

$$\frac{d}{dt}\boldsymbol{M}_N^k = \sum_{j=0}^{k-1} \boldsymbol{M}_N^j \dot{\boldsymbol{M}}_N \boldsymbol{M}_N^{k-1-j}. \tag{690}$$

Substitution of equation 682 gives

$$\begin{aligned}&\boldsymbol{A}_N^\top \boldsymbol{M}_N^j \dot{\boldsymbol{M}}_N \boldsymbol{M}_N^{k-1-j} \boldsymbol{A}_N \\ &\quad = -\boldsymbol{Q}_{j,N}\boldsymbol{Q}_{k-j,N} - \boldsymbol{Q}_{j+1,N}\boldsymbol{Q}_{k-1-j,N}.\end{aligned} \tag{691}$$

Summing equation 691 over j and adding equation 689 proves equation 686. □

For $k = 0$, Lemma C.32 reads

$$\dot{\boldsymbol{Q}}_{0,N} = 4\boldsymbol{Q}_{0,N} - 4\boldsymbol{Q}_{0,N}^2 - 2\boldsymbol{Q}_{1,N}. \tag{692}$$

Since $\boldsymbol{M}_N \succeq 0$, we have

$$\boldsymbol{Q}_{1,N} \succeq 0. \tag{693}$$

The hierarchy equation 686 is already closed at the matrix level. Thus no moment-closure approximation is needed. The only remaining issue is to prove that its random solution becomes scalar in the active m-dimensional coordinates.

C.5.3 Uniform a priori bounds

The first bound is a contraction property of the inactive block.

Lemma C.33 (Inactive contraction)**.** *For every $t \geq 0$ for which the solution exists,*

$$\|\boldsymbol{B}_N(t)\|_{\mathrm{op}} \leq \|\boldsymbol{B}_N(0)\|_{\mathrm{op}}, \tag{694}$$

and therefore

$$\|\boldsymbol{M}_N(t)\|_{\mathrm{op}} \leq \|\boldsymbol{M}_N(0)\|_{\mathrm{op}}. \tag{695}$$

Proof. For every deterministic $\boldsymbol{v} \in \mathbb{R}^{p-m}$,

$$\begin{aligned}\frac{d}{dt}\|\boldsymbol{B}_N(t)\boldsymbol{v}\|^2 &= 2\boldsymbol{v}^\top \boldsymbol{B}_N^\top \dot{\boldsymbol{B}}_N \boldsymbol{v} \\ &= -2\boldsymbol{v}^\top \boldsymbol{B}_N^\top \boldsymbol{A}_N \boldsymbol{A}_N^\top \boldsymbol{B}_N \boldsymbol{v} \\ &= -2\|\boldsymbol{A}_N^\top \boldsymbol{B}_N \boldsymbol{v}\|^2 \leq 0. \end{aligned} \tag{696}$$

Taking the supremum over unit $\boldsymbol{v}$ gives equation 694; squaring yields equation 695. □

The active block is also uniformly bounded.

Lemma C.34 (Active operator-norm bound)**.** *For the orthogonal-data flow,*

$$\|\boldsymbol{A}_N(t)\|_{\mathrm{op}}^2 = \lambda_{\max}(\boldsymbol{Q}_{0,N}(t)) \leq \max\left\{1, \|\boldsymbol{A}_N(0)\|_{\mathrm{op}}^2\right\}. \tag{697}$$

Proof. Let $\lambda_N(t) = \lambda_{\max}(\boldsymbol{Q}_{0,N}(t))$. At any time at which λ_N is differentiable, choose a normalized top eigenvector $\boldsymbol{v}$. Using equation 692 and equation 693,

$$\begin{aligned}\dot{\lambda}_N(t) &= \boldsymbol{v}^\top \dot{\boldsymbol{Q}}_{0,N} \boldsymbol{v} \\ &= 4\lambda_N - 4\lambda_N^2 - 2\boldsymbol{v}^\top \boldsymbol{Q}_{1,N} \boldsymbol{v} \\ &\leq 4\lambda_N(1-\lambda_N). \end{aligned} \tag{698}$$

The same inequality holds for the upper Dini derivative. Comparison with the scalar logistic equation gives equation 697. □

Combining Lemmas C.33 and C.34, for every $k \geq 0$,

$$\|\boldsymbol{Q}_{k,N}(t)\|_{\mathrm{op}} \leq \|\boldsymbol{A}_N(t)\|_{\mathrm{op}}^2 \|\boldsymbol{M}_N(t)\|_{\mathrm{op}}^k. \tag{699}$$

C.5.4 A Banach algebra for the hierarchy

Fix $L > 0$ and choose $r > 0$ such that

$$rL < 1. \tag{700}$$

For a sequence $\mathbf{R} = (\boldsymbol{R}_0, \boldsymbol{R}_1, \ldots)$ of $m \times m$ matrices, set

$$\|\mathbf{R}\|_r := \sum_{k=0}^{\infty} r^k \|\boldsymbol{R}_k\|_{\mathrm{op}}. \tag{701}$$

Denote the resulting Banach space by $\mathcal{B}_r^{(m)}$.

For two sequences define the convolution

$$(\mathbf{R} * \mathbf{S})_k := \sum_{j=0}^{k} \boldsymbol{R}_j \boldsymbol{S}_{k-j}. \tag{702}$$

The norm equation 701 is submultiplicative:

$$\|\mathbf{R} * \mathbf{S}\|_r \leq \|\mathbf{R}\|_r \|\mathbf{S}\|_r. \tag{703}$$

The left-shift operator

$$(D\mathbf{R})_k := \boldsymbol{R}_{k+1} \tag{704}$$

obeys

$$\|D\mathbf{R}\|_r \le \frac{1}{r}\|\mathbf{R}\|_r. \tag{705}$$

Let $\mathcal{F}_m$ denote the vector field given componentwise by the right-hand side of equation 686. The preceding estimates imply that, on every ball

$$\|\mathbf{R}\|_r \le R, \qquad \|\mathbf{S}\|_r \le R, \tag{706}$$

there is a constant $K_{r,R} < \infty$, independent of m, such that

$$\|\mathcal{F}_m(\mathbf{R}) - \mathcal{F}_m(\mathbf{S})\|_r \le K_{r,R}\|\mathbf{R} - \mathbf{S}\|_r. \tag{707}$$

For example, one may take

$$K_{r,R} = 4 + \frac{2}{r} + 12R. \tag{708}$$

The precise numerical constant is immaterial; what matters is that it does not depend on m.

Indeed, the linear terms are controlled by equation 705, while every quadratic term is controlled by equation 703 and the elementary inequality

$$\|\boldsymbol{R}\boldsymbol{S} - \widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{S}}\| \le \|\boldsymbol{R} - \widetilde{\boldsymbol{R}}\|\|\boldsymbol{S}\| + \|\widetilde{\boldsymbol{R}}\|\|\boldsymbol{S} - \widetilde{\boldsymbol{S}}\|. \tag{709}$$

The scalar subspace is exactly invariant. More precisely, if $\mathbf{c} = (c_0, c_1, \dots)$ is a scalar sequence and

$$\iota_m(\mathbf{c}) := (c_0\boldsymbol{I}_m, c_1\boldsymbol{I}_m, \dots), \tag{710}$$

then there is a scalar vector field $\mathcal{F}_{\mathrm{sc}}$ such that

$$\mathcal{F}_m(\iota_m(\mathbf{c})) = \iota_m\big(\mathcal{F}_{\mathrm{sc}}(\mathbf{c})\big). \tag{711}$$

Explicitly,

$$\dot{c}_0 = 4c_0 - 4c_0^2 - 2c_1, \tag{712}$$

and, for $k \ge 1$,

$$\dot{c}_k = 4c_k - 4c_0c_k - 2c_{k+1} - 2\sum_{j=0}^{k-1} c_jc_{k-j}. \tag{713}$$

Thus the desired scalar closure is not an approximation at the level of the limiting vector field: it is an exact invariant manifold of the full matrix hierarchy.

C.5.5 Scalarization of the random initial condition

Under the rotated Gaussian initialization,

$$\boldsymbol{A}_N(0) = \sqrt{\frac{\rho_N}{n}}\,\boldsymbol{Z}_{A,N}, \qquad \boldsymbol{B}_N(0) = \sqrt{\frac{\rho_N}{n}}\,\boldsymbol{Z}_{B,N}, \tag{714}$$

where $\boldsymbol{Z}_{A,N} \in \mathbb{R}^{n\times m}$ and $\boldsymbol{Z}_{B,N} \in \mathbb{R}^{n\times(p-m)}$ have independent standard Gaussian entries and are independent.

Consequently,

$$\boldsymbol{M}_N(0) = \frac{\rho_N}{n}\boldsymbol{Z}_{B,N}\boldsymbol{Z}_{B,N}^\top. \tag{715}$$

For $k \ge 0$, set

$$s_{k,N} := \frac{\rho_N}{n}\operatorname{Tr}\big(\boldsymbol{M}_N(0)^k\big). \tag{716}$$

Lemma C.35 (Conditional scalarization)**.** *For each fixed $k \ge 0$,*

$$\|\boldsymbol{Q}_{k,N}(0) - s_{k,N}\boldsymbol{I}_m\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{717}$$

Proof. Condition on $\boldsymbol{M}_N(0)$ and abbreviate it by $\boldsymbol{M}$. From equation 714,

$$\boldsymbol{Q}_{k,N}(0) = \frac{\rho_N}{n} \boldsymbol{Z}_{A,N}^\top \boldsymbol{M}^k \boldsymbol{Z}_{A,N}. \tag{718}$$

Let

$$\boldsymbol{H}_{k,N} := \frac{1}{n} \boldsymbol{Z}_{A,N}^\top \boldsymbol{M}^k \boldsymbol{Z}_{A,N} - \frac{1}{n} \operatorname{Tr}(\boldsymbol{M}^k) \boldsymbol{I}_m. \tag{719}$$

For a deterministic unit vector $\boldsymbol{v} \in \mathbb{R}^m$, $\boldsymbol{Z}_{A,N}\boldsymbol{v}$ is a standard Gaussian vector $\boldsymbol{g} \in \mathbb{R}^n$, so

$$\boldsymbol{v}^\top \boldsymbol{H}_{k,N} \boldsymbol{v} = \frac{1}{n} \left(\boldsymbol{g}^\top \boldsymbol{M}^k \boldsymbol{g} - \operatorname{Tr}(\boldsymbol{M}^k) \right). \tag{720}$$

Fix $L < \infty$ and work on the event

$$\|\boldsymbol{M}\|_{\mathrm{op}} \le L. \tag{721}$$

The standard Gaussian quadratic-form estimate implies that, for every $\varepsilon > 0$,

$$\mathbb{P}\left(|\boldsymbol{v}^\top \boldsymbol{H}_{k,N} \boldsymbol{v}| > \varepsilon \,\middle|\, \boldsymbol{M} \right) \le 2 \exp\left(-c_k n \min\{\varepsilon^2, \varepsilon\} \right), \tag{722}$$

where $c_k > 0$ depends on k and L, but not on n or m.

Let $\mathcal{N}_m$ be a $1/4$-net of the Euclidean unit sphere in $\mathbb{R}^m$ with

$$|\mathcal{N}_m| \le 9^m. \tag{723}$$

For every symmetric matrix $\boldsymbol{H}$,

$$\|\boldsymbol{H}\|_{\mathrm{op}} \le 2 \max_{\boldsymbol{v} \in \mathcal{N}_m} |\boldsymbol{v}^\top \boldsymbol{H} \boldsymbol{v}|. \tag{724}$$

A union bound using equation 722 therefore gives

$$\mathbb{P}\left(\|\boldsymbol{H}_{k,N}\|_{\mathrm{op}} > 2\varepsilon \mid \boldsymbol{M} \right) \le 2 \exp\left(m \log 9 - c_k n \min\{\varepsilon^2, \varepsilon\} \right). \tag{725}$$

By equation 660, the right-hand side tends to zero.

It remains to remove the conditioning event equation 721. Since equation 715 is Gaussian Wishart with

$$\frac{p-m}{n} = \psi_N \left(1 - \frac{1}{\phi_N} \right) \longrightarrow \psi, \tag{726}$$

the largest eigenvalue of $\boldsymbol{M}_N(0)$ is tight; in fact,

$$\|\boldsymbol{M}_N(0)\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} \rho(1 + \sqrt{\psi})^2. \tag{727}$$

Choosing $L > \rho(1 + \sqrt{\psi})^2$ completes the proof after multiplying $\boldsymbol{H}_{k,N}$ by ρ_N. □

The scalar coefficients in Lemma C.35 have deterministic limits. By Marchenko–Pastur moment convergence applied to equation 715,

$$\frac{1}{n} \operatorname{Tr}\left(\boldsymbol{M}_N(0)^k \right) \xrightarrow{\mathbb{P}} \rho^k \sum_{j=1}^{k} \operatorname{Nar}(k, j) \psi^j, \qquad k \ge 1. \tag{728}$$

Hence

$$s_{0,N} \longrightarrow \rho, \tag{729}$$

and

$$s_{k,N} \xrightarrow{\mathbb{P}} c_k(0) := \rho^{k+1} \sum_{j=1}^{k} \operatorname{Nar}(k, j) \psi^j, \qquad k \ge 1. \tag{730}$$

The convergence can be strengthened from each fixed k to the weighted Banach norm.

Lemma C.36 (Banach-norm scalarization at initialization)**.** *There exists $r > 0$ such that*

$$\|\mathbf{Q}_N(0) - \iota_m(\mathbf{c}(0))\|_r \xrightarrow{\mathbb{P}} 0, \tag{731}$$

where

$$\mathbf{c}(0) := \left(c_0(0), c_1(0), \ldots \right). \tag{732}$$

Proof. Choose $L > \rho(1+\sqrt{\psi})^2$. By standard Gaussian singular-value estimates, for some deterministic $C < \infty$,

$$\mathbb{P}\left(\|\boldsymbol{M}_N(0)\|_{\mathrm{op}} \le L, \quad \|\boldsymbol{A}_N(0)\|_{\mathrm{op}}^2 \le C\right) \longrightarrow 1. \tag{733}$$

On this event,

$$\|\boldsymbol{Q}_{k,N}(0)\|_{\mathrm{op}} \le CL^k. \tag{734}$$

The deterministic moments in equation 730 satisfy, after increasing L if necessary,

$$|c_k(0)| \le C'L^k. \tag{735}$$

Choose $r > 0$ with $rL < 1$. For every K,

$$\begin{aligned}\sum_{k>K} r^k \left\|\boldsymbol{Q}_{k,N}(0) - c_k(0)\boldsymbol{I}_m\right\|_{\mathrm{op}} \\ \le (C + C') \sum_{k>K} (rL)^k.\end{aligned} \tag{736}$$

The right-hand side tends to zero as $K \to \infty$, uniformly on the high-probability event equation 733. For each fixed K, the contribution of $0 \le k \le K$ tends to zero by Lemma C.35 and equation 730. First let $N \to \infty$, then $K \to \infty$. □

C.5.6 Propagation of scalarity

Let $\mathbf{c}(t)$ be the maximal solution in the scalar weighted space to equation 712–equation 713 with initial condition equation 729–equation 730. Local existence and uniqueness follow from the Banach-space Picard–Lindelof theorem and the local Lipschitz estimate equation 707.

We next prove both global existence on every finite interval and convergence of the matrix hierarchy to this solution.

Proposition C.37 (Finite-time propagation of isotropy)**.** *For the $r > 0$ chosen in Lemma C.36, and every $T < \infty$,*

$$\sup_{0\le t\le T} \left\|\mathbf{Q}_N(t) - \iota_m(\mathbf{c}(t))\right\|_r \xrightarrow{\mathbb{P}} 0. \tag{737}$$

In particular, the scalar Banach-space solution exists uniquely on every finite time interval.

Proof. On the event equation 733, Lemmas C.33 and C.34 imply

$$\sup_{t\ge 0} \|\mathbf{Q}_N(t)\|_r \le C_A \sum_{k=0}^{\infty} (rL)^k = \frac{C_A}{1-rL}, \tag{738}$$

where

$$C_A := \max\{1, C\}. \tag{739}$$

Let R be larger than twice the right-hand side of equation 738. Up to the first exit time of either $\mathbf{Q}_N$ or $\iota_m(\mathbf{c})$ from the radius-R ball, equation 707 and the integral form of the two ODEs give

$$\begin{aligned}&\left\|\mathbf{Q}_N(t) - \iota_m(\mathbf{c}(t))\right\|_r \\ &\quad\le \left\|\mathbf{Q}_N(0) - \iota_m(\mathbf{c}(0))\right\|_r + K_{r,R} \int_0^t \left\|\mathbf{Q}_N(s) - \iota_m(\mathbf{c}(s))\right\|_r ds.\end{aligned} \tag{740}$$

Gronwall's inequality yields

$$\sup_{0\le s\le t} \left\|\mathbf{Q}_N(s) - \iota_m(\mathbf{c}(s))\right\|_r \le e^{K_{r,R}t} \left\|\mathbf{Q}_N(0) - \iota_m(\mathbf{c}(0))\right\|_r. \tag{741}$$

By Lemma C.36, the right-hand side tends to zero in probability for every fixed t.

We now remove the stopping time carefully. Write

$$B_r := \frac{C_A}{1-rL}, \qquad d_R := R - B_r > 0. \tag{742}$$

Fix a deterministic $T < \infty$, and set

$$\varepsilon_T := \frac{d_R}{2} e^{-K_{r,R}T} > 0. \tag{743}$$

Let $\mathcal{E}_N(T)$ be the event on which equation 733 holds and

$$\|\mathbf{Q}_N(0) - \iota_m(\mathbf{c}(0))\|_r < \varepsilon_T. \tag{744}$$

By Lemma C.36,

$$\mathbb{P}(\mathcal{E}_N(T)) \longrightarrow 1. \tag{745}$$

In particular, $\mathcal{E}_N(T)$ is nonempty for all sufficiently large N. On this event the matrix hierarchy obeys

$$\sup_{t \geq 0} \|\mathbf{Q}_N(t)\|_r \leq B_r, \tag{746}$$

and, up to the first time at which the deterministic scalar trajectory $\iota_m(\mathbf{c}(t))$ reaches the radius-R boundary, equation 741 gives

$$\sup_{0 \leq s \leq t} \|\mathbf{Q}_N(s) - \iota_m(\mathbf{c}(s))\|_r < \frac{d_R}{2}, \qquad 0 \leq t \leq T. \tag{747}$$

If the scalar trajectory reached that boundary at some first time $\tau \leq T$, then on any realization in $\mathcal{E}_N(T)$,

$$\begin{aligned} R = \|\iota_m(\mathbf{c}(\tau))\|_r &\leq \|\mathbf{Q}_N(\tau)\|_r + \|\mathbf{Q}_N(\tau) - \iota_m(\mathbf{c}(\tau))\|_r \\ &< B_r + \frac{d_R}{2} < R, \end{aligned} \tag{748}$$

a contradiction. Since the scalar trajectory is deterministic, the existence of such a realization is enough to rule out the boundary hitting time altogether. Thus the scalar solution remains in the radius-R ball throughout $[0, T]$; by local Banach-space existence it cannot blow up before T. The stopped estimate is therefore valid on all of $[0, T]$, and Lemma C.36 now gives equation 737. Since T is arbitrary, the scalar solution exists uniquely on every finite interval. □

Taking any fixed coordinate k in equation 737 gives

$$\sup_{0 \leq t \leq T} \|\boldsymbol{Q}_{k,N}(t) - c_k(t)\boldsymbol{I}_m\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{749}$$

C.5.7 Identification of the scalar hierarchy

Set

$$a(t) := c_0(t). \tag{750}$$

We first verify that $a(t)$ is strictly positive at every finite time, so that the normalized moments are well defined.

At finite N,

$$0 \preceq \boldsymbol{Q}_{1,N}(t) \preceq \|\boldsymbol{M}_N(t)\|_{\mathrm{op}} \boldsymbol{Q}_{0,N}(t). \tag{751}$$

Passing to the scalar limit and using equation 695 gives, for some finite L,

$$0 \leq c_1(t) \leq L a(t). \tag{752}$$

From equation 712,

$$\dot{a} = 4a(1-a) - 2c_1 \geq a(4 - 4a - 2L). \tag{753}$$

Since $a(0) = \rho > 0$, scalar comparison implies

$$a(t) > 0 \qquad \text{for every finite } t. \tag{754}$$

We may therefore define

$$\mu_k(t) := \frac{c_k(t)}{a(t)}, \qquad \mu_0(t) \equiv 1. \tag{755}$$

Equation (712) immediately becomes equation 665. For $k \geq 1$, substitute $c_k = a\mu_k$ into equation 713:

$$\dot{a}\,\mu_k + a\dot{\mu}_k = 4a\mu_k - 4a^2\mu_k - 2a\mu_{k+1} - 2a^2 \sum_{j=0}^{k-1} \mu_j \mu_{k-j}. \tag{756}$$

Subtracting $\dot{a}\,\mu_k$ and using equation 665 yields precisely equation 666. The initial conditions equation 667 follow from equation 729 and equation 730.

This proves the hierarchy claimed in Theorem C.31. Moreover, if a solution is a moment sequence of probability measures whose supports are uniformly contained in $[0, L_*]$ on $[0, T]$, then

$$|\mu_k(t)| \leq L_*^k \qquad \text{for every } k \geq 0 \text{ and } 0 \leq t \leq T. \tag{757}$$

For every $r < L_*^{-1}$, the corresponding sequence belongs to the weighted Banach space used above. Thus the Banach-space uniqueness in Proposition C.37 implies uniqueness in the bounded-support class stated in Theorem C.31.

C.5.8 The active spectral measure and uniqueness

The moment interpretation can also be made rigorous. Diagonalize

$$\boldsymbol{M}_N(t) = \sum_{\alpha=1}^{n} \lambda_{\alpha,N}(t) \boldsymbol{v}_{\alpha,N}(t) \boldsymbol{v}_{\alpha,N}(t)^\top. \tag{758}$$

Let

$$a_N(t) := \frac{1}{m} \operatorname{Tr} \boldsymbol{Q}_{0,N}(t), \tag{759}$$

and, whenever $a_N(t) > 0$, define

$$\nu_{\phi_{N,t}} := \frac{1}{m a_N(t)} \sum_{\alpha=1}^{n} \|\boldsymbol{A}_N(t)^\top \boldsymbol{v}_{\alpha,N}(t)\|^2 \delta_{\lambda_{\alpha,N}(t)}. \tag{760}$$

This is a probability measure because

$$\sum_{\alpha=1}^{n} \|\boldsymbol{A}_N^\top \boldsymbol{v}_{\alpha,N}\|^2 = \operatorname{Tr}(\boldsymbol{A}_N^\top \boldsymbol{A}_N) = m a_N. \tag{761}$$

Its moments are exactly

$$\int \lambda^k \, d\nu_{\phi_{N,t}}(\lambda) = \frac{\operatorname{Tr} \boldsymbol{Q}_{k,N}(t)}{m a_N(t)}. \tag{762}$$

By Lemma C.33, these measures have uniformly bounded support with probability tending to one. By equation 749,

$$\frac{1}{m} \operatorname{Tr} \boldsymbol{Q}_{k,N}(t) \xrightarrow{\mathbb{P}} c_k(t), \tag{763}$$

uniformly on compact time intervals. More explicitly, from any subsequence of N one may extract a further subsequence along which the support bound and the moment convergences for all integer $k \geq 0$ hold almost surely, using a diagonal argument. Compactness of probability measures on the resulting fixed compact interval then gives a further weakly convergent subsequence. Bounded support allows passage of every polynomial moment to the weak limit, so every such limit ν_{ϕ_t} satisfies

$$\mu_k(t) = \int \lambda^k \, d\nu_{\phi_t}(\lambda), \qquad k \geq 0. \tag{764}$$

The compactly supported moment problem is determinate, hence all subsequential limits coincide. This identifies a unique active spectral measure ν_{ϕ_t}.

Equivalently, for $|z|$ sufficiently small one may introduce

$$H(t, z) := \sum_{k=0}^{\infty} \mu_k(t) z^k = \int \frac{1}{1 - z\lambda} \, d\nu_{\phi_t}(\lambda). \tag{765}$$

Summing equation 666 gives

$$\partial_t H = -2\frac{H-1}{z} - 2aH(H-1) + 2\mu_1 H, \qquad \mu_1 = \partial_z H(t,0), \tag{766}$$

where the apparent singularity at $z = 0$ is removable. The local analytic uniqueness of equation 766, equivalently the Banach-space uniqueness proved above, supplies a rigorous meaning to the infinite moment hierarchy.

C.5.9 Train loss for orthogonal data

For equation 675, the empirical loss satisfies the exact identity

$$\widehat{L}_N(t) = \frac{1}{2m}\|\boldsymbol{I}_m - \boldsymbol{Q}_{0,N}(t)\|_F^2 + \frac{1}{2m}\operatorname{Tr}\boldsymbol{Q}_{1,N}(t). \tag{767}$$

Indeed, the two terms are respectively the active reconstruction error and the leakage into the inactive block.

By equation 749,

$$\sup_{0\le t\le T}\left|\frac{1}{m}\operatorname{Tr}(\boldsymbol{I}_m - \boldsymbol{Q}_{0,N}(t))^2 - (1-a(t))^2\right| \xrightarrow{\mathbb{P}} 0, \tag{768}$$

and

$$\sup_{0\le t\le T}\left|\frac{1}{m}\operatorname{Tr}\boldsymbol{Q}_{1,N}(t) - a(t)\mu_1(t)\right| \xrightarrow{\mathbb{P}} 0. \tag{769}$$

Substitution in equation 767 proves equation 670 for orthogonal data.

C.5.10 Population loss and its first correction for orthogonal data

The population loss admits the exact block decomposition

$$\begin{aligned} L_N(t) = \frac{1}{2p}\|\boldsymbol{I}_m - \boldsymbol{Q}_{0,N}(t)\|_F^2 + \frac{1}{p}\operatorname{Tr}\boldsymbol{Q}_{1,N}(t) \\ + \frac{1}{2p}\|\boldsymbol{I}_{p-m} - \boldsymbol{B}_N(t)^\top \boldsymbol{B}_N(t)\|_F^2. \end{aligned} \tag{770}$$

Define

$$F_{\perp,N}(t) := \frac{1}{2p}\|\boldsymbol{I}_{p-m} - \boldsymbol{B}_N(t)^\top \boldsymbol{B}_N(t)\|_F^2 \tag{771}$$

and

$$G_N(t) := \frac{1}{2m}\|\boldsymbol{I}_m - \boldsymbol{Q}_{0,N}(t)\|_F^2 + \frac{1}{m}\operatorname{Tr}\boldsymbol{Q}_{1,N}(t). \tag{772}$$

Then

$$L_N(t) = F_{\perp,N}(t) + \frac{1}{\phi_N}G_N(t). \tag{773}$$

Differentiating equation 771 and using equation 681, one obtains the exact identity

$$\dot{F}_{\perp,N}(t) = \frac{2}{p}\left[\operatorname{Tr}\boldsymbol{Q}_{1,N}(t) - \operatorname{Tr}\boldsymbol{Q}_{2,N}(t)\right]. \tag{774}$$

Indeed, if

$$\boldsymbol{C}_N := \boldsymbol{B}_N^\top \boldsymbol{B}_N, \tag{775}$$

then

$$\dot{\boldsymbol{C}}_N = -2\boldsymbol{B}_N^\top \boldsymbol{A}_N \boldsymbol{A}_N^\top \boldsymbol{B}_N, \tag{776}$$

and cyclicity of the trace gives

$$\begin{aligned} \dot{F}_{\perp,N} &= \frac{2}{p}\operatorname{Tr}\left[(\boldsymbol{I}_{p-m} - \boldsymbol{C}_N)\boldsymbol{B}_N^\top \boldsymbol{A}_N \boldsymbol{A}_N^\top \boldsymbol{B}_N\right] \\ &= \frac{2}{p}\left[\operatorname{Tr}(\boldsymbol{A}_N^\top \boldsymbol{M}_N \boldsymbol{A}_N) - \operatorname{Tr}(\boldsymbol{A}_N^\top \boldsymbol{M}_N^2 \boldsymbol{A}_N)\right]. \end{aligned} \tag{777}$$

Integrating equation 774 and using $m/p = 1/\phi_N$ gives the exact identity

$$\phi_N\big(L_N(t) - L_N(0)\big) = G_N(t) - G_N(0) + 2\int_0^t \left[\frac{1}{m}\operatorname{Tr}\boldsymbol{Q}_{1,N}(s) - \frac{1}{m}\operatorname{Tr}\boldsymbol{Q}_{2,N}(s)\right] ds. \tag{778}$$

By equation 749,

$$G_N(t) \xrightarrow{\mathbb{P}} G(t) := \frac{1}{2}(1-a(t))^2 + a(t)\mu_1(t), \tag{779}$$

uniformly for $t \in [0,T]$, and

$$\frac{1}{m}\operatorname{Tr}\boldsymbol{Q}_{2,N}(t) \xrightarrow{\mathbb{P}} a(t)\mu_2(t). \tag{780}$$

At time zero,

$$G(0) = \frac{1}{2}(1-\rho)^2 + \psi\rho^2. \tag{781}$$

Passing to the limit in equation 778 proves equation 673–equation 674 in the orthogonal-data model.

The leading population limit follows from the fact that equation 778 implies

$$\sup_{0\le t\le T} |L_N(t) - L_N(0)| = O_{\mathbb{P}}(\phi_N^{-1}). \tag{782}$$

At initialization,

$$L_N(0) = \frac{1}{2} - \frac{1}{p}\operatorname{Tr}(\boldsymbol{U}_N(0)^\top \boldsymbol{U}_N(0)) + \frac{1}{2p}\operatorname{Tr}\left[(\boldsymbol{U}_N(0)^\top \boldsymbol{U}_N(0))^2\right]. \tag{783}$$

Marchenko–Pastur convergence for the Gaussian initialization gives

$$\frac{1}{p}\operatorname{Tr}(\boldsymbol{U}_N(0)^\top \boldsymbol{U}_N(0)) \xrightarrow{\mathbb{P}} \rho, \tag{784}$$

and

$$\frac{1}{p}\operatorname{Tr}\left[(\boldsymbol{U}_N(0)^\top \boldsymbol{U}_N(0))^2\right] \xrightarrow{\mathbb{P}} \rho^2(1+\psi). \tag{785}$$

Equations (782) to (785) prove equation 672 for orthogonal data.

C.5.11 Transfer from orthogonalized to general near-isometric data

We now return to an arbitrary training matrix satisfying equation 48.

On the event that $\boldsymbol{X}_N$ has full column rank, write its thin polar decomposition as

$$\boldsymbol{X}_N = \sqrt{p}\,\boldsymbol{Q}_N \boldsymbol{\Gamma}_N^{1/2}, \qquad \boldsymbol{Q}_N^\top \boldsymbol{Q}_N = \boldsymbol{I}_m. \tag{786}$$

Since equation 48 implies $\lambda_{\min}(\boldsymbol{\Gamma}_N) \to 1$ in probability, this event has probability tending to one.

Extend $\boldsymbol{Q}_N$ to an orthogonal matrix

$$\boldsymbol{R}_N = \begin{bmatrix}\boldsymbol{Q}_N & \boldsymbol{Q}_N^\perp\end{bmatrix} \in O(p), \tag{787}$$

and replace $\boldsymbol{U}_N$ by $\boldsymbol{U}_N\boldsymbol{R}_N$. Conditional on $\boldsymbol{X}_N$, the law of the rotated initialization is still the one specified in Eq. (36). Moreover, because this conditional law does not depend on $\boldsymbol{X}_N$, the rotated initialization remains independent of $\boldsymbol{\Gamma}_N$.

In the rotated coordinates,

$$\frac{1}{\phi_N}\boldsymbol{S}_N = \begin{pmatrix}\boldsymbol{\Gamma}_N & 0\\ 0 & 0\end{pmatrix}. \tag{788}$$

Writing again $\boldsymbol{U}_N = [\boldsymbol{A}_N\ \boldsymbol{B}_N]$, direct block multiplication in the common flow equation 49 with $\gamma = \phi_N^{-1}$ gives

$$\dot{\boldsymbol{A}}_N = \boldsymbol{A}_N\left[\boldsymbol{\Gamma}_N(\boldsymbol{I}_m - \boldsymbol{Q}_{0,N}) + (\boldsymbol{I}_m - \boldsymbol{Q}_{0,N})\boldsymbol{\Gamma}_N\right] - \boldsymbol{M}_N\boldsymbol{A}_N\boldsymbol{\Gamma}_N, \tag{789}$$

and

$$\dot{\boldsymbol{B}}_N = -\boldsymbol{A}_N \boldsymbol{\Gamma}_N \boldsymbol{A}_N^\top \boldsymbol{B}_N. \tag{790}$$

For $\boldsymbol{\Gamma}_N = \boldsymbol{I}_m$, these equations reduce exactly to equation 680–equation 681.

We now couple the $\boldsymbol{\Gamma}_N$-dynamics and the orthogonal-data dynamics using the same initial matrices $\boldsymbol{A}_N(0), \boldsymbol{B}_N(0)$. Denote the two solutions by

$$(\boldsymbol{A}_N^\Gamma, \boldsymbol{B}_N^\Gamma) \qquad \text{and} \qquad (\boldsymbol{A}_N^I, \boldsymbol{B}_N^I). \tag{791}$$

Lemma C.38 (Finite-time stability under Gram perturbations)**.** *For every fixed $T < \infty$, on events whose probability tends to one,*

$$\sup_{0\le t\le T} \left(\|\boldsymbol{A}_N^\Gamma(t) - \boldsymbol{A}_N^I(t)\|_{\mathrm{op}} + \|\boldsymbol{B}_N^\Gamma(t) - \boldsymbol{B}_N^I(t)\|_{\mathrm{op}} \right) \le C_T \|\boldsymbol{\Gamma}_N - \boldsymbol{I}_m\|_{\mathrm{op}}, \tag{792}$$

where C_T is deterministic and independent of N.

Proof. By the Gaussian singular-value bounds used above, with probability tending to one,

$$\|\boldsymbol{A}_N(0)\|_{\mathrm{op}} + \|\boldsymbol{B}_N(0)\|_{\mathrm{op}} \le K_0 \tag{793}$$

for a deterministic $K_0 < \infty$. The orthogonal-data trajectory then satisfies a deterministic bound

$$\sup_{0\le t\le T} \left(\|\boldsymbol{A}_N^I(t)\|_{\mathrm{op}} + \|\boldsymbol{B}_N^I(t)\|_{\mathrm{op}} \right) \le K_T \tag{794}$$

by Lemmas C.33 and C.34.

On the set

$$\|\boldsymbol{A}\|_{\mathrm{op}} + \|\boldsymbol{B}\|_{\mathrm{op}} \le 2K_T, \qquad \|\boldsymbol{\Gamma}\|_{\mathrm{op}} \le 2, \tag{795}$$

the polynomial vector field in equation 789–equation 790 is uniformly Lipschitz in $(\boldsymbol{A}, \boldsymbol{B})$ and Lipschitz in $\boldsymbol{\Gamma}$. Write $\mathcal{F}_A$ and $\mathcal{F}_B$ for its two block components. Hence there is a deterministic $C = C(T, K_T)$ such that

$$\begin{aligned} &\|\mathcal{F}_A(\boldsymbol{A}, \boldsymbol{B}; \boldsymbol{\Gamma}) - \mathcal{F}_A(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}}; \boldsymbol{I}_m)\|_{\mathrm{op}} \\ &\quad + \|\mathcal{F}_B(\boldsymbol{A}, \boldsymbol{B}; \boldsymbol{\Gamma}) - \mathcal{F}_B(\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{B}}; \boldsymbol{I}_m)\|_{\mathrm{op}} \\ &\quad \le C \left(\|\boldsymbol{A} - \widetilde{\boldsymbol{A}}\|_{\mathrm{op}} + \|\boldsymbol{B} - \widetilde{\boldsymbol{B}}\|_{\mathrm{op}} + \|\boldsymbol{\Gamma} - \boldsymbol{I}_m\|_{\mathrm{op}} \right). \end{aligned} \tag{796}$$

Applying Lemma C.3 up to the first exit of the $\boldsymbol{\Gamma}_N$-trajectory from the radius-$2K_T$ ball yields

$$\sup_{0\le t\le T} \left(\|\boldsymbol{A}_N^\Gamma(t) - \boldsymbol{A}_N^I(t)\|_{\mathrm{op}} + \|\boldsymbol{B}_N^\Gamma(t) - \boldsymbol{B}_N^I(t)\|_{\mathrm{op}} \right) \le C_T \|\boldsymbol{\Gamma}_N - \boldsymbol{I}_m\|_{\mathrm{op}}. \tag{797}$$

For N sufficiently large, equation 48 makes the right-hand side smaller than K_T, so the presumed first exit cannot occur. This proves equation 792. □

Since

$$\boldsymbol{M}_N^\Gamma = \boldsymbol{B}_N^\Gamma (\boldsymbol{B}_N^\Gamma)^\top, \qquad \boldsymbol{M}_N^I = \boldsymbol{B}_N^I (\boldsymbol{B}_N^I)^\top, \tag{798}$$

Lemma C.38 implies

$$\sup_{0\le t\le T} \|\boldsymbol{M}_N^\Gamma(t) - \boldsymbol{M}_N^I(t)\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{799}$$

For each fixed k,

$$(\boldsymbol{M}_N^\Gamma)^k - (\boldsymbol{M}_N^I)^k = \sum_{j=0}^{k-1} (\boldsymbol{M}_N^\Gamma)^j (\boldsymbol{M}_N^\Gamma - \boldsymbol{M}_N^I)(\boldsymbol{M}_N^I)^{k-1-j}, \tag{800}$$

so the uniform operator-norm bounds and equation 792 imply

$$\sup_{0\le t\le T} \left\| (\boldsymbol{A}_N^\Gamma)^\top (\boldsymbol{M}_N^\Gamma)^k \boldsymbol{A}_N^\Gamma - (\boldsymbol{A}_N^I)^\top (\boldsymbol{M}_N^I)^k \boldsymbol{A}_N^I \right\|_{\mathrm{op}} \xrightarrow{\mathbb{P}} 0. \tag{801}$$

Combining equation 801 with equation 749 proves equation 669 for the original, non-Gaussian training matrix.

C.5.12 TRAIN LOSS FOR GENERAL NEAR-ISOMETRIC DATA

In the rotated coordinates equation 786, the empirical loss has the exact form

$$\widehat{L}_N(t) = \frac{1}{2m} \operatorname{Tr}\left\{\boldsymbol{\Gamma}_N \left[(\boldsymbol{I}_m - \boldsymbol{Q}_{0,N}(t))^2 + \boldsymbol{Q}_{1,N}(t)\right]\right\}. \tag{802}$$

To verify this, note that the upper-left block of $\boldsymbol{E}_N^2$ is

$$(\boldsymbol{E}_N^2)_{11} = (\boldsymbol{I}_m - \boldsymbol{Q}_{0,N})^2 + \boldsymbol{Q}_{1,N}. \tag{803}$$

Equation (802) then follows directly from equation 786 and equation 38.

For every matrix $\boldsymbol{R} \in \mathbb{R}^{m\times m}$,

$$\left|\frac{1}{m} \operatorname{Tr}\big[(\boldsymbol{\Gamma}_N - \boldsymbol{I}_m)\boldsymbol{R}\big]\right| \le \|\boldsymbol{\Gamma}_N - \boldsymbol{I}_m\|_{\mathrm{op}} \|\boldsymbol{R}\|_{\mathrm{op}}. \tag{804}$$

Therefore equation 48, equation 669, and the uniform operator-norm bounds imply equation 670.

C.5.13 POPULATION CORRECTION FOR GENERAL NEAR-ISOMETRIC DATA

The population decomposition equation 770 is independent of $\boldsymbol{\Gamma}_N$, but the inactive block now evolves according to equation 790. Repeating the differentiation leading to equation 774 gives the exact identity

$$\dot{F}_{\perp,N}(t) = \frac{2}{p} \operatorname{Tr}\left[\boldsymbol{\Gamma}_N \big(\boldsymbol{Q}_{1,N}(t) - \boldsymbol{Q}_{2,N}(t)\big)\right]. \tag{805}$$

Consequently,

$$\begin{aligned}\phi_N\big(L_N(t) - L_N(0)\big) = G_N(t) - G_N(0) \\ + 2\int_0^t \frac{1}{m} \operatorname{Tr}\left[\boldsymbol{\Gamma}_N \big(\boldsymbol{Q}_{1,N}(s) - \boldsymbol{Q}_{2,N}(s)\big)\right] ds.\end{aligned} \tag{806}$$

For $k = 1, 2$,

$$\begin{aligned}&\left|\frac{1}{m} \operatorname{Tr}(\boldsymbol{\Gamma}_N \boldsymbol{Q}_{k,N}) - a\mu_k\right| \\ &\quad \le \|\boldsymbol{\Gamma}_N\|_{\mathrm{op}} \|\boldsymbol{Q}_{k,N} - a\mu_k \boldsymbol{I}_m\|_{\mathrm{op}} + |a\mu_k| \left|\frac{1}{m} \operatorname{Tr} \boldsymbol{\Gamma}_N - 1\right|.\end{aligned} \tag{807}$$

Both terms tend to zero uniformly on $[0, T]$ in probability, by equation 48 and equation 669. Therefore equation 806 converges uniformly to equation 674, proving equation 673.

This also explains why an $O(\phi_N^{-1/2})$ operator-norm error in $\boldsymbol{\Gamma}_N$ does not contaminate the $O(\phi_N^{-1})$ population correction. After multiplication by ϕ_N, the Gram perturbation enters only through an m-normalized trace. In particular,

$$\left|\frac{1}{m} \operatorname{Tr}\left[(\boldsymbol{\Gamma}_N - \boldsymbol{I}_m)\boldsymbol{Q}_{k,N}\right]\right| \le \|\boldsymbol{\Gamma}_N - \boldsymbol{I}_m\|_{\mathrm{op}} \|\boldsymbol{Q}_{k,N}\|_{\mathrm{op}} = o_{\mathbb{P}}(1). \tag{808}$$

No second-order expansion of the Gram matrix is required.

Finally, equation 673 implies

$$\sup_{0\le t\le T} |L_N(t) - L_N(0)| = O_{\mathbb{P}}(\phi_N^{-1}), \tag{809}$$

and the initialization limit equation 784–equation 785 is independent of the training data. Thus equation 672 also holds for general data satisfying equation 48. This completes the proof of Theorem C.31.

C.5.14 Instantiation for isotropic Gaussian training data

We record explicitly how Theorem C.31 specializes to the Gaussian setting considered in the main text. Suppose that

$$\boldsymbol{X}_N(i,j) \overset{\text{iid}}{\sim} \mathcal{N}(0,1). \tag{810}$$

The standard Gaussian singular-value inequality gives, for every $t \geq 0$,

$$\mathbb{P}\left(s_{\max}(\boldsymbol{X}_N) > \sqrt{p} + \sqrt{m} + t\right) \leq e^{-t^2/2}, \tag{811}$$

and

$$\mathbb{P}\left(s_{\min}(\boldsymbol{X}_N) < \sqrt{p} - \sqrt{m} - t\right) \leq e^{-t^2/2}. \tag{812}$$

Since $m/p = \phi_N^{-1} \to 0$, these inequalities imply

$$\left\| \frac{1}{p} \boldsymbol{X}_N^\top \boldsymbol{X}_N - \boldsymbol{I}_m \right\|_{\text{op}} = O_{\mathbb{P}}\left(\sqrt{\frac{m}{p}} + \frac{m}{p}\right) = O_{\mathbb{P}}(\phi_N^{-1/2}). \tag{813}$$

Thus the sole data assumption equation 48 is satisfied.

Consequently, for isotropic Gaussian training data the rigorous limiting train loss is

$$\widehat{\mathcal{L}}(t) = \frac{1}{2}(1 - a(t))^2 + \frac{1}{2} a(t) \mu_1(t), \tag{814}$$

with a, μ_k determined by equation 665–equation 667, and the centered population correction is

$$\phi_N \big(L_N(t) - L_N(0)\big) \xrightarrow{\mathbb{P}} C(t), \tag{815}$$

uniformly on every compact time interval.

For comparison with the averaged notation used in the main text, the Gaussian initialization admits the exact finite-dimensional identity

$$\mathbb{E} L_N(0) = \frac{1}{2}\left[1 - 2\rho_N + \left(1 + \psi_N + \frac{1}{n}\right)\rho_N^2\right]. \tag{816}$$

Indeed,

$$\mathbb{E}\left[\frac{1}{p} \operatorname{Tr}(\boldsymbol{U}_N(0)^\top \boldsymbol{U}_N(0))\right] = \rho_N, \tag{817}$$

while the Gaussian fourth-moment identity gives

$$\mathbb{E}\left[\frac{1}{p} \operatorname{Tr}\big(\boldsymbol{U}_N(0)^\top \boldsymbol{U}_N(0)\big)^2\right] = \rho_N^2 \left(1 + \psi_N + \frac{1}{n}\right). \tag{818}$$

Moreover,

$$\frac{1/n}{1/\phi_N} = \frac{\psi_N}{m} \longrightarrow 0. \tag{819}$$

We now justify the expectation upgrade, including the scaled population correction. Define

$$R_N := \max\left\{1, \|\boldsymbol{U}_N(0)\|_{\text{op}}^2\right\}. \tag{820}$$

Lemma C.2, applied to the common flow equation 49 with $\gamma = \phi_N^{-1}$, gives the pathwise bound

$$\sup_{t \geq 0} \|\boldsymbol{U}_N(t)\|_{\text{op}}^2 \leq R_N. \tag{821}$$

After the rotation used above, this implies, for $k = 0, 1, 2$,

$$\sup_{t \geq 0} \|\boldsymbol{Q}_{k,N}(t)\|_{\text{op}} \leq R_N^{k+1}. \tag{822}$$

Indeed, $\|\boldsymbol{A}_N\|_{\text{op}} \leq \|\boldsymbol{U}_N\|_{\text{op}}$ and $\|\boldsymbol{M}_N\|_{\text{op}} = \|\boldsymbol{B}_N\|_{\text{op}}^2 \leq R_N$. Consequently, the quantity G_N in equation 772 satisfies

$$\sup_{t \geq 0} |G_N(t)| \leq \frac{1}{2}(1 + R_N)^2 + R_N^2. \tag{823}$$

Using the exact identity equation 806, we obtain for every fixed $T < \infty$

$$\begin{aligned}\sup_{0\le t\le T}\left|\phi_N\big(L_N(t)-L_N(0)\big)\right| &\le (1+R_N)^2+2R_N^2\\ &\quad+2T\|\mathbf{\Gamma}_N\|_{\mathrm{op}}\left(R_N^2+R_N^3\right)\\ &\le C_T\left(1+\|\mathbf{\Gamma}_N\|_{\mathrm{op}}\right)\left(1+R_N^3\right).\end{aligned} \tag{824}$$

Similarly, equation 802 gives

$$\sup_{0\le t\le T}\widehat{L}_N(t)\le\frac{1}{2}\|\mathbf{\Gamma}_N\|_{\mathrm{op}}\left[(1+R_N)^2+R_N^2\right]. \tag{825}$$

For Gaussian $\boldsymbol{U}_N(0)$ and $\boldsymbol{X}_N$, the singular-value tail bounds used above imply, for every fixed $q<\infty$,

$$\sup_N\mathbb{E}R_N^q<\infty,\qquad \sup_N\mathbb{E}\|\mathbf{\Gamma}_N\|_{\mathrm{op}}^q<\infty. \tag{826}$$

Hence the right-hand sides of equation 824 and equation 825 have a uniformly bounded moment of some order strictly larger than one. They are therefore uniformly integrable. Combining this fact with the uniform convergence in probability in equation 670 and equation 673 yields

$$\mathbb{E}\left[\sup_{0\le t\le T}\left|\widehat{L}_N(t)-\frac{1}{2}(1-a(t))^2-\frac{1}{2}a(t)\mu_1(t)\right|\right]\longrightarrow 0 \tag{827}$$

and

$$\mathbb{E}\left[\sup_{0\le t\le T}\left|\phi_N\big(L_N(t)-L_N(0)\big)-C(t)\right|\right]\longrightarrow 0. \tag{828}$$

In particular,

$$\mathbb{E}\widehat{L}_N(t)\longrightarrow\frac{1}{2}(1-a(t))^2+\frac{1}{2}a(t)\mu_1(t), \tag{829}$$

uniformly on compact time intervals, and the exact initialization identity equation 816 together with equation 828 gives

$$\begin{aligned}\mathbb{E}L_N(t)&=\frac{1}{2}\left[1-2\rho_N+\left(1+\psi_N+\frac{1}{n}\right)\rho_N^2\right]\\ &\quad+\frac{1}{\phi_N}C(t)+o(\phi_N^{-1}),\end{aligned} \tag{830}$$

uniformly for $0\le t\le T$. If $\rho_N-\rho=o(\phi_N^{-1})$ and $\psi_N-\psi=o(\phi_N^{-1})$, then equation 830 can equivalently be written as

$$\begin{aligned}\mathbb{E}L_N(t)&=\frac{1}{2}\left[1-2\rho+(1+\psi)\rho^2\right]\\ &\quad+\frac{1}{\phi_N}\left\{\frac{1}{2}(1-a(t))^2+a(t)\mu_1(t)-\frac{1}{2}\left[1-2\rho+(1+2\psi)\rho^2\right]\right.\\ &\quad\left.+2\int_0^t a(s)\big[\mu_1(s)-\mu_2(s)\big]ds\right\}+o(\phi_N^{-1}),\end{aligned} \tag{831}$$

which is the $1/\phi$ formula stated in the main text, now with a rigorous finite-time justification.

Universality with respect to the training distribution. The proof shows that Gaussianity of the training samples is not used anywhere except as one convenient sufficient condition for equation 48. Therefore all conclusions of Theorem C.31 hold for any independent training-sample model, or even for deterministic data matrices, for which

$$\left\|\frac{1}{p}\boldsymbol{X}_N^\top\boldsymbol{X}_N-\boldsymbol{I}_m\right\|_{\mathrm{op}}\longrightarrow 0 \tag{832}$$

in the relevant deterministic or probabilistic sense.

For example, this includes broad families of isotropic non-Gaussian distributions with sufficiently strong norm and inner-product concentration. Isotropy by itself is not sufficient. If

$$\boldsymbol{x} = \sqrt{\tau}\,\sqrt{p}\,\boldsymbol{u}, \qquad \boldsymbol{u} \sim \mathrm{Unif}(\mathbb{S}^{p-1}), \qquad \mathbb{E}\tau = 1, \tag{833}$$

and τ has nonvanishing fluctuations as $p \to \infty$, then $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \boldsymbol{I}_p$, but

$$\frac{\|\boldsymbol{x}\|^2}{p} = \tau \tag{834}$$

does not concentrate. In the small-data regime the resulting Gram matrix is approximately diagonal with entries $\tau_1, \dots, \tau_m$, rather than approximately $\boldsymbol{I}_m$. In that case different examples evolve at different effective rates, and the scalar hierarchy equation 665–equation 666 need not remain valid.

C.5.15 Remarks on anisotropic data

The training-loss part of the argument extends further than the isotropic population setting. The empirical dynamics depends on the data, after rotation to their span, only through the normalized Gram matrix $\boldsymbol{\Gamma}_N$. Thus if an anisotropic data model is normalized in such a way that it still satisfies

$$\|\boldsymbol{\Gamma}_N - \boldsymbol{I}_m\|_{\mathrm{op}} \longrightarrow 0, \tag{835}$$

then the same limiting hierarchy and the same train-loss formula continue to hold.

The population loss is different. If the test covariance is a non-scalar matrix $\boldsymbol{\Sigma}_N$, then

$$L_{\Sigma,N}(U) = \frac{1}{2p}\,\mathrm{Tr}\left[\boldsymbol{\Sigma}_N(\boldsymbol{I}_p - \boldsymbol{U}^\top\boldsymbol{U})^2\right]. \tag{836}$$

After rotating to the training span, $\boldsymbol{\Sigma}_N$ acquires the block form

$$\boldsymbol{R}_N^\top \boldsymbol{\Sigma}_N \boldsymbol{R}_N = \begin{pmatrix} \boldsymbol{\Sigma}_{\parallel\parallel,N} & \boldsymbol{\Sigma}_{\parallel\perp,N} \\ \boldsymbol{\Sigma}_{\perp\parallel,N} & \boldsymbol{\Sigma}_{\perp\perp,N} \end{pmatrix}. \tag{837}$$

The population correction then depends not only on $\boldsymbol{A}_N^\top \boldsymbol{M}_N^k \boldsymbol{A}_N$, but also on weighted observables coupling these blocks to $\boldsymbol{A}_N$ and $\boldsymbol{B}_N$. Consequently, the train hierarchy can remain universal while the population correction changes.

A rigorous anisotropic extension should therefore augment the present matrix hierarchy by weighted order parameters describing the joint asymptotic geometry of the training span and $\boldsymbol{\Sigma}_N$. If, for example, the compressed covariance becomes asymptotically scalar,

$$\left\|\boldsymbol{Q}_N^\top \boldsymbol{\Sigma}_N \boldsymbol{Q}_N - \gamma \boldsymbol{I}_m\right\|_{\mathrm{op}} \longrightarrow 0, \tag{838}$$

and the corresponding active–inactive weighted traces admit deterministic limits, one expects a finite enlargement of the hierarchy and an anisotropy-dependent analogue of equation 674. For a fully general deterministic spectral profile $\boldsymbol{\Sigma}_N$, however, scalar closure should not be expected; a matrix-valued or measure-valued weighted hierarchy is the natural object.

Finally, all results in this section are finite-time statements: the limit $N \to \infty$ is uniform on every fixed interval $[0, T]$. The subsequent limit $t \to \infty$, and in particular any interchange of the large-dimension and large-time limits, requires an additional analysis of the deterministic moment dynamics and is not asserted here.

D Experimental details

This appendix describes the finite-dimensional experiments used to test the limiting dynamics derived in Section 4. We organize the discussion by learning regime. Several data constructions recur across regimes and are described first.

Unless stated otherwise, the model is initialized according to the same isotropic Gaussian law as in the theory,

$$U_{\alpha i}(0) \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, \sigma^2),$$

with the regime-specific rescaled initialization parameter held fixed. Whenever several independent initializations are used, curves show the empirical mean and error bars show one standard deviation unless specified otherwise.

D.1 Recurring experimental setups

Synthetic Gaussian data. For synthetic experiments we sample

$$X \in \mathbb{R}^{p\times m}, \qquad X_{i\mu} \overset{\text{iid}}{\sim} \mathcal{N}(0,1).$$

Thus the population covariance is exactly isotropic,

$$\mathbb{E}[xx^\top] = I_p,$$

and the empirical covariance

$$S = \frac{1}{m} XX^\top$$

has the Marchenko–Pastur spectral limit at fixed $\phi = p/m$. When $m < p$, its m nonzero eigenvalues are obtained numerically from the smaller matrix $X^\top X/m$.

The narrow-latent large-time experiment integrates an exact reduced finite-dimensional gradient-flow system. In that experiment a single Gaussian training matrix is fixed and the weight initialization is resampled across seeds.

Globally whitened CIFAR-10. CIFAR-10 images are treated as unlabeled vectors. Each 32×32 RGB image is flattened, giving the original dimension

$$p_0 = 32 \cdot 32 \cdot 3 = 3072,$$

and pixel intensities are divided by 255.

Whenever we refer to globally whitened CIFAR-10, the whitening map is fit using the entire set of

$$m_{\text{full}} = 50\,000$$

training images, before any training subset is selected. Let

$$\bar{x} = \frac{1}{m_{\text{full}}} \sum_{\mu=1}^{m_{\text{full}}} x_\mu$$

and

$$C_{\text{full}} = \frac{1}{m_{\text{full}}} \sum_{\mu=1}^{m_{\text{full}}} (x_\mu - \bar{x})(x_\mu - \bar{x})^\top = V\Lambda V^\top.$$

The whitened coordinates are

$$z = \Lambda^{-1/2} V^\top (x - \bar{x}).$$

The same training mean and the same whitening transformation are applied to held-out test images; no test-set statistics enter the preprocessing map.

If only a subset of the training data is used for optimization, that subset is selected after the global whitening map has been fixed. Thus the covariance of the complete whitened training split is approximately the identity, while the covariance S of a finite subset remains nontrivial. This construction removes the large anisotropy of the original CIFAR-10 coordinates while retaining a nondegenerate finite-sample spectrum.

For the mean-field experiment we retain only the $p = 512$ largest-variance PCA directions before whitening, for computational efficiency. The narrow-latent and large-data experiments use the full image dimension unless stated otherwise.

D.2 Large-data experiments

We use all $50\,000$ CIFAR-10 training images and all $10\,000$ test images, with the global whitening construction described above and $p = 3072$. Up to numerical precision,

$$S_{\text{tr}} = I_p.$$

We fix

$$\rho = n\sigma^2 = 1$$

and use

$$n \in \{2p, p, p/2\} = \{6144, 3072, 1536\}, \qquad \psi = \frac{p}{n} \in \left\{\frac{1}{2}, 1, 2\right\}.$$

The same Gaussian seed is used for the three widths.

With $\eta = p$, the exact finite-dimensional gradient flow reduces to

$$\dot{U} = 2U(I_p - U^\top U).$$

Writing $A = U^\top U$ gives the exact Gram dynamics

$$\dot{A} = 4A(I_p - A).$$

Therefore, if

$$A(0) = V \operatorname{diag}(a_1(0), \ldots, a_p(0))V^\top,$$

the eigenvectors remain fixed and

$$a_j(t) = \frac{a_j(0)}{a_j(0) + (1 - a_j(0))e^{-4t}}.$$

We use this formula directly, so the finite-width CIFAR-10 curves contain no time-discretization error. When $n < p$, the $p - n$ exact zero eigenvalues are retained analytically.

The finite-width training loss is

$$\widehat{L}_{\mathrm{tr}}(t) = \frac{1}{2p} \sum_{j=1}^{p} (1 - a_j(t))^2.$$

If

$$T = \frac{1}{m_{\mathrm{te}}} X_{\mathrm{te}} X_{\mathrm{te}}^\top$$

is the empirical covariance of the globally whitened CIFAR-10 test set and v_j is the jth eigenvector of $A(0)$, we set

$$w_j = v_j^\top T v_j$$

and evaluate

$$L_{\mathrm{te}}(t) = \frac{1}{2p} \sum_{j=1}^{p} w_j (1 - a_j(t))^2.$$

The theoretical comparison is the proportional large-data prediction from Appendix C.4,

$$L_{\psi,\rho}(t) = \frac{1}{2} \int \big(\Theta_t(\rho\lambda) - 1\big)^2 \, d\mu_\psi(\lambda), \qquad \Theta_t(s) = \frac{s}{s + (1 - s)e^{-4t}},$$

with μ_ψ the Marchenko–Pastur law. The noncritical cases are evaluated by Gauss–Legendre quadrature; at $\psi = 1$ we use the stabilized closed form described in Appendix C.4.1. All plotted training, test, and theoretical curves are divided by their respective initial values.

D.3 Mean-field experiments

We use a globally whitened CIFAR-10 subset for the train/test comparison shown in the main text. Global whitening is fit on all $50\,000$ training images, after which we retain the top

$$p = 512$$

PCA coordinates and select a fixed subset of

$$m = 256$$

whitened training examples using subset seed 0. All $10\,000$ test images are transformed using the same full-training whitening map.

We fix

$$\rho = 0.5$$

and repeat the exact finite-width experiment for

$$\frac{n}{p} \in \{4, 16, 64\}, \qquad n \in \{2048, 8192, 32768\}.$$

For each width we use 16 independent Gaussian initializations,

$$A(0) \stackrel{d}{=} \frac{\rho}{n} Z^\top Z, \qquad Z_{\alpha i} \overset{\text{iid}}{\sim} \mathcal{N}(0, 1).$$

Let

$$S = Q \operatorname{diag}(\lambda_1, \ldots, \lambda_p) Q^\top$$

be the covariance of the training subset, and let T be the covariance of the whitened held-out test split. Define

$$a_t(\lambda) = \frac{\rho}{\rho + (1 - \rho) e^{-4\lambda t}}$$

and

$$w_i = q_i^\top T q_i.$$

The two mean-field comparison curves are

$$\widehat{L}_{\text{MF}}(t) = \frac{1}{2p} \sum_{i=1}^{p} \lambda_i \big(1 - a_t(\lambda_i)\big)^2,$$

and

$$L_{\text{te,MF}}(t) = \frac{1}{2p} \sum_{i=1}^{p} w_i \big(1 - a_t(\lambda_i)\big)^2.$$

Thus both theoretical curves use the actual finite-sample spectrum of the training subset; no Marchenko–Pastur approximation is made.

For the finite-width trajectories we evolve the exact Gram dynamics

$$\dot{A} = 2(SA + AS) - SA^2 - A^2 S - 2ASA.$$

Writing $E = I - A$ in the S eigenbasis gives a semilinear equation whose linear part is diagonal entrywise. We integrate it with the adaptive ETD2 scheme using

$$\text{rtol} = 5 \times 10^{-4}, \qquad \text{atol} = 10^{-11},$$

and record 161 equally spaced output times on $0 \le t \le 4$. Experimental error bars show one standard deviation over the 16 initialization seeds. The theory curve is common to the three values of n/p; increasing n/p tests convergence toward the mean-field limit.

D.4 Narrow-latent experiments

The narrow-latent experiments serve two purposes. The CIFAR-10 experiment tests the finite-time scalar reduction and the initialization-dependent phase structure. A separate Gaussian experiment tests the derived large-time loss asymptotics.

Globally whitened CIFAR-10 subset: phase structure. We globally whiten CIFAR-10 using the full training split and then choose a subset of

$$m = 1024$$

examples. The experiment uses

$$p = 3072, \qquad n = 1, \qquad \phi = \frac{p}{m} = 3,$$

and 16 Gaussian initialization seeds.

Because $n = 1$, finite-width row competition is absent and the narrow-latent order parameters can be followed directly. Let $\lambda_1, \ldots, \lambda_m$ be the nonzero eigenvalues of the subset covariance. In its

eigenbasis write C_i for the active coordinates of the single weight row and let M denote its total squared norm in the data nullspace. Then

$$a = \sum_{i=1}^{m} C_i^2 + M, \qquad r = \sum_{i=1}^{m} \lambda_i C_i^2,$$

and the original finite-dimensional gradient flow is exactly

$$\dot{C}_i = C_i(2\lambda_i - r - a\lambda_i), \qquad \dot{M} = -2rM.$$

We additionally integrate

$$\dot{q} = 2(2-a), \qquad q(0) = 0,$$

which allows the finite-width spectral clock to be compared directly with the limiting one.

The finite-width scaled loss corrections are

$$\widehat{F}_N(t) = \psi\left(\widehat{L}(t) - \frac{\operatorname{Tr} S}{2p}\right),$$

and

$$F_N(t) = \psi\left(L(t) - \frac{1}{2}\right), \qquad \psi = \frac{p}{n}.$$

For $n = 1$ these reduce exactly to

$$\widehat{F}_N = r\left(\frac{a}{2} - 1\right), \qquad F_N = a\left(\frac{a}{2} - 1\right).$$

The population observable here is the isotropic population loss from the theorem, not reconstruction loss on the CIFAR-10 test split. This choice isolates the train/population distinction predicted by the theory from additional anisotropy of the empirical CIFAR-10 test covariance.

We consider four initialization scenarios:

scenario	$\kappa = p\sigma^2$
monotonic alignment	0.5
nonmonotonic alignment	4
trapped	10
critical	κ_c

Writing

$$w_0 = \frac{p - \operatorname{rank}(S)}{p},$$

the critical value is the solution $\kappa_c > 2$ of

$$\frac{\kappa_c^2}{\kappa_c - 1} w_0^2 = 4.$$

For $p = 3072$ and a full-rank subset of size 1024, $w_0 = 2/3$ and $\kappa_c \simeq 7.854$.

For the limiting comparison we use the actual empirical subset spectrum. Specifically,

$$M_S(q) = \frac{1}{p}\left[p - m + \sum_{i=1}^{m} e^{q\lambda_i}\right],$$

$$m_S(q) = \frac{\sum_i \lambda_i e^{q\lambda_i}}{p - m + \sum_i e^{q\lambda_i}},$$

and solve

$$\dot{a} = 4a\, m_S(q)(1-a), \qquad \dot{q} = 2(2-a), \qquad r = a\, m_S(q).$$

The associated limiting loss curves are

$$\widehat{F}(t) = r(t)\left(\frac{a(t)}{2} - 1\right), \qquad F(t) = a(t)\left(\frac{a(t)}{2} - 1\right).$$

The training-loss plot also shows the finite-sample rank-one PCA optimum,

$$\widehat{F}_{\mathrm{PCA}} = -\frac{\lambda_{\max}(S)}{2}.$$

This makes the optimization failure in the trapped phase visible directly: the limiting trapped trajectory approaches $\widehat{F} = 0$ rather than the PCA optimum. The plots of $a(t)$, $q(t)$, and $r(t)$ additionally display the nonmonotonic-alignment mechanism: for $\kappa > 2$, q first decreases, reaches its minimum when $a \simeq 2$, and subsequently reverses direction in the alignment phase.

We integrate the exact $n = 1$ system using RK4 with the step size scaled by the empirical upper spectral edge,

$$\Delta t = \frac{0.01}{\lambda_{\max}(S)},$$

up to $t = 30$.

Gaussian data: large-time tails. To test the narrow-latent large-time formulas without residual finite-sample anisotropy from CIFAR-10, we use a larger Gaussian experiment with

$$p = 3072 \times 8 = 24576, \qquad m = 1024 \times 8 = 8192, \qquad n = 1,$$

so again $\phi = 3$. A single Gaussian training matrix is generated with data seed 12345, and we average over eight independent weight initializations.

The experiment evaluates alignment, critical, and trapped initializations,

$$\kappa_{\mathrm{align}} = 4, \qquad \kappa_{\mathrm{crit}} = \kappa_c, \qquad \kappa_{\mathrm{trap}} = 9.$$

The exact finite-width active/null-space dynamics are integrated with RK4 using

$$\Delta t = \frac{0.02}{\lambda_{\max}^{\mathrm{emp}}}, \qquad 0 \le t \le 10.$$

At $\phi = 3$ the asymptotic comparison uses the Marchenko–Pastur edges

$$\lambda_{\pm} = (1 \pm \sqrt{3})^2.$$

For alignment, the derived loss residuals are

$$\widehat{F}(t) - \widehat{F}(\infty) \sim \frac{3}{8t},$$

and

$$F(t) - F(\infty) \sim \frac{1}{2} K_a^2 t^6 e^{-8\lambda_+ t}.$$

D.5 Small-data experiments

We validate the small-data prediction of Appendix C.5 using globally whitened CIFAR-10. The experiment is designed to separate the two geometric requirements involved in the comparison. The whitening transformation is fit globally, using the complete CIFAR-10 training split, so that held-out test images can be evaluated in the same approximately isotropic coordinate system. The much smaller subset actually used for optimization is then orthogonalized separately, so that the active-space near-isometry assumed by the small-data theory holds exactly at finite dimension.

Global whitening and dimensionality reduction. Each 32×32 RGB CIFAR-10 image is first flattened into $\mathbb{R}^{3072}$ and rescaled by $1/255$. We compute the empirical mean and covariance using all 50 000 training images,

$$\bar{x} = \frac{1}{m_{\mathrm{full}}} \sum_{\mu=1}^{m_{\mathrm{full}}} x_\mu, \qquad C_{\mathrm{full}} = \frac{1}{m_{\mathrm{full}}} \sum_{\mu=1}^{m_{\mathrm{full}}} (x_\mu - \bar{x})(x_\mu - \bar{x})^\top.$$

Writing

$$C_{\mathrm{full}} = V \Lambda V^\top,$$

we retain the $p = 1024$ principal directions with the largest eigenvalues and apply the whitening transformation

$$z = \Lambda_p^{-1/2} V_p^\top (x - \bar{x}).$$

The inverse square root is formed only on numerically stable directions: eigenvalues below 10^{-8} times the largest eigenvalue are discarded before the top p directions are selected.

The same mean $\bar{x}$ and the same whitening transformation are applied to all $10\,000$ CIFAR-10 test images. Thus no test-set statistics enter either preprocessing or training. Let

$$T_{\mathrm{te}} = \frac{1}{m_{\mathrm{te}}} X_{\mathrm{te}} X_{\mathrm{te}}^\top$$

denote the empirical second-moment matrix of the resulting globally whitened test data.

Orthogonalized training subsets. After the global whitening map has been fixed, we randomly select 128 whitened training examples using subset seed 0. Arrange them as the columns of

$$X_{\mathrm{sel}} \in \mathbb{R}^{p \times 128}$$

and compute a thin QR factorization

$$X_{\mathrm{sel}} = QR, \qquad Q^\top Q = I_{128}.$$

We replace the selected training vectors by

$$X_{\mathrm{orth}} = \sqrt{p}\, Q.$$

Consequently,

$$\frac{1}{p} X_{\mathrm{orth}}^\top X_{\mathrm{orth}} = I_{128}$$

up to floating-point precision.

We consider the nested subset sizes

$$m \in \{128, 32, 8\},$$

where, for each m, the training matrix X_m consists of the first m columns of X_{orth}. Thus

$$\frac{1}{p} X_m^\top X_m = I_m$$

for every experiment. Only the small training subset is orthogonalized. The held-out test images are merely globally whitened and are otherwise left unchanged.

We take

$$p = n = 1024, \qquad \psi = \frac{p}{n} = 1, \qquad \rho = n\sigma^2 = 1.$$

The same Gaussian initialization is used for all three values of m,

$$U_{\alpha i}(0) \overset{\mathrm{iid}}{\sim} \mathcal{N}\Big(0, \frac{\rho}{n}\Big),$$

with initialization seed 1. Sharing the initialization isolates the effect of changing the amount of training data.

Exact finite-width dynamics. For each m, we train on the orthogonalized matrix X_m using the small-data learning-rate scaling

$$\eta = m.$$

Rather than forming the $p \times p$ empirical covariance explicitly, we evaluate the original weight dynamics through matrix products with X_m. Writing

$$E_X = X_m - U^\top U X_m,$$

the gradient flow is evaluated as

$$\dot{U} = \frac{1}{p} \left[(UX_m) E_X^\top + (UE_X) X_m^\top \right].$$

This is an algebraic reassociation of the original tied-autoencoder gradient flow and introduces no dynamical approximation.

We integrate this equation by fourth-order Runge–Kutta with

$$\Delta t = 5 \times 10^{-3}, \qquad 0 \le t \le 1,$$

and record the trajectory every two integration steps.

The training loss is evaluated exactly as

$$\widehat{L}_m(t) = \frac{1}{2pm} \left\| X_m - U(t)^\top U(t) X_m \right\|_F^2 .$$

For the globally whitened test covariance T_{te}, we compute the held-out reconstruction loss

$$L_{\mathrm{te}}(t) = \frac{1}{2p} \operatorname{Tr}\left[T_{\mathrm{te}} (I - U(t)^\top U(t))^2 \right] .$$

This expression is algebraically identical to averaging the reconstruction loss over all 10 000 globally whitened test images. In the numerical implementation it is evaluated through T_{te}, avoiding repeated passes through the full test set.

Small-data theoretical prediction. For each m we solve the limiting hierarchy of Appendix C.5,

$$\dot{a} = 4a(1-a) - 2a\mu_1,$$

and

$$\dot{\mu}_k = -2\mu_{k+1} - 2a \sum_{j=0}^{k-1} \mu_j \mu_{k-j} + 2\mu_1 \mu_k, \qquad k \ge 1, \qquad \mu_0 = 1.$$

To account for the finite dimension of the inactive subspace, the initial moments use

$$\psi_\perp = \frac{p-m}{n}$$

instead of its limiting value ψ. Explicitly,

$$\mu_k(0) = \rho^k \sum_{j=1}^{k} N_{k,j} \psi_\perp^j,$$

where $N_{k,j}$ are the Narayana numbers.

We truncate the hierarchy at

$$K = 32, \qquad \mu_{K+1} = 0,$$

and integrate it with DOP853 using relative and absolute tolerances 10^{-9} and 10^{-11}, respectively.

For the training loss we use the finite-m correction

$$\widehat{L}_m^{\mathrm{th}}(t) = \frac{1}{2}(1-a(t))^2 + \frac{1}{2} a(t) \mu_1(t) + \frac{\rho^2 (m+1)}{2n}.$$

For the population prediction, define

$$J(t) = 2 \int_0^t a(s) \big(\mu_1(s) - \mu_2(s) \big) \, ds$$

and

$$\epsilon = \frac{m}{p}.$$

The finite-size isotropic prediction used in the experiment is

$$L_{\mathrm{iso}}^{\mathrm{th}}(t) = L_0 + \epsilon \left[\frac{1}{2}(1-a(t))^2 + a(t)\mu_1(t) - C_0 + J(t) \right],$$

where

$$L_0 = \frac{1}{2}\left[1 - 2\rho + \rho^2\left(1 + \frac{p+1}{n}\right)\right]$$

and

$$C_0 = \frac{1}{2}\left[1 - 2\rho + (1+2\psi)\rho^2\right].$$

Appendix C.5 derives this expression for an isotropic population covariance, whereas the finite globally whitened CIFAR-10 test covariance T_{te} is not exactly scalar. We therefore define

$$c_{\mathrm{te}} = \frac{1}{p}\operatorname{Tr} T_{\mathrm{te}}$$

and use

$$L_{\mathrm{te}}^{\mathrm{th}}(t) = c_{\mathrm{te}} L_{\mathrm{iso}}^{\mathrm{th}}(t)$$

as the theoretical comparison curve. This formula would be exact if $T_{\mathrm{te}} = c_{\mathrm{te}} I_p$.

We additionally report

$$\lambda_{\min}(T_{\mathrm{te}}), \qquad \lambda_{\max}(T_{\mathrm{te}}), \qquad \frac{\|T_{\mathrm{te}} - c_{\mathrm{te}} I_p\|_{\mathrm{op}}}{c_{\mathrm{te}}}$$

as diagnostics of the residual test-set anisotropy. Thus the training comparison directly tests the small-data theorem, while the held-out population comparison should be interpreted as testing its isotropic prediction after global whitening, with the remaining discrepancy also reflecting finite test-covariance anisotropy.

The resulting figures show, separately, the training and held-out population losses for

$$m \in \{128, 32, 8\},$$

with solid curves denoting the exact finite-width dynamics and dashed curves denoting the corresponding small-data predictions.