GLOBAL EXPONENTIAL CONVERGENCE OF TWO-LAYER LINEAR NETWORK TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

We prove global exponential (linear) convergence with an explicit rate in the rich scaling for wide two-layer linear networks trained with smooth Polyak–Łojasiewicz predictor losses. Gradient flow in the factors closes exactly in terms of a finite-dimensional Bures flow of the neuron law covariance, in which the predictor dynamics are preconditioned by hidden covariance blocks. Mean-field conservation laws provide uniform spectral lower bounds on the hidden preconditioning blocks when the initial covariance satisfies a spectral support gap condition. This condition encompasses positive definiteness while still allowing for singular initializations. For an initial covariance $\Sigma_0 = \sigma^2 \mathrm{Id}$, the loss converges to the global minimum with linear rate at least $4\sigma^2\kappa$, where κ is the PL constant. We establish stability of this rate under finite-width sampling, as well as global convergence of factor gradient descent for an explicit stepsize interval depending on smoothness, the initial loss, and conserved spectral margins. Our argument extends layerwise to deep linear ResNets, subject to a residual-path bound. In the case of heavy-ball momentum, training dynamics close instead over positions and velocities in terms of a lifted phase covariance. Linear convergence holds under an explicit condition on the energy and damping, specifying a window of admissible dampings. For two-scale white initializations, this interval is nonempty for sufficiently large position scales, with a fixed initial loss gap and velocity covariance. Numerical experiments illustrate the covariance geometry and compare the predicted and observed rates.

1 INTRODUCTION

Two-layer linear networks are nonconvex in their weight factors, even when the loss is a simple function of the predictor matrix. Training moves in the lifted space of incoming and outgoing neuron weights, whose second moments act as a time-dependent preconditioner on the predictor. These moments may in principle become degenerate and slow down the dynamics, and it is not clear that the dynamics necessarily converge to a globally optimal predictor. We consider this question in the rich mean-field scaling, in which both sides of each rank-one neuron evolve, and smooth predictor losses satisfying a Polyak–Łojasiewicz (PL) inequality, which subsume the standard quadratic loss.

Central to our analysis are width-independent conservation laws for the weight covariance: the signed covariance spectrum, relative to the input-output block split, supplies conserved supporting inequalities and lower bounds on each hidden-block. The sum of these lower bounds conditions the predictor dynamics: a positive sum is sufficient to guarantee global convergence of the predictor loss with an explicit rate. This analysis covers every positive-definite initialization and some singular ones. For initial variance $\sigma^2 \mathrm{Id}$, the loss decays with rate at least $4\sigma^2\kappa$, where κ is the PL constant.

The covariance viewpoint also makes the result robust to finite-width sampling and factor gradient descent. For deep linear ResNets, the same conservation argument applied layerwise yields linear convergence conditional on a uniform bound on the residual path. With heavy-ball momentum, closure is lifted to the joint phase covariance of positions and velocities. Conformally transported spectral laws describe its geometry, while energy dissipation gives an explicit initialization and damping certificate for linear convergence. Special values of the conserved quantities further permit lower-dimensional intrinsic dynamics; these reductions are described in Appendix G.

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.

Related work Mean-field analyses describe the collective evolution of neuron weight distributions in wide two-layer networks (Chizat and Bach, 2018; Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018). These formulations treat broad nonlinear feature maps and often prove asymptotic global optimality or propagation-of-chaos type statements under support, homogeneity, noise, or regularity assumptions. In the present linear setting, exact second-moment closure connects training to Bures–Wasserstein covariance geometry (Takatsu, 2011; Bhatia et al., 2019), making a sharp quantitative linear convergence rate available. This closure distinguishes our analysis from the NTK (Jacot et al., 2018) and lazy training (Chizat et al., 2019) regime: both incoming and outgoing weights move, and the evolving representation is essential to the rate.

Deep linear networks have served as a tractable model of the nonconvexity created by matrix products, with results on their landscapes, exact dynamics, and convergence (Saxe et al., 2014; Kawaguchi, 2016; Laurent and von Brecht, 2018; Arora et al., 2019). Finite-width analyses of two-layer matrix factorization furnish convergence guarantees under small, asymmetric, or general initializations (Ye and Du, 2021; Xu et al., 2023), while recent work treats discrete-time low-rank adaptation directly (Mu and Klabjan, 2026). Our result instead starts from a population covariance satisfying a positive *support gap* condition, in particular any positive-definite covariance, and obtains an explicit global rate that survives independent finite-width sampling. Related work identifies Gramian differences and balancedness as invariants of factor training (Du et al., 2018; Tarmoun et al., 2021) and develops a general differential framework for identifying conservation laws (Marcotte et al., 2023). Our mean-field laws are functionals of the neuron distribution that induce compatible particle laws at every width. The conserved spectrum of the signed covariance yields hidden-block lower bounds without requiring a special balancedness identity.

Residual networks (He et al., 2016a;b) were introduced as architectures built from perturbations of the identity, and their infinite depth limit leads naturally to neural ODEs (He et al., 2016a; Chen et al., 2018). Mean-field formulations have established global optimality and convergence results under expressivity, regularity, or local PL conditions (Lu et al., 2020; Ding et al., 2021; Isobe, 2023; Barboni et al., 2025). Our linear analysis provides a complementary condition: uniform conditioning of the forward and backward residual products is sufficient for convergence.

Momentum turns first-order descent into an inertial, second-order iteration; the heavy-ball method of Polyak (1964) and Nesterov acceleration (Nesterov, 1983) have mature convergence theories, including continuous-time interpretations (Su et al., 2016) and robust-control analyses (Lessard et al., 2016). For deep linear networks, Wang et al. (2021); Liu et al. (2022) establish acceleration results under initialization and overparameterization conditions. Here the appropriate state variable is the phase covariance, over which the dynamics close. We give an explicit energy-damping certificate for global convergence with general position-velocity correlations.

Our contributions Section 2 identifies width-independent spectral conservation laws and derives uniform hidden-block bounds from their support margins (Propositions 2.1 and 2.3). Their Wasserstein interpretation and local completeness among smooth covariance laws common to all split losses are developed in Appendices B and F. Section 3 proves global exponential convergence with an explicit rate under a smooth PL predictor loss from every initial covariance with positive support gap (Theorem 3.1). The rate survives independent finite-width sampling at width proportional to the population covariance rank, which can be as small as $\min(d_u, d_v)$ (Corollary 3.4 and Remark 3.5). We establish global linear convergence of factor gradient descent under an explicit stepsize bound by controlling the accumulated conservation error (Theorem 3.6). Section 4 extends the layerwise argument to finite- and continuous-depth linear ResNets under residual-path control (Theorems 4.1 and 4.2), and discusses why that control is not automatic (Remark C.19 in Appendix C.6). Section 5 shows that heavy-ball momentum dynamics closes on the phase covariance and identifies conformal conservation laws (Proposition 5.1). Global exponential convergence is proven under an initial-energy and damping condition, including a two-scale white specialization (Theorem 5.2 and Corollary 5.3).

Setup Writing training in terms of the predictor separates the loss geometry from the nonconvex factorization. A width-n two-layer linear network has factors $U := (u_1, \dots, u_n) \in \mathbb{R}^{d_u \times n}$ and $V := (v_1, \dots, v_n) \in \mathbb{R}^{d_v \times n}$. Write $m := d_u + d_v$ and $W := \binom{U}{V}$. Its predictor is $\Sigma^{vu} := VU^\top/n$, and its factor objective is $\mathcal{L}(U, V) := \ell(VU^\top/n)$ for a differentiable predictor loss $\ell : \mathbb{R}^{d_v \times d_u} \to \mathbb{R}$. For example, if a data distribution ρ on pairs (z, y) satisfies $y = \Sigma^{vu}_\star z$ for a teacher $\Sigma^{vu}_\star$, squared

regression loss is

$$\ell_{\mathrm{quad}}(\Sigma^{vu}) := \tfrac{1}{2}\int \|(\Sigma^{vu} - \Sigma^{vu}_\star)z\|^2 \,\mathrm{d}\rho(z,y).$$

More generally, a self-adjoint positive semidefinite operator $\mathcal{Q}$ on $\mathbb{R}^{d_v\times d_u}$ defines a quadratic loss $\ell(S) := \frac{1}{2}\langle \mathcal{Q}[S-\Sigma^{vu}_\star], S-\Sigma^{vu}_\star\rangle$. The regression example corresponds to $\mathcal{Q}[X] = XC_z$, where $C_z := \int zz^\top \,\mathrm{d}\rho(z,y)$ is the input second moment.

We state the assumptions on the predictor loss: assume throughout that $\ell : \mathbb{R}^{d_v\times d_u} \to \mathbb{R}$ is C^1 and bounded below, and write $\ell_\star := \inf_S \ell(S)$.

Assumption 1.1 (Smooth PL predictor loss)**.** *There exist constants $L > 0$ and $\kappa > 0$ such that the loss ℓ is L-smooth and satisfies the κ-Polyak–Łojasiewicz inequality: $\|\nabla\ell(S) - \nabla\ell(S')\|_{\mathrm{F}} \le L\|S-S'\|_{\mathrm{F}}$ for all S, S', and $\|\nabla\ell(S)\|_{\mathrm{F}}^2 \ge 2\kappa\big(\ell(S)-\ell_\star\big)$ for all S.*

We take $0 < \kappa \le L$, as holds automatically for any nonconstant loss satisfying Assumption 1.1. For a nonzero quadratic Hessian $\mathcal{Q} \succeq 0$, one can take $L = \lambda_{\max}(\mathcal{Q})$ and $\kappa = \lambda^+_{\min}(\mathcal{Q})$. The operator $\mathcal{Q}$ may have a nontrivial kernel.

2 Mean Field Conservation Laws

Gradient flow and covariance closure Although training evolves all neuron weights, its effect on the predictor is governed by a fixed-size covariance matrix. We use the width-normalized gradient flow $(\dot U_t, \dot V_t) = -n\nabla_{(U,V)}\mathcal{L}(U_t, V_t)$, namely

$$\dot U_t = -E_t^\top V_t, \qquad \dot V_t = -E_t U_t, \qquad E_t := \nabla\ell(V_t U_t^\top/n). \tag{1}$$

The factor n fixes a common time scale across widths; equivalently this is Euclidean gradient flow of $n\mathcal{L}$. With $W_t := \binom{U_t}{V_t}$, define the uncentered covariance

$$\Sigma_t := \tfrac{1}{n}W_tW_t^\top = \begin{pmatrix}\Sigma_t^u & \Sigma_t^{uv}\\ \Sigma_t^{vu} & \Sigma_t^v\end{pmatrix}, \qquad \Sigma_t^u := \tfrac{U_tU_t^\top}{n}, \quad \Sigma_t^v := \tfrac{V_tV_t^\top}{n}, \quad \Sigma_t^{vu} := \tfrac{V_tU_t^\top}{n}.$$

The hidden blocks Σ_t^u, Σ_t^v do not enter the loss, but govern the predictor's motion. We use the same notation Σ whether the covariance is specified directly or realized by finite-width factors: the closed equation depends only on this matrix. Set $G_E := \left(\begin{smallmatrix}0 & E^\top\\ E & 0\end{smallmatrix}\right)$, $G_t := G_{E_t}$, and $G(S) := G_{\nabla\ell(S)}$, so $\dot W_t = -G_tW_t$. The split structure of the gradient is encoded by the signature matrix

$$J := \mathrm{diag}(\mathrm{Id}_{d_u}, -\mathrm{Id}_{d_v}), \qquad G_tJ + JG_t = 0.$$

Our first result combines covariance closure with width-independent spectral conservation laws.

Proposition 2.1 (Covariance closure and conservation laws)**.** *Let $\mathcal{H}_k(\Sigma) := \mathrm{tr}\big((J\Sigma)^k\big)$. On its interval of existence, the factor flow (1) closes on the covariance Σ_t as a Bures equation, for which the characteristic polynomial of $J\Sigma$ is preserved:*

$$\dot\Sigma_t = -(G_t\Sigma_t + \Sigma_tG_t), \qquad \tfrac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}_k(\Sigma_t) = 0, \quad k = 1,\dots,m, \tag{2}$$

The covariance equation (2) is the Bures gradient flow (Takatsu, 2011; Bhatia et al., 2019) of $F(\Sigma) := \ell(\Sigma^{vu})$, since $\nabla_\Sigma F = G/2$. Its fundamental matrix Φ_t, defined by $\dot\Phi_t = -G_t\Phi_t$ and $\Phi_0 = \mathrm{Id}$, satisfies

$$\Phi_t^\top J\Phi_t = \Phi_tJ\Phi_t^\top = J, \qquad \Sigma_t = \Phi_t\Sigma_0\Phi_t^\top. \tag{3}$$

The functions $\mathcal{H}_k$ are conserved at every width n, defining mean-field laws. Appendix B gives their measure formulation, the corresponding Wasserstein PDE, and describes a larger family of laws.

Conserved spectral margins The spectral conservation laws (2) give uniform lower bounds on the spectrum of the hidden blocks (Σ^u, Σ^v) through two margins of the full covariance.

Definition 2.2 (Covariance support margins and gap)**.** *For $\Sigma \succeq 0$, the support margins and their sum are*

$$a_\pm(\Sigma) := \max\{a \ge 0 : \Sigma \mp aJ \succeq 0\}, \qquad g(\Sigma) := a_+(\Sigma) + a_-(\Sigma). \tag{4}$$

These maxima are finite and nonnegative. If $\lambda_1 \le \cdots \le \lambda_m$ are the eigenvalues of $\Sigma^{1/2}J\Sigma^{1/2}$, *including zeros*, then

$$a_+(\Sigma) = \lambda_{d_v+1}, \qquad a_-(\Sigma) = -\lambda_{d_v}, \qquad g(\Sigma) = \lambda_{d_v+1} - \lambda_{d_v}. \tag{5}$$

Thus $g(\Sigma)$ is an indexed spectral gap. Equivalently, it is the length of the positive-semidefinite interval $[-a_-(\Sigma), a_+(\Sigma)]$ of the pencil $\Sigma - tJ$; such intervals are standard in matrix-pencil theory (Nguyen and Nguyen, 2024). Appendix C proves (5), including at singular covariances. For $\Sigma \succ 0$, $a_+(\Sigma)$ and $-a_-(\Sigma)$ are the signed eigenvalues of $\Sigma^{1/2} J \Sigma^{1/2}$ nearest zero on their respective sides, and $g(\Sigma) \geq 2\lambda_{\min}(\Sigma) > 0$.

Proposition 2.3 (Transported lower bounds)**.** *On its interval of existence, every solution of* (2) *from* $\Sigma_0 \succeq 0$ *preserves* $a_\pm(\Sigma_t) = a_\pm(\Sigma_0)$ *and satisfies*

$$\Sigma_t \mp a_\pm(\Sigma_0) J \succeq 0, \qquad \Sigma_t^u \succeq a_+(\Sigma_0)\mathrm{Id}_{d_u}, \qquad \Sigma_t^v \succeq a_-(\Sigma_0)\mathrm{Id}_{d_v}.$$

The margins $a_\pm(\Sigma)$ need not equal the minimum eigenvalues of the initial hidden blocks: they satisfy $a_+(\Sigma) \leq \lambda_{\min}(\Sigma^u)$ and $a_-(\Sigma) \leq \lambda_{\min}(\Sigma^v)$, with equality when $\Sigma^{vu} = 0$. For instance, letting $\Sigma = \left(\begin{smallmatrix} 1 & r \\ r & 1 \end{smallmatrix}\right)$, $|r| \leq 1$, both margins are $\sqrt{1-r^2}$, whereas both hidden variances are one. The support gap $g(\Sigma)$ controls predictor dissipation and can remain positive even when the full covariance is degenerate. Under special conditions on the conserved spectrum of $J\Sigma$, elimination of hidden blocks becomes possible: Appendix G describes these intrinsic dynamics.

Numerical illustration: the 2×2 covariance cone For scalar input and output, every covariance has the form $\Sigma = \left(\begin{smallmatrix} p & r \\ r & q \end{smallmatrix}\right) \succeq 0$. Take the loss $F(\Sigma) := (r-\beta)^2/2$, with scalar teacher β. Writing $(a, b) := ((p+q)/2, (p-q)/2)$ identifies $\Sigma \succeq 0$ with the Lorentz cone $a^2 \geq r^2 + b^2$, $a \geq 0$. The Bures flow (24) becomes $(\dot{a}, \dot{b}, \dot{r}) = -(2r(r-\beta), 0, 2a(r-\beta))$, preserving $b = b_0$ and $\alpha^2 := a^2 - r^2 - b^2 = pq - r^2$.

When $p = q$, set $s := p$. Then $b = 0$ and trajectories follow hyperbolas $s^2 - r^2 = \alpha^2$ in $s \geq |r|$ (Figure 1(a)). For $\alpha > 0$, they converge to $(r, s) = (\beta, \sqrt{\beta^2 + \alpha^2})$; the spectral-margin bound is exact for the minimum s on the leaf. Unequal diagonal blocks give intersections of the planes $b = b_0$ with determinant leaves in the three-dimensional cone (panel (b)). If $b_0^2 + \alpha^2 > 0$, the endpoint is $(r_\infty, a_\infty) = (\beta, \sqrt{\beta^2 + b_0^2 + \alpha^2})$.

The invariants $2b$ and $-\alpha^2$ are the trace and determinant of $J\Sigma$, whose eigenvalues are $b_0 \pm \sqrt{b_0^2 + \alpha^2}$. Thus $g(\Sigma_0) = 2\sqrt{b_0^2 + \alpha^2}$: the convergence bound covers both panels, including singular leaves with $b_0 \neq 0$. Remark C.8 in Appendix C.3 treats the zero-support-gap boundary. Figure 1 is obtained by integrating $\dot{r} = -2\sqrt{r^2 + b_0^2 + \alpha^2}(r - \beta)$.

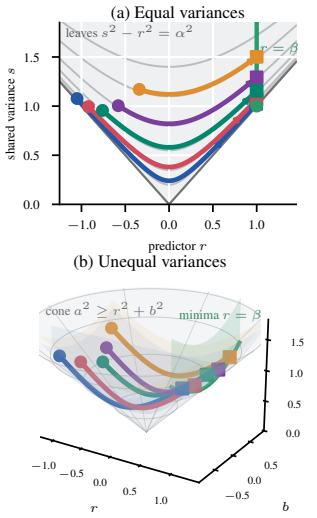

Figure 1: **Bures trajectories in the covariance cone.** (a) Equal diagonal blocks: invariant hyperbolas. (b) Unequal blocks: fixed-b sections of determinant leaves. The displayed trajectories converge to $r = \beta$.

3 Linear Convergence of Training Dynamics

Convergence of gradient flow A positive conserved support gap $g(\Sigma_0) > 0$ guarantees global linear convergence for the Bures flow (2), including some cases where the initial covariance Σ_0 is singular. Write $\Delta_t := \ell(\Sigma_t^{vu}) - \ell_\star$ for the loss gap and $\mathcal{S}_\star$ for the set of global predictor minimizers.

Theorem 3.1 (Global convergence from a positive support gap)**.** *Under Assumption 1.1, let* $\Sigma_0 \succeq 0$ *be a possibly singular covariance with* $g(\Sigma_0) > 0$*. Then the solution of* (2) *is global and satisfies*

$$\Delta_t \leq \Delta_0 e^{-2\kappa g(\Sigma_0) t}, \qquad t \geq 0. \tag{6}$$

As $t \to \infty$, $\Sigma_t \to \Sigma_\infty \succeq 0$ *with* $\operatorname{rank}\Sigma_\infty = \operatorname{rank}\Sigma_0$ *and* $\Sigma_\infty^{vu} \in \mathcal{S}_\star$*. The predictor* Σ^{vu} *and any factor realization (as per* (1)*) converge linearly with rate* $\kappa g(\Sigma_0)$ *in Frobenius norm.*

The only nondegeneracy requirement is $g(\Sigma_0) > 0$. Positive definiteness is sufficient but not necessary: $\Sigma_0 \succ 0$ implies $g(\Sigma_0) \geq 2\lambda_{\min}(\Sigma_0) > 0$ and then $\Sigma_\infty \succ 0$. The exponent separates loss geometry (captured by the PL constant κ) from initial covariance conditioning (the support gap $g(\Sigma_0)$), with no extra dimension factor. The trajectory selects a global minimizer according to its initialization and dynamics. Proposition C.4 supplies all explicit prefactors and bounds.

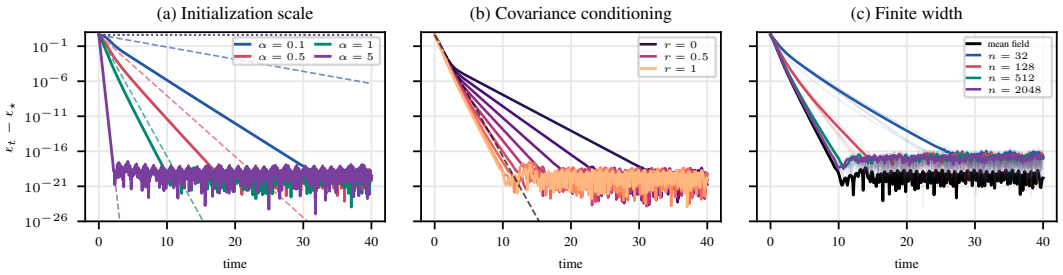

Figure 2: Numerical convergence for quadratic loss with $d_u = 16$, $d_v = 9$, a random teacher, and whitened input. **(a)** Bures flow from $\Sigma_0 = \alpha \mathrm{Id}_m$ at several white scales. Dashed lines show the spectral-margin rate $4\alpha\kappa$ from Theorem 3.1; dotted lines show the coarse rate $4\kappa c_{\mathrm{coarse}}$ from the determinant comparison in Appendix C.3 (Remark C.7), using the exact initial sublevel radius $M = 2\|\Sigma_\star^{vu}\|_{\mathrm{F}}$. **(b)** Rate deterioration when one matched diagonal pair of Σ_0 is scaled by r: $r = 1$ is the white reference and $r = 0$ is rank-deficient. **(c)** Bures flow from the reference covariance $\widetilde{\Sigma}_0 = \mathrm{Id}_m$ compared with finite-width flows initialized by i.i.d. samples from $\mathcal{N}(0, \widetilde{\Sigma}_0)$, for $n = 32, 128, 512, 2048$ and five independent trials.

Remark 3.2 (Singular covariance and predictor rank)**.** *The initialization $\Sigma_0 = \mathrm{diag}(\mathrm{Id}_{d_u}, 0_{d_v})$ has $a_+(\Sigma_0) = 1$, $a_-(\Sigma_0) = 0$, and $g(\Sigma_0) = 1$. Its rank is only d_u, yet the theorem applies. There is no rank contradiction: a full-rank predictor needs rank $\min(d_u, d_v)$, not m. More generally, $a_+(\Sigma_0) > 0$ implies $\mathrm{rank}\,\Sigma_0 \geq d_u$, while $a_-(\Sigma_0) > 0$ implies $\mathrm{rank}\,\Sigma_0 \geq d_v$. The condition cannot be dropped altogether: for $d_u = d_v = 2$, $\Sigma_0 = \mathrm{diag}(1, 0, 1, 0)$ and $\ell(S) := \left\|S - e_2 e_2^\top\right\|_{\mathrm{F}}^2/2$, the flow is stationary and nonoptimal. Here $g(\Sigma_0) = 0$, despite the nonzero signed eigenvalues $+1, -1$.*

Remark 3.3 (Quadratic specialization)**.** *For $\ell(S) := \frac{1}{2}\langle \mathcal{Q}[S - \Sigma_\star^{vu}], S - \Sigma_\star^{vu}\rangle$ with $\mathcal{Q} \succeq 0$ a nonzero self-adjoint linear operator, the exponent is $2g(\Sigma_0)\lambda_{\min}^+(\mathcal{Q})$. For a white initialization $\Sigma_0 = \sigma^2 \mathrm{Id}_m$, one has $a_+(\Sigma_0) = a_-(\Sigma_0) = \sigma^2$, giving a rate of $4\sigma^2\lambda_{\min}^+(\mathcal{Q})$. The operator $\mathcal{Q}$ is allowed to have a nontrivial kernel: the predictor Σ^{vu} converges to a point in $\Sigma_\star^{vu} + \ker \mathcal{Q}$. If $\mathcal{Q}$ is positive definite, the predictor converges to the teacher $\Sigma_\star^{vu}$.*

For a quadratic predictor loss, Figure 2 compares the initialization scale, loss of conditioning, and the finite-width sampling we discuss next. The dotted curves use a coarse determinant-based rate, derived for comparison in Appendix C.3. If the support gap $g(\Sigma_0)$ vanishes, trace and initial loss alone cannot guarantee a uniform positive rate (Example C.9). Figure 6 in Appendix H provides the counterpart for a non-quadratic smoothed-Lasso predictor loss.

Convergence under finite width sampling Theorem 3.1 applies generally to neuron law covariances with positive support gap. We now focus on finite-width networks, whose initial empirical covariance is $\Sigma_0 := W_0 W_0^\top/n$. For random initialization, the goal is to make this empirical covariance close to a prescribed population covariance $\widetilde{\Sigma}_0$, thereby retaining its support gap and convergence rate. Draw the columns $w_i := (u_i, v_i)$ of $W_0 := \binom{U_0}{V_0}$ independently from the same *subgaussian* distribution: $w_i := \widetilde{\Sigma}_0^{1/2}\xi_i$, where $\mathbb{E}[\xi_i \xi_i^\top] = \mathrm{Id}_m$ and $\sup_{\|z\|=1}\|\langle z, \xi_i\rangle\|_{\psi_2} \leq K$. Here $\|X\|_{\psi_2} := \inf\{c > 0 : \mathbb{E}e^{X^2/c^2} \leq 2\}$ denotes the standard subgaussian norm (Vershynin, 2018); standard Gaussian sampling has dimension-independent K. Thus $\mathbb{E}[w_i w_i^\top] = \widetilde{\Sigma}_0$, whereas $\Sigma_0 = W_0 W_0^\top/n$ is the empirical covariance of finitely many sampled neurons. The loss gap along the sampled trajectory remains $\Delta_t = \ell(\Sigma_t^{vu}) - \ell_\star$ and we write $\widetilde{\Delta}_0 := \ell(\widetilde{\Sigma}_0^{vu}) - \ell_\star$ for its population reference.

Corollary 3.4 (Finite width preserves the population rate)**.** *Under Assumption 1.1, let $\widetilde{\Sigma}_0 \succeq 0$ with $g(\widetilde{\Sigma}_0) > 0$, set $r_0 := \mathrm{rank}\,\widetilde{\Sigma}_0$, and fix $\varepsilon, \delta \in (0, 1)$. There is a universal $C > 0$ such that, if*

$$n \geq CK^4\varepsilon^{-2}\big(r_0 + \log(2/\delta)\big), \tag{7}$$

then with probability at least $1 - \delta$, it holds that $(1-\varepsilon)\widetilde{\Sigma}_0 \preceq \Sigma_0 \preceq (1+\varepsilon)\widetilde{\Sigma}_0$. On this event, $g(\Sigma_0) \geq (1-\varepsilon)g(\widetilde{\Sigma}_0)$, the empirical factor flow (1) is global, and $\Delta_t \leq \Delta_0 e^{-2\kappa(1-\varepsilon)g(\widetilde{\Sigma}_0)t}$ for all $t \geq 0$. The factors (U_t, V_t) converge to a limit with globally optimal predictor $\frac{1}{n}V_\infty U_\infty^\top = \Sigma_\infty^{vu}$.

Remark 3.5 (Width and initialization geometry)**.** *The condition $g(\widetilde{\Sigma}_0) > 0$ forces $r_0 \geq \min(d_u, d_v)$, and this minimum is attainable. If $d_u \leq d_v$, take $\widetilde{\Sigma}_0 = \mathrm{diag}(2\sigma^2\mathrm{Id}_{d_u}, 0_{d_v})$, $\sigma > 0$: then $r_0 = d_u$ and*

$g(\widetilde{\Sigma}_0) = 2\sigma^2$. *With Gaussian sampling, Corollary 3.4 gives width proportional to* $\varepsilon^{-2}(d_u + \log(2/\delta))$ *and rate* $4(1-\varepsilon)\kappa\sigma^2$*, the same as for full-rank population covariance* $\sigma^2 \mathrm{Id}_m$*. Here every* v_i *starts at zero, so* $\Sigma_0^{vu} = 0$ *and* $\Delta_0 = \widetilde{\Delta}_0$ *exactly. Exchanging the blocks gives dependence on* $\min(d_u, d_v)$ *in general. This improvement changes the initialization: for the full-rank white law,* $r_0 = m$*. Its empirical support gap can vanish at small width even though its population support gap is positive (Example C.11).*

For the sampling scheme in Corollary 3.4, the prefactor is the empirical initial loss gap Δ_0, which need not equal the population loss gap $\widetilde{\Delta}_0$. Controlling it is a separate accuracy requirement: for isotropic Gaussian sampling and the standard quadratic loss, the expected excess initial error is $\sigma^4 d_u d_v/(2n)$ (Proposition C.13). Theorem C.12 gives a separate accuracy budget and a deterministic loss envelope. For population variance $\sigma^2 \mathrm{Id}_m$ and $\varepsilon = 1/2$, the certified exponent is $2\kappa\sigma^2$.

Convergence of gradient descent Unlike finite-width sampling of neurons, discretized gradient descent in the factor matrices (U, V) with a finite step size changes the covariance evolution itself. We show that sufficiently small step sizes retain global linear convergence by controlling the accumulated conservation error. We consider discrete-time gradient descent (GD) on the factor matrices, using an Euler discretization of (1) with step size τ. As in the gradient flow setting, define $\Sigma_k := W_k W_k^\top/n$, $\Sigma_k^{vu} := V_k U_k^\top/n$, and $\Delta_k := \ell(\Sigma_k^{vu}) - \ell_\star$. Setting $E_k := \nabla\ell(\Sigma_k^{vu})$ and $G_k := G_{E_k}$, the iterates read

$$U_{k+1} = U_k - \tau E_k^\top V_k, \qquad V_{k+1} = V_k - \tau E_k U_k. \tag{8}$$

Equivalently, $W_{k+1} = W_k - n\tau\nabla_W \mathcal{L}(U_k, V_k)$: this is GD for $n\mathcal{L}$ with stepsize τ, or for $\mathcal{L}$ with stepsize $n\tau$. Writing $W_{k+1} = M_k W_k$ with $M_k := \mathrm{Id} - \tau G_k$ gives

$$\Sigma_{k+1} = M_k \Sigma_k M_k = \Sigma_k - \tau(G_k\Sigma_k + \Sigma_k G_k) + \tau^2 G_k \Sigma_k G_k. \tag{9}$$

The additional term $\tau^2 G_k\Sigma_k G_k$ arises from taking the GD step in the individual factors (U, V) and is absent from an Euler step taken directly on the Bures ODE.

A positive initial support gap $g(\Sigma_0)$ still guarantees global linear convergence for sufficiently small step sizes τ. The condition on τ is given below and depends only on the initial covariance, initial loss gap, and loss parameters; it controls both descent and the accumulated conservation error.

Theorem 3.6 (Small-step global convergence of factor descent)**.** *Under Assumption 1.1, let* $\Sigma_0 := W_0 W_0^\top/n$ *satisfy* $g(\Sigma_0) > 0$*, and suppose* $\Delta_0 > 0$*. Set* $I_\star := 4\sqrt{\Delta_0/\kappa}/g(\Sigma_0)$*. If*

$$0 < \tau \le \tau_{\mathrm{GD}} := g(\Sigma_0) e^{-2I_\star} \min\left\{ \frac{1}{4\|\Sigma_0\|_{\mathrm{op}}\sqrt{L\Delta_0} + 9\sqrt{2}L\|\Sigma_0\|_{\mathrm{op}}^2 e^{2I_\star}}, \quad \frac{\sqrt{2}-1}{4\Delta_0} \right\}, \tag{10}$$

then, with $q := 1 - \kappa\tau g(\Sigma_0)/\sqrt{2} \in (0,1)$ *it holds that* $\Delta_k \le q^k \Delta_0$*. The factors* W_k *and covariance* Σ_k *converge at rate* $q^{k/2}$ *to limits* W_∞ *and* $\Sigma_\infty \succeq 0$*, with* $\operatorname{rank}\Sigma_\infty = \operatorname{rank}\Sigma_0$ *and* $\Sigma_\infty^{vu} \in \mathcal{S}_\star$*; explicit prefactors are given in (46).*

The support-gap condition $g(\Sigma_0) > 0$ is implied by $\Sigma_0 \succ 0$, in which case Σ_∞ is also positive definite. The threshold τ_{GD} is positive for every initialization with $g(\Sigma_0) > 0$ and $\Delta_0 < \infty$; no initial loss gap assumption or dimension-dependent term is needed. The first restriction ensures descent, and the second controls the accumulated conservation error.

Appendix C.5 gives the explicit expression for τ_{GD} when $\Sigma_0 = \sigma^2 \mathrm{Id}_m$ and a comparison with the step-size condition obtained by linearizing near an equilibrium (Remark C.16). Proposition C.17 shows that sufficiently small step sizes yield a certified loss-decay exponent arbitrarily close to $2\kappa g(\Sigma_0)$, the gradient-flow exponent in Theorem 3.1. Corollary C.18 gives a step size chosen before sampling the factors, using the loss parameters and the bounds on the initial support gap, covariance norm, and loss gap from Theorem C.12.

4 Deep Linear Residual Networks

We now turn to deep linear residual networks (ResNets), including a continuous-depth formulation. To stabilize optimization as depth increases, the networks considered here use skip connections between successive layers and scale each residual update by the inverse depth. In this section, we demonstrate that linear ResNets also enjoy global exponential convergence rates that remain uniform

across depths, provided the average conserved layer support gap is bounded away from zero and the residual predictors satisfy an a priori operator-norm bound independent of depth throughout training. The latter is an additional hypothesis: layerwise conservation alone does not control forward and backward propagation. For simplicity, we state the results for gradient flow, as in Section 3.

Finite-depth setting We consider a depth-S, width-n network with matching input and output dimensions $d_u = d_v = d$, and set $h := 1/S$. For $s = 0, \dots, S-1$, layer s has factors $U_s, V_s \in \mathbb{R}^{d\times n}$, covariance $\Sigma_s := W_s W_s^\top/n$ with $W_s := \binom{U_s}{V_s}$, and predictor $\Sigma_s^{vu} := V_s U_s^\top/n$. The network maps an input $x^0 \in \mathbb{R}^d$ to its output x^S through

$$x^{s+1} = x^s + \tfrac{1}{Sn} V_s U_s^\top x^s = A_s x^s, \qquad A_s := \mathrm{Id} + h\Sigma_s^{vu}, \qquad s = 0, \dots, S-1.$$

Thus $x^S = P_S x^0$, where the endpoint predictor is the ordered product

$$P_S := A_{S-1} \cdots A_0 = \left(\mathrm{Id} + \tfrac{V_{S-1}U_{S-1}^\top}{Sn}\right) \cdots \left(\mathrm{Id} + \tfrac{V_0 U_0^\top}{Sn}\right).$$

Backpropagation in the factors couples the layers through these products, which can be singular even when every layer covariance is positive definite. Let $P_{>s} := A_{S-1} \cdots A_{s+1}$ and $P_{<s} := A_{s-1} \cdots A_0$, with empty products equal to Id. For $E := \nabla \ell(P_S)$, the backpropagated gradient is $\widetilde{E}_s := P_{>s}^\top E P_{<s}^\top$. In residual mean-field time, each layer obeys

$$\dot{\Sigma}_s = -(\widetilde{G}_s \Sigma_s + \Sigma_s \widetilde{G}_s), \qquad \widetilde{G}_s := G_{\widetilde{E}_s}. \tag{11}$$

This is the covariance equation corresponding to factor gradient flow for $nS\ell(P_S)$. Since $\widetilde{G}_s J + J\widetilde{G}_s = 0$, the signed margins are conserved separately at every layer.

Theorem 4.1 (Finite-depth convergence from conserved layer margins)**.** *Under Assumption 1.1 for the endpoint loss, suppose $\Sigma_s(0) \succ 0$ for all s, and write $\overline{g}_S := S^{-1}\sum_{s=0}^{S-1} g(\Sigma_s(0))$ for the average initial support gap. Suppose $\sup_s \|\Sigma_s^{vu}(t)\|_{\mathrm{op}} \le B < S$ for some constant $0 \le B < \infty$ on the maximal existence interval, and define the propagation lower bound $\underline{\sigma}_S := (1 - B/S)^{S-1}$. Then the flow is global and $\Delta_t := \ell(P_S(t)) - \ell_\star \le \Delta_0 e^{-2\kappa \overline{g}_S \underline{\sigma}_S^2 t}$. Every layer's covariance and factors converge, and the limiting endpoint is globally optimal.*

The rate combines the average support gap $\overline{g}_S$ with $\underline{\sigma}_S^2$. For fixed B, this product-conditioning factor tends to e^{-2B} and is at least e^{-4B} when $S \ge 2B$. For fixed B, a positive lower bound on the average layer support gap suffices for a convergence rate uniform in depth.

Continuous-depth limit Formally in the limit $S \to \infty$, residual products become propagators of a linear ODE. Theorem 4.1 can be extended to a continuum of residual layers when these propagators stay uniformly conditioned, but requires a well-defined global training trajectory. Let $s \in [0, 1]$ denote depth. At time t, the network maps x to $P(t)x$, where $P(t) := X_1(t)$ and $X_s(t)$ solves

$$\partial_s X_s(t) = \Sigma_s^{vu}(t) X_s(t), \qquad X_0(t) = \mathrm{Id}.$$

Write $P_{b,a}(t)$ for the propagator from depth a to depth b. The backpropagated gradient at depth s is $\widetilde{E}_s := P_{1,s}^\top \nabla \ell(P) P_{s,0}^\top$. We assume the covariance profile $s \mapsto \Sigma_s(t)$ is defined for all training times $t \ge 0$ and depends locally absolutely continuously on t as an element of $L^1([0,1]; \mathbb{S}^{2d})$. The layer covariances are positive semidefinite and satisfy (11) in L^1. Its precise time-integral form is (50) in Appendix C.6.

Theorem 4.2 (Continuous-depth convergence from averaged margins)**.** *Under Assumption 1.1 for the endpoint loss, consider a covariance profile satisfying the assumptions above, with $\Sigma_s(0) \succ 0$ for almost every s. Set $\overline{g} := \int_0^1 g(\Sigma_s(0))\,\mathrm{d}s \in (0, \infty)$. If $\sup_{t\ge 0}\int_0^1 \|\Sigma_s^{vu}(t)\|_{\mathrm{op}}\,\mathrm{d}s \le B < \infty$, then*

$$\ell(P(t)) - \ell_\star \le (\ell(P(0)) - \ell_\star) e^{-2\kappa \overline{g} e^{-2B} t}. \tag{12}$$

As $t \to \infty$, $\Sigma_s(t) \to \Sigma_s^\infty$ for almost every depth s, and profiles converge in $L^1([0,1]; \mathbb{S}^{2d})$. The network defined by the limiting profile attains loss $\ell_\star$. Moreover, $\Sigma_s^\infty \succ 0$ for almost every s, although its smallest eigenvalue need not be bounded away from zero uniformly over depth.

The need for additional residual-path boundedness assumptions arises because the loss depends only on the endpoint $s = 1$: see Remark C.19 in Appendix C.6.

5 HEAVY-BALL MOMENTUM CONVERGENCE

We now turn our attention to the analogous dynamics in an inertial setting that models heavy-ball momentum-based optimization. Upon introducing momentum, the system no longer closes on the position covariance of the neuron weights (u, v) alone: one must instead consider the phase space of positions and velocities tracked jointly. We show that, in terms of the *phase covariance*, dynamics are again closed for linear networks. We derive conservation identities describing this phase geometry; energy dissipation and control of position deformation give the convergence certificate below.

Dynamics and phase closure Inertia adds velocity as a state variable, so the natural moment over which the dynamics close is the joint position-velocity covariance. For $W := \binom{U}{V}$, predictor $\Sigma^{vu} := VU^\top/n$, and damping $\gamma > 0$, write $E := \nabla\ell(\Sigma^{vu})$. The factor dynamics are

$$\ddot{W} + \gamma\dot{W} + G(\Sigma^{vu})W = 0, \qquad G(\Sigma^{vu}) := \left(\begin{smallmatrix} 0 & E^\top \\ E & 0 \end{smallmatrix}\right). \tag{13}$$

Write $(P, Q) := (\dot{U}, \dot{V})$, so the columns of W and $\dot{W}$ are positions $x_i := (u_i, v_i)$ and velocities $y_i := (p_i, q_i)$ respectively. In this section $\Sigma := n^{-1}\binom{W}{\dot{W}}\binom{W}{\dot{W}}^\top$ denotes the full $2m$-dimensional phase covariance. Its blocks are $\Sigma^{(u,v)} := WW^\top/n$, $\Sigma^{(p,q)} := \dot{W}\dot{W}^\top/n$, and $\Xi := W\dot{W}^\top/n$; the predictor remains the vu block of $\Sigma^{(u,v)}$. With $G_t := G(\Sigma_t^{vu})$, the closed equation is

$$\dot{\Sigma}_t = \mathsf{M}_t\Sigma_t + \Sigma_t\mathsf{M}_t^\top, \qquad \Sigma_t = \begin{pmatrix} \Sigma_t^{(u,v)} & \Xi_t \\ \Xi_t^\top & \Sigma_t^{(p,q)} \end{pmatrix}, \qquad \mathsf{M}_t := \begin{pmatrix} 0 & \mathrm{Id}_m \\ -G_t & -\gamma\mathrm{Id}_m \end{pmatrix}. \tag{14}$$

Appendix C.7 gives the particle equations, measure formulation, and blockwise expansion.

Conformal laws and mechanical energy Without damping, the fundamental matrix generated by M_t in (14) is symplectic; damping rescales the associated bilinear form exponentially in time.

Proposition 5.1 (Conformal phase conservation laws)**.** *Let* $J_c := \left(\begin{smallmatrix} 0 & \mathrm{Id}_m \\ -\mathrm{Id}_m & 0 \end{smallmatrix}\right)$. *The phase fundamental matrix* $\dot{\Psi}_t = \mathsf{M}_t\Psi_t$, $\Psi_0 = \mathrm{Id}_{2m}$ *generated by* M_t *satisfies*

$$(\Psi_t)^\top J_c\Psi_t = e^{-\gamma t}J_c, \qquad \Sigma_t = \Psi_t\Sigma_0(\Psi_t)^\top.$$

Consequently, the spectrum of $e^{\gamma t}J_c\Sigma_t$ *and all its trace powers are constant. At the characteristic level, these identities imply that, for any two trajectories of the same flow,* $z_a := (x_a, y_a)$, $z_b := (x_b, y_b)$, $e^{\gamma t}z_a^\top J_c z_b$ *is conserved.*

Special spectral values permit reduced phase dynamics, we describe these in Appendix G. As before, write $\Delta_t := \ell(\Sigma_t^{vu}) - \ell_\star$ for the predictor loss gap. Define $\mathcal{K}_t := \mathrm{tr}(\Sigma_t^{(p,q)})/2$ and $\mathcal{E}_t := \Delta_t + \mathcal{K}_t$ as the kinetic and mechanical excess energies. Differentiating $\mathcal{E}_t$ along (14) gives $\dot{\mathcal{E}}_t = -2\gamma\mathcal{K}_t$. This identity gives global existence of the solution to (14) from every $\Sigma_0 \succeq 0$ (Proposition C.22). The remaining task is to derive spectral lower bounds for $\Sigma_t^{(u,v)}$.

Global convergence We seek an initialization condition on Σ_0 that keeps $\Sigma_t^{(u,v)}$ uniformly elliptic for the whole trajectory. Let $a_0 := \lambda_{\min}(\Sigma_0^{(u,v)}) > 0$ and $b_0 := \lambda_{\max}(\Sigma_0^{(u,v)})$. A deformation budget $s > 0$ specifies a spectral interval $[a_s, b_s] := [a_0e^{-2s}, b_0e^{2s}]$ for $\Sigma_t^{(u,v)}$. For $\mathcal{E}_0 > 0$, put $\rho_s := 4b_sL + 2\sqrt{2L\mathcal{E}_0}$ and define

$$\lambda_s(\gamma) := \tfrac{a_s\kappa\gamma}{\rho_s + 2\gamma^2}, \qquad \mathfrak{S}_s(\gamma) := \tfrac{4}{\lambda_s(\gamma)}\sqrt{\tfrac{3\mathcal{E}_0}{2a_s}}. \tag{15}$$

Here, λ_s is the exponential decay rate of $\mathcal{E}_t$, and $\mathfrak{S}_s$ bounds the resulting total deformation in $\Sigma_t^{(u,v)}$, conditional on the spectral interval $[a_s, b_s]$.

Theorem 5.2 (Global linear rate from a phase-energy certificate)**.** *Under Assumption 1.1, let* $\Sigma_0 \succeq 0$ *with* $\Sigma_0^{(u,v)} \succ 0$ *and* $\mathcal{E}_0 > 0$. *If* $\mathfrak{S}_s(\gamma) < s$ *for some* $s > 0$, *then for all* $t \geq 0$,

$$a_s\mathrm{Id}_m \preceq \Sigma_t^{(u,v)} \preceq b_s\mathrm{Id}_m, \qquad \mathcal{E}_t \leq 3\mathcal{E}_0e^{-\lambda_s(\gamma)t}.$$

The covariance $\Sigma_t^{(u,v)}$ *and any associated factors* (U_t, V_t) *evolving under* (13) *converge exponentially with rate* $\lambda_s/2$ *to limits with optimal predictor and* $\Sigma_\infty^{(u,v)} \succ 0$; $\Sigma_t^{(p,q)}$ *tends to zero.*

Corollary 5.3 (Two-scale white initialization). *Under Assumption 1.1, let* $\Sigma_0^{(u,v)} = \sigma_{vu}^2 \mathrm{Id}_m$, $\Xi_0 = 0$, *and* $\Sigma_0^{(p,q)} \succeq 0$ *have trace* $m\sigma_{qp}^2$, *with* $\sigma_{vu} > 0$, $\sigma_{qp} \geq 0$. *Then* $\mathcal{E}_0 = \Delta + m\sigma_{qp}^2/2$, *where* $\Delta := \ell(0) - \ell_\star$. *If* $\mathcal{E}_0 > 0$, *the certificate is*

$$2\gamma^2 - \frac{s\kappa\sigma_{vu}^3 e^{-3s}}{4\sqrt{3\mathcal{E}_0/2}}\gamma + 4L\sigma_{vu}^2 e^{2s} + 2\sqrt{2L\mathcal{E}_0} < 0.$$

For fixed $s, \gamma > 0, \Delta, \sigma_{qp}, m, L, \kappa$, *sufficiently large* σ_{vu} *satisfies this condition.*

Appendix E gives explicit bounds on the admissible γ interval and analyzes the large-σ_{vu} asymptotics.

Numerical view of the bootstrap certificate Writing $\mathfrak{G}_s(\gamma) := \mathfrak{S}_s(\gamma) - s$, Figure 3 shows the boundaries $\mathfrak{G}_s(\gamma) = 0$ for $1 \leq \sigma_{vu} \leq 10^3$, with $(d_u, d_v) = (3, 2)$, $\kappa = L = 1$, $\sigma_{qp} = 1$, and $\ell(0) - \ell_\star = 1$. The regions where $\mathfrak{G}_s(\gamma) < 0$ specify the pairs (s, γ) for which Theorem 5.2 guarantees spectral bounds and exponential energy decay. For the two-scale initialization, sufficiently large σ_{vu} yields a nonempty interval of γ satisfying Corollary 5.3. For fixed initialization and damping, a smaller admissible s gives a tighter tube and faster rate; Figure 5 illustrates the limiting rate at the feasibility boundary analyzed in Appendix E.

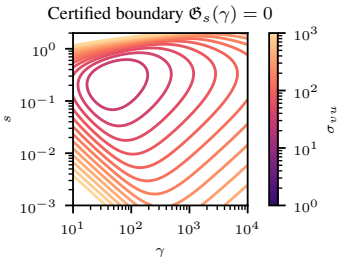

Figure 3: Two-scale white momentum initialization. Curves show $\mathfrak{G}_s(\gamma) = 0$ for the color-bar scale σ_{vu}; their interiors are certified damping windows.

Figure 4 compares empirical losses with the upper bounds on $\mathcal{E}_t$ from Corollary D.3. Each upper bound curve uses an explicit, approximately optimal rate certified by Theorem 5.2 (see Proposition D.2); no rate is fitted to the empirical trajectory. All displayed losses stay below their upper bounds. Notably in this inertial setting, the predictor loss may oscillate even while the mechanical energy decreases.

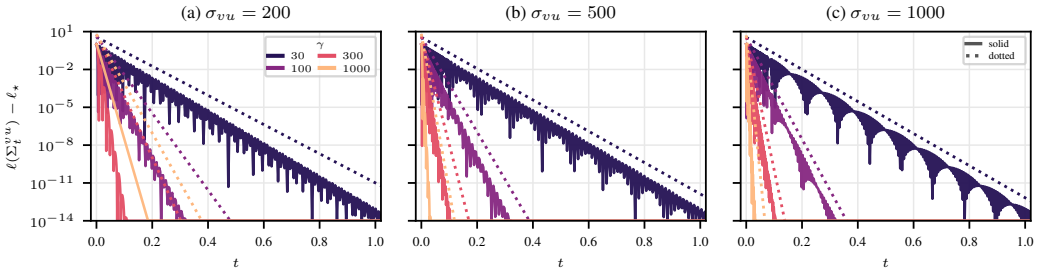

Figure 4: Empirical predictor-loss trajectories (solid) and the corresponding upper bounds on mechanical excess energy (dotted) for two-scale white initializations with varying σ_{vu} and damping γ: **(a)** $\sigma_{vu} = 200$, **(b)** $\sigma_{vu} = 500$, and **(c)** $\sigma_{vu} = 1000$. Each theoretical curve uses the explicit, approximately optimal rate on a tube certified by Theorem 5.2.

6 Conclusion

The present work addresses two-layer linear networks, making use of the exact closure of mean-field training on the neuron law covariance as a finite-dimensional Bures flow. This flow enjoys conservation laws that control the ellipticity of the predictor preconditioner through a conserved spectral support gap, yielding global linear convergence whenever this gap is positive. In particular, this is guaranteed for every positive-definite initial covariance and for smooth PL predictor losses. We show that this guarantee survives finite-width sampling of neurons, and extends to discrete-time gradient descent in the factors through approximate versions of the same conservation laws. The covariance viewpoint clarifies two further extensions: for deep linear ResNets, the same conservation law holds layerwise and gives linear convergence conditional on bounded residual paths. Heavy-ball momentum dynamics close in the joint phase covariance, where conformal conservation laws and mechanical energy lead to an explicit damping certificate from general initial phase moments. The certificate includes two-scale white initializations without imposing a phase-covariance identity. Covariance conditioning is the common quantitative mechanism throughout our analysis.

AI USE STATEMENT

In this work, we used generative AI tools for aiding and polishing writing, research ideation and execution, drafting sections of the paper, and proving mathematical claims. We have not used generative AI tools for retrieval and discovery (e.g. finding related work) or generating synthetic datasets. We have reviewed all AI-assisted work, having verified the claims, proofs and other material presented in this paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2019. arXiv:1810.02281.

Raphaël Barboni, Gabriel Peyré, and François-Xavier Vialard. Understanding the training of infinitely deep and wide ResNets with conditional optimal transport. *Communications on Pure and Applied Mathematics*, 78(11):2149–2205, 2025. doi: 10.1002/cpa.70004. arXiv:2403.12887.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019. doi: 10.1016/j.exmath.2018.01.002.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Lénaïc Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in Neural Information Processing Systems*, volume 31, 2018. arXiv:1805.09545.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019. arXiv:1812.07956.

Zhiyan Ding, Shi Chen, Qin Li, and Stephen Wright. On the global convergence of gradient descent for multi-layer ResNets in the mean-field regime, 2021. arXiv:2110.02926.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, volume 31, 2018. arXiv:1806.00900.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016a.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016b. doi: 10.1007/978-3-319-46493-0_38.

Noboru Isobe. A convergence result of a continuous model of deep learning via a Łojasiewicz–Simon inequality, 2023. arXiv:2311.15365v3, revised 2026.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. arXiv:1806.07572.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, volume 9851 of *Lecture Notes in Computer Science*, pages 795–811. Springer, 2016. doi: 10.1007/978-3-319-46128-1_50. Corrected version: arXiv:1608.04636v4.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, volume 29, pages 586–594, 2016. arXiv:1605.07110.

Thomas Laurent and James von Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2902–2907, 2018. arXiv:1712.01473.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016. doi: 10.1137/15M1009597.

Xin Liu, Wei Tao, and Zhisong Pan. A convergence analysis of Nesterov's accelerated gradient method in training deep linear neural networks. *Information Sciences*, 612:898–925, 2022. doi: 10.1016/j.ins.2022.08.090. arXiv:2204.08306.

Yiping Lu, Chao Ma, Yulong Lu, Jianfeng Lu, and Lexing Ying. A mean-field analysis of deep ResNet and beyond: Towards provable optimization via overparameterization from depth. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6426–6436, 2020.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: Conservation laws for gradient flows. In *Advances in Neural Information Processing Systems*, volume 36, pages 63210–63221, 2023. arXiv:2307.00144.

Sibylle Marcotte, Gabriel Peyré, and Rémi Gribonval. Intrinsic training dynamics of deep neural networks. In *International Conference on Learning Representations*, 2026. arXiv:2508.07370.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018. doi: 10.1073/pnas.1806579115. arXiv:1804.06561.

Siqiao Mu and Diego Klabjan. On the convergence rate of LoRA gradient descent. In *International Conference on Machine Learning*, 2026. arXiv:2512.18248.

Yurii E. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983.

Van-Bong Nguyen and Thi Ngan Nguyen. Positive semidefinite interval of matrix pencil and its applications to the generalized trust region subproblems. *Linear Algebra and its Applications*, 680: 371–390, 2024. doi: 10.1016/j.laa.2023.10.015. arXiv:2302.14352.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. doi: 10.1016/0041-555 3(64)90137-5.

Grant M. Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: Long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018. NeurIPS proceedings.

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

Weijie Su, Stephen Boyd, and Emmanuel J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17 (153):1–43, 2016.

Asuka Takatsu. Wasserstein geometry of Gaussian measures. *Osaka Journal of Mathematics*, 48(4): 1005–1026, 2011. doi: 10.18910/4973.

Salma Tarmoun, Guilherme Franca, Benjamin D. Haeffele, and René Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10153–10161, 2021.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, first edition, 2018. doi: 10.1017/9781108231596.

Jun-Kun Wang, Chi-Heng Lin, and Jacob D. Abernethy. A modular analysis of provable acceleration via Polyak's momentum: Training a wide ReLU network and a deep linear network. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10816–10827. PMLR, 2021. PMLR.

Ziqing Xu, Hancheng Min, Salma Tarmoun, Enrique Mallada, and René Vidal. Linear convergence of gradient descent for finite width over-parametrized linear networks with general initialization. In *International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2262–2284, 2023.

Tian Ye and Simon S. Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. In *Advances in Neural Information Processing Systems*, volume 34, pages 1429–1439, 2021. arXiv:2106.14289.

A NOTATION TABLE

The table collects the symbols, their meanings, and their first definitions.

Notation	Meaning	First reference
d_u, d_v, n, m	Input/output dimensions, width, and $m := d_u + d_v$	Section 1
$U, V, W := \binom{U}{V}, \mathcal{L}$	Factor matrices, stacked positions, and factor loss	Section 1
$\Sigma, \Sigma^u, \Sigma^v, \Sigma^{vu}$	Uncentered moment, hidden blocks, and predictor	Section 2
$\ell, \ell_\star, \mathcal{S}_\star$	Predictor loss, infimum, and minimizer set	Assumption 1.1; Section 3
L, κ, λ	Smoothness, PL, and QG constants	Assumption 1.1; Eq. (23)
$\mathcal{Q}, C_z, \Sigma_\star^{vu}$	Quadratic Hessian, input second moment, teacher	Section 1
$\mu, \widehat{\mu}_W$	Neuron law and empirical law	Appendix B
E, G_E	Predictor gradient and symmetric split generator	Section 2
J, Φ_t	Input/output signature and position fundamental matrix	Section 2
$\mathcal{H}, \mathcal{H}_{\varphi,r}$	Mean-field law and Gram-test law	Def. B.2; Eq. (19)
$\mathcal{H}_k, \Gamma$	Spectral laws and labeled Gram imbalance	Propositions 2.1, B.3
$a_\pm(\Sigma), g(\Sigma)$	Covariance support margins and gap	Def. 2.2
Δ_t, I_0, B_∞	Loss gap, force-length bound, covariance upper bound	Section 3; Proposition C.4
$\varepsilon, \zeta, \beta$	Rate tolerance, absolute accuracy, concentration tolerance	Corollary 3.4; Eq. (36)
$\widetilde{\Sigma}_0, \widetilde{\Delta}_0, r_0$	Population covariance, reference loss gap, and rank; empirical covariance is Σ_0	Corollary 3.4
$\underline{g}, \overline{\Delta}, \overline{B}_0$	Empirical support-gap lower bound, loss-gap and covariance-norm upper bounds	Eqs. (36), (38)
$\tau, I_\star$	GD step and force-length bound	Eq. (8); Theorem 3.6
$\alpha_\star, M_\nabla, B_\star$	Proof-local GD descent scale, initial gradient bound, and covariance bound	Eq. (43)
$M_k, \mathcal{J}_k, \mathcal{D}_k$	Factor update, transported signature, and its defect	Eq. (8), (40)
$S, A_s, P_S, \widetilde{E}_s$	Residual depth, layer map, endpoint, and backpropagated gradient	Section 4
$\overline{g}_S, \overline{g}$	Finite/continuous depth averages of $g(\Sigma_s(0))$	Theorems 4.1, 4.2
s, B	Continuous depth and residual-path bound	Section 4
$\underline{\sigma}_S, \overline{\sigma}_S$	Lower and upper residual-propagation bounds	Section 4; Appendix C.6
$x := (u, v), y := (p, q)$	Phase position and velocity variables	Section 5
$\Sigma, \Sigma^{(u,v)}, \Xi, \Sigma^{(p,q)}$	Full phase moment and its position, mixed, velocity blocks	Eq. (14)
γ, J_c	Damping and canonical symplectic matrix	Eq. (13), Proposition 5.1
M_t, Ψ_t	Phase generator and characteristic matrix	Eq. (14); Proposition 5.1
D, Q	Velocity regression and residual covariance (Schur complement)	Eq. (56)
$\mathcal{K}, \mathcal{E}$	Kinetic and total mechanical excess energies; $\mathcal{E}_t = \Delta_t + \mathcal{K}_t$	Section 5
$s, a_s, b_s, \rho_s, \lambda_s, \mathfrak{S}_s$	Deformation budget, position tube, force bound, rate, certificate; s here is not depth	Eq. (15)
σ_{vu}, σ_{qp}	White-position scale and matched per-coordinate velocity scale	Corollary 5.3
$\varepsilon, \eta, q, \mathcal{V}$	Momentum energy correction, Young parameter, equivalence constant, Lyapunov energy	Proposition D.1
$\mathcal{N}, \mathfrak{C}_0$	Mean squared particle force and tightened initial energy prefactor	Lemma C.24, Corollary D.4
$\mathfrak{o}(J), \xi_A$	Matrix Lie algebra and infinitesimal congruence field	Appendix F

B Measure Formulation and Wasserstein Conservation Laws

The covariance equations in the main text have the same meaning at both finite and infinite widths. In this section, we introduce the measure flow that captures training dynamics, define a law shared by all widths, and lifts the spectral invariants to functionals of a neuron distribution.

B.1 Neuron distributions and Wasserstein flow

We describe the configuration of any number of neurons by their law: for $w := (u, v) \in \mathbb{R}^m$, let $\mu \in \mathcal{P}_2(\mathbb{R}^m)$ and write its uncentered second moment

$$\Sigma(\mu) := \int ww^\top \, \mathrm{d}\mu(w). \tag{16}$$

For n neurons $W := (w_1, \dots, w_n)$, the neuron law is an empirical distribution $\widehat{\mu}_W := n^{-1} \sum_{i=1}^n \delta_{w_i}$, for which $\Sigma(\widehat{\mu}_W) = WW^\top/n$. A general covariance $\Sigma \succeq 0$ is also realized by a centered Gaussian distribution. The loss functional is $\mathcal{F}(\mu) := \ell(\Sigma^{vu}(\mu))$, so $\mathcal{F}(\widehat{\mu}_W) = \mathcal{L}(U, V)$.

For a functional with a sufficiently regular first variation, the Wasserstein gradient is the spatial gradient $\nabla_{W_2}\mathcal{F}(\mu)(w) := \nabla_w(\delta\mathcal{F}/\delta\mu)(\mu, w)$ and its negative is the velocity in the continuity equation

$$\partial_t \mu_t + \nabla_w \cdot (\mu_t b_t) = 0, \qquad b_t := -\nabla_{W_2}\mathcal{F}(\mu_t). \tag{17}$$

The PDE is understood distributionally, including for empirical measures. In the present split model, $\delta\mathcal{F}/\delta\mu = \frac{1}{2} w^\top G_t w$ up to an additive constant. Consequently,

$$\partial_t \mu_t = \nabla_w \cdot (\mu_t G_t w), \qquad \mu_t = (\Phi_t)_{\#}\mu_0, \tag{18}$$

where Φ_t is the fundamental matrix in (3). For the empirical neuron law, the characteristics of (17) are exactly $\dot{u}_i = -E_t^\top v_i$, $\dot{v}_i = -E_t u_i$.

Conversely, given a solution Σ_t of (2), transporting any initial distribution with second moment Σ_0 by $\dot{\Phi}_t = -G_t\Phi_t$ produces a Wasserstein gradient flow with second moment Σ_t. Indeed, its second moment and Σ_t solve the same linear equation with the same initial data. Matching covariances alone does not ensure that a measure curve follows this gradient flow.

B.2 Covariance-dependent objectives

Moment closure is not specific to the off-diagonal predictor: any sufficiently regular objective depending on second moments has a closed Bures equation.

Proposition B.1 (Second-moment closure)**.** *Let F be a differentiable scalar function on an open neighborhood of $\mathbb{S}_+^m$ in the symmetric matrices, and define $\mathcal{F}(\mu) := F(\Sigma(\mu))$. Let μ_t be a sufficiently regular Wasserstein gradient flow of $\mathcal{F}$, and set $\Sigma_t := \Sigma(\mu_t)$ and $D_t := \nabla_\Sigma F(\Sigma_t)$. Then the characteristics satisfy $\dot{w}_t = -2D_t w_t$, and the second moment evolves autonomously as $\dot{\Sigma}_t = -2(D_t\Sigma_t + \Sigma_t D_t)$. Thus the Wasserstein flow of $\mathcal{F}$ closes exactly on the Bures gradient flow of F on second moments, with the normalization used throughout the paper.*

Proof of Proposition B.1. The first variation of $\mathcal{F}$ at μ_t, up to an additive constant, is $\frac{\delta\mathcal{F}}{\delta\mu}(\mu_t, w) = \langle D_t, ww^\top \rangle = w^\top D_t w$. Taking its Euclidean gradient in w gives $2D_t w$ and hence the claimed characteristic equation. Finally, writing the flow map as $w_0 \mapsto w_t$ and differentiating $\Sigma_t = \int w_t w_t^\top \, \mathrm{d}\mu_0(w_0)$ gives

$$\dot{\Sigma}_t = \int \left(\dot{w}_t w_t^\top + w_t \dot{w}_t^\top\right) \mathrm{d}\mu_0(w_0) = -2(D_t\Sigma_t + \Sigma_t D_t).$$

□

B.3 Width consistency and covariance-level laws

We seek conservation laws defined on neuron distributions that remain conserved when evaluated on finite networks of any width.

Definition B.2 (Mean-field conservation law)**.** *A scalar functional $\mathcal{H}(\mu)$ is a mean-field conservation law for a family of flows if it is constant along every admissible trajectory of every flow in that family, for every admissible initialization. It is covariance-level if $\mathcal{H}(\mu) = h(\Sigma(\mu))$.*

For differentiable functionals, the chain rule expresses conservation as orthogonality:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{H}(\mu_t) = -\int \nabla_{W_2}\mathcal{H}(\mu_t)\cdot\nabla_{W_2}\mathcal{F}(\mu_t)\,\mathrm{d}\mu_t = 0,$$

whenever taking gradients and the chain rule are well-defined. The laws below also have a direct characteristic proof. A single mean-field law $\mathcal{H}$ induces laws $h_n(W) := \mathcal{H}(\widehat{\mu}_W)$ for every width, in the finite-dimensional sense of Marcotte et al. (2023). Unrelated laws found separately at each width need not arise this way. Conversely, conservation at every finite width passes to the measure flow whenever the empirical flows converge in W_2 and $\mathcal{H}$ is continuous in that metric. For covariance-level laws no limiting argument is needed: every positive semidefinite covariance has an empirical realization, and the common closed moment equation gives the equivalence with Bures laws.

B.4 Gram-function laws and their spectral restriction

Preservation of pairwise indefinite products yields functionals that need not depend only on covariance.

Proposition B.3 (Fundamental mean-field laws)**.** *Let $\mu_t = (\Phi_t)_{\#}\mu_0$ be the characteristic flow of $\dot{w}_t = -G_t w_t$, where $G_t^\top = G_t$ and $G_tJ + JG_t = 0$. For every $r \geq 1$ and Borel function $\varphi : \mathbb{S}^r \to \mathbb{R}$, the functional*

$$\mathcal{H}_{\varphi,r}(\mu) := \int \varphi\big((w_a^\top J w_b)_{1\leq a,b\leq r}\big)\prod_{a=1}^{r}\mathrm{d}\mu(w_a) \tag{19}$$

is conserved whenever this integral is absolutely integrable at initialization. In particular, the spectral laws $\mathcal{H}_k(\Sigma(\mu))$ of (2) are conserved for every $\mu_0 \in \mathcal{P}_2(\mathbb{R}^m)$. For particles, the Gram imbalance $\Gamma_t := V_t^\top V_t - U_t^\top U_t = -W_t^\top J W_t$ is constant and

$$\mathcal{H}_k(\Sigma_t) = \mathrm{tr}\big((-\Gamma_t/n)^k\big)\,, \qquad k \geq 1. \tag{20}$$

The nonzero spectra of $J\Sigma_t$ and $-\Gamma_t/n$ coincide.

Proof of Proposition B.3. Along characteristics, $\frac{\mathrm{d}}{\mathrm{d}t}(w_a^\top J w_b) = -w_a^\top(G_tJ + JG_t)w_b = 0$, so (19) is conserved. Bounded Borel tests are integrable, and cyclic polynomial tests need only second moments because each sampled variable appears quadratically. For the cyclic polynomial test function, with $w_{k+1} := w_1$, repeated integration gives

$$\int\prod_{\ell=1}^{k}\big(w_\ell^\top J w_{\ell+1}\big)\prod_{\ell=1}^{k}\mathrm{d}\mu(w_\ell) = \mathrm{tr}\big((J\Sigma)^k\big)\,.$$

Thus (2) is a special case of (19). The first m trace powers determine the characteristic polynomial by Newton identities. For empirical measures, $J\Sigma = (JW)(W^\top/n)$ while $(W^\top/n)(JW) = -\Gamma/n$. The nonzero eigenvalues of AB and BA coincide, which gives the spectral statement and (20). □

C Technical details and proofs

C.1 Loss estimates

Lemma C.1 (Smooth PL consequences)**.** *Under Assumption 1.1, the minimizer set $\mathcal{S}_\star := \{S : \ell(S) = \ell_\star\}$ is nonempty, and*

$$\|\nabla\ell(S)\|_{\mathrm{F}}^2 \leq 2L(\ell(S) - \ell_\star), \qquad \mathrm{dist}(S,\mathcal{S}_\star)^2 \leq 2(\ell(S) - \ell_\star)/\kappa. \tag{21}$$

For every perturbation H,

$$\sqrt{\ell(S+H) - \ell_\star} \leq \sqrt{\ell(S) - \ell_\star} + \sqrt{L/2}\,\|H\|_{\mathrm{F}}\,. \tag{22}$$

Proof of Lemma C.1. Apply the descent lemma at $S - \nabla\ell(S)/L$ for the first inequality. Since $\nabla\ell$ is globally Lipschitz, the Euclidean gradient flow $\dot{S} = -\nabla\ell(S)$ exists for all $t \geq 0$. Writing $h := \ell(S) - \ell_\star$ gives $\dot{h} = -\|\nabla\ell\|^2$ and, where $h > 0$, $-\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{h} \geq \sqrt{\kappa/2}\left\|\dot{S}\right\|$. Thus its total remaining length is at most $\sqrt{2h/\kappa}$. The flow has a limit, with loss gap zero by PL, proving attainment and the distance bound. Finally, the descent lemma and the first inequality bound the perturbed loss gap by $(\sqrt{h} + \sqrt{L/2}\,\|H\|_{\mathrm{F}})^2$. □

Remark C.2 (Quadratic growth and convexity). *Lemma C.1 proves the quadratic-growth bound*

$$\ell(S) - \ell_\star \geq \frac{\lambda}{2}\operatorname{dist}(S, \mathcal{S}_\star)^2 \qquad \textit{with } \lambda = \kappa. \tag{23}$$

If a larger quadratic-growth constant λ is known, it improves the distance bounds while leaving the loss rates, which use κ, unchanged. The PL-to-quadratic-growth implication does not need convexity (Karimi et al., 2016). Conversely, convex quadratic growth with constant λ implies PL with constant $\lambda/4$: for a nearest minimizer $S_\star$, $\Delta \leq \langle\nabla\ell(S), S - S_\star\rangle \leq \|\nabla\ell(S)\|_{\mathrm{F}}\sqrt{2\Delta/\lambda}$. The best PL and quadratic-growth constants need not coincide. For a nonzero quadratic Hessian $\mathcal{Q} \succeq 0$, $\lambda = \kappa = \lambda^+_{\min}(\mathcal{Q})$ is valid and $\mathcal{S}_\star = \Sigma^{vu}_\star + \ker\mathcal{Q}$.

C.2 Covariance closure and conservation laws

Direct matrix differentiation proves closure and conservation at every width. The measure interpretation is given separately in Appendix B.

Proof of Proposition 2.1. Since $\dot{W}_t = -G_t W_t$ and $G_t^\top = G_t$, differentiating $W_t W_t^\top/n$ gives (2). For any covariance solution, define $\dot{\Phi}_t = -G_t\Phi_t$, $\Phi_0 = \mathrm{Id}$. Uniqueness for the resulting linear covariance equation yields $\Sigma_t = \Phi_t\Sigma_0\Phi_t^\top$, so positive semidefiniteness is preserved. The identity $G_t J + J G_t = 0$ gives $\frac{\mathrm{d}}{\mathrm{d}t}(\Phi_t^\top J\Phi_t) = 0$. Invertibility of Φ_t then gives both identities in (3). Equivalently, $\frac{\mathrm{d}}{\mathrm{d}t}(J\Sigma_t) = [G_t, J\Sigma_t]$, so $\frac{\mathrm{d}}{\mathrm{d}t}\operatorname{tr}((J\Sigma_t)^k) = 0$ for every $k \geq 1$. The first m trace powers determine the characteristic polynomial. For later use, the block equations are

$$\begin{aligned} \dot{\Sigma^u}_t &= -\Sigma_t^{uv}E_t - E_t^\top\Sigma_t^{vu}, \qquad \dot{\Sigma^v}_t = -\Sigma_t^{vu}E_t^\top - E_t\Sigma_t^{uv}, \\ \dot{\Sigma^{vu}}_t &= -E_t\Sigma_t^u - \Sigma_t^v E_t. \end{aligned} \tag{24}$$

□

Proof of Proposition 2.3. For any $c \in \mathbb{R}$, the characteristic congruence gives

$$\Sigma_t - cJ = \Phi_t(\Sigma_0 - cJ)\Phi_t^\top.$$

Invertibility of Φ_t preserves the entire feasible interval of c, so both endpoints are conserved even when Σ_0 is singular. Taking principal blocks proves the hidden bounds. □

Spectral formula (5). The feasible sets in Definition 2.2 contain zero, are closed intervals, and are bounded above by the corresponding hidden-block minimum eigenvalues. Thus the maxima exist. For $\Sigma \succ 0$, put $H := \Sigma^{1/2}J\Sigma^{1/2}$. Sylvester's inertia theorem gives d_v negative and d_u positive eigenvalues. Since $H^{-1} = \Sigma^{-1/2}J\Sigma^{-1/2}$, $\Sigma - cJ = \Sigma^{1/2}(\mathrm{Id} - cH^{-1})\Sigma^{1/2}$ gives $a_+(\Sigma) = \lambda_{d_v+1}(H)$ and $a_-(\Sigma) = -\lambda_{d_v}(H)$. For singular Σ, apply this to $\Sigma + \epsilon\mathrm{Id}$. As $\epsilon \downarrow 0$, each margin decreases to a finite limit that is feasible for Σ by closedness, and hence equals its margin. Continuity of the square root and of ordered symmetric eigenvalues proves (5) without deleting any zero eigenvalues. Finally H and $J\Sigma$ have the same characteristic polynomial by the AB, BA identity, even when they are not similar. □

Remark C.3 (Sharpness of the supporting constants). *For every $\Sigma \succeq 0$, testing the supporting inequalities on the two coordinate subspaces gives the upper bounds in*

$$\lambda_{\min}(\Sigma) \leq a_+(\Sigma) \leq \lambda_{\min}(\Sigma^u), \qquad \lambda_{\min}(\Sigma) \leq a_-(\Sigma) \leq \lambda_{\min}(\Sigma^v).$$

The lower bounds follow from $\mathrm{Id} \mp J \succeq 0$. When $\Sigma^{vu} = 0$, the two upper bounds are attained. Consequently $2\lambda_{\min}(\Sigma) \leq g(\Sigma) \leq 2\|\Sigma\|_{\mathrm{op}}$. Both margins are positive exactly when $\Sigma \succ 0$: if both

are positive and $x \in \ker \Sigma$, the supporting inequalities force $x^\top Jx = 0$ and then $(\Sigma - a_+(\Sigma)J)x = 0$, so $x = 0$. The convergence theorem requires only their sum to be positive, so it also covers some singular covariances. Each positive margin makes the corresponding principal block positive definite, giving the rank bounds in Remark 3.2.

C.3 Linear Convergence

We prove the spectral-margin theorem and its quantitative trajectory bounds first. Conditional and determinant-based criteria follow as supplementary comparisons.

Proposition C.4 (Quantitative trajectory bounds)**.** *Under Theorem 3.1, set $a_0 := \lambda_{\min}(\Sigma_0)$, $B_0 := \|\Sigma_0\|_{\mathrm{op}}$, $I_0 := g(\Sigma_0)^{-1}\sqrt{2\Delta_0/\kappa}$, and $B_\infty := B_0 e^{2I_0}$. Then*

$$a_0 e^{-2I_0}\mathrm{Id}_m \preceq \Sigma_t \preceq B_\infty \mathrm{Id}_m, \tag{25}$$

$$\|\Sigma_t - \Sigma_\infty\|_{\mathrm{op}} \le 2B_\infty I_0 e^{-\kappa g(\Sigma_0)t}, \tag{26}$$

$$\|\Sigma_t^{vu} - \Sigma_\infty^{vu}\|_{\mathrm{F}} \le 2B_\infty I_0 e^{-\kappa g(\Sigma_0)t}. \tag{27}$$

For the characteristic lift of any initial law in $\mathcal{P}_2$ with this second moment, there is a limit μ_∞ and

$$W_2(\mu_t, \mu_\infty) \le \min\left\{\sqrt{\operatorname{tr}\Sigma_0}\, e^{I_0}, \sqrt{2B_\infty}\right\} I_0 e^{-\kappa g(\Sigma_0)t}. \tag{28}$$

The same bound applies to $\|W_t - W_\infty\|_{\mathrm{F}}/\sqrt{n}$ for particles. The smallest positive covariance eigenvalue also satisfies $\lambda^+_{\min}(\Sigma_t) \ge e^{-2I_0}\lambda^+_{\min}(\Sigma_0)$, including at $t = \infty$.

Proof of Theorem 3.1 and Proposition C.4. The predictor equation (24) gives the exact dissipation identity on every positive semidefinite trajectory:

$$\frac{\mathrm{d}}{\mathrm{d}t}\ell(\Sigma_t^{vu}) = -\operatorname{tr}(E_t^\top E_t \Sigma_t^u) - \operatorname{tr}(E_t^\top \Sigma_t^v E_t) \le 0. \tag{29}$$

The transported hidden-block bounds from Proposition 2.3 therefore give $-\dot\Delta \ge g(\Sigma_0)\,\|E\|_{\mathrm{F}}^2 \ge 2\kappa g(\Sigma_0)\Delta$. On any finite existence interval, the preceding dissipation and PL give, where the loss gap is positive,

$$-\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{\Delta_t} \ge \frac{g(\Sigma_0)\,\|E_t\|_{\mathrm{F}}^2}{2\sqrt{\Delta_t}} \ge g(\Sigma_0)\sqrt{\kappa/2}\,\|E_t\|_{\mathrm{F}}.$$

At zero loss gap the flow is stationary. Hence, for every finite $T > t$,

$$\int_t^T \|E_r\|_{\mathrm{F}}\,\mathrm{d}r \le \frac{1}{g(\Sigma_0)}\sqrt{\frac{2}{\kappa}}(\sqrt{\Delta_t} - \sqrt{\Delta_T}) \le I_0 e^{-\kappa g(\Sigma_0)t}.$$

Together with smoothness, this gives

$$\|G_t\|_{\mathrm{op}} = \|E_t\|_{\mathrm{op}} \le \sqrt{2L\Delta_0}e^{-\kappa g(\Sigma_0)t}, \qquad \int_t^T \|G_r\|_{\mathrm{op}}\,\mathrm{d}r \le I_0 e^{-\kappa g(\Sigma_0)t}. \tag{30}$$

Both Φ_t and Φ_t^{-1} have norm at most e^{I_0}, even when Σ_0 is singular. Their congruence representation proves (25) before the maximal existence time. The covariance vector field is locally Lipschitz on all symmetric matrices, and the upper bound prevents finite-time blow-up; hence the solution is global. Integrating $\left\|\dot\Sigma_t\right\|_{\mathrm{op}} \le 2B_\infty \|G_t\|_{\mathrm{op}}$ proves the Cauchy property and (26). Integrability of the force also makes Φ_t Cauchy; its limit is invertible since $\sigma_{\min}(\Phi_t) \ge e^{-I_0}$. Thus $\Sigma_\infty = \Phi_\infty \Sigma_0 \Phi_\infty^\top$ has the same rank as Σ_0 and is optimal by (6). The singular-value bound for congruence gives the claimed positive eigenvalue floor, without requiring full rank. Directly in predictor Frobenius norm, $\left\|\dot{\Sigma^{vu}}_t\right\|_{\mathrm{F}} \le 2B_\infty \|E_t\|_{\mathrm{F}}$; its integral proves (27) without a dimension factor. Finally $\|\Phi_t - \Phi_\infty\|_{\mathrm{op}} \le e^{I_0} I_0 e^{-\kappa g(\Sigma_0)t}$. Coupling the two characteristic images of the same initial particle gives the first bound in (28). For the second, use

$$\left(\int \|G_t w\|^2\,\mathrm{d}\mu_t(w)\right)^{1/2} = \sqrt{\operatorname{tr}(G_t\Sigma_t G_t)} \le \sqrt{2B_\infty}\,\|E_t\|_{\mathrm{F}},$$

and integrate along the same characteristics. The identical estimate holds for empirical factors after division by $\sqrt{n}$. If $\Delta_0 = 0$, the first-order flow is stationary and the same formulas hold with $I_0 = 0$. □

Alternative nondegeneracy criteria The same argument can use any independently established hidden-block lower bounds. The common-bound formulation below uses $2c$ in place of $g(\Sigma_0)$. For $\Sigma_0 \succ 0$, taking $c := \min(a_+(\Sigma_0), a_-(\Sigma_0))$ recovers a possibly weaker rate than Theorem 3.1.

Lemma C.5 (Conditional hidden-block criterion)**.** *Under Assumption 1.1, start (2) from $\Sigma_0 \succ 0$. If $\Sigma_t^u \succeq c\mathrm{Id}$ and $\Sigma_t^v \succeq c\mathrm{Id}$ for some $c > 0$ throughout its maximal existence interval, the flow is global and*

$$\Delta_t \le \Delta_0 e^{-4\kappa c t}. \tag{31}$$

Moreover Σ_t converges to $\Sigma_\infty \succ 0$ with $\Sigma_\infty^{vu} \in \mathcal{S}_\star$.

Quadratic growth further gives $\operatorname{dist}(\Sigma_t^{vu}, \mathcal{S}_\star)^2 \le 2\Delta_0 e^{-4\kappa c t}/\lambda$, where $\lambda = \kappa$ is always available.

Proof of Lemma C.5. The assumed hidden bounds give $-\dot\Delta \ge 2c\,\|E\|_{\mathrm{F}}^2 \ge 4c\kappa\Delta$. On every positive loss gap interval, differentiating $\sqrt{\Delta_t}$ gives

$$-\frac{\mathrm{d}}{\mathrm{d}t}\sqrt{\Delta_t} \ge c\sqrt{2\kappa}\,\|E_t\|_{\mathrm{F}}, \qquad \int_t^T \|G_r\|_{\mathrm{op}}\,\mathrm{d}r \le \frac{\sqrt{\Delta_t} - \sqrt{\Delta_T}}{c\sqrt{2\kappa}}.$$

The same bound holds through zero loss gap by stationarity and is valid before the maximal existence time. Both characteristic matrices Φ_t and Φ_t^{-1} are bounded by the exponential of this integral. Their covariance congruence gives

$$\|\Sigma_t\|_{\mathrm{op}} \le \|\Sigma_0\|_{\mathrm{op}} \exp\left(\frac{1}{c}\sqrt{\frac{2\Delta_0}{\kappa}}\right), \qquad \lambda_{\min}(\Sigma_t) \ge \lambda_{\min}(\Sigma_0) \exp\left(-\frac{1}{c}\sqrt{\frac{2\Delta_0}{\kappa}}\right). \tag{32}$$

These bounds rule out finite-time blow-up. The tail of the integral proves convergence to a positive-definite covariance with optimal predictor. Quadratic growth gives the distance bound. □

For coercive losses, determinant conservation and the second spectral law yield an alternative, generally much coarser, source of such bounds.

Proposition C.6 (Coercivity prevents covariance collapse)**.** *Under Assumption 1.1, suppose ℓ is coercive and $\Sigma_0 \succ 0$. Let $M := \sup\{\|S\|_{\mathrm{F}} : \ell(S) \le \ell(\Sigma_0^{vu})\}$. On its maximal existence interval the Bures flow satisfies $c\mathrm{Id}_m \preceq \Sigma_t \preceq B\mathrm{Id}_m$, where $B := \sqrt{\operatorname{tr}((J\Sigma_0)^2) + 2M^2} + M$ and $c := \det\Sigma_0 / B^{m-1} > 0$. In particular both hidden blocks are bounded below by $c\mathrm{Id}$.*

Proof of Proposition C.6. The characteristic congruence (3) preserves positive definiteness on the maximal existence interval. Loss dissipation gives $\|\Sigma_t^{vu}\|_{\mathrm{F}} \le M$. Since $\operatorname{tr} G_t = 0$, Jacobi's formula yields $\det\Sigma_t = \det\Sigma_0 > 0$. The second spectral law gives

$$\|\Sigma_t^u\|_{\mathrm{F}}^2 + \|\Sigma_t^v\|_{\mathrm{F}}^2 = \operatorname{tr}((J\Sigma_0)^2) + 2\,\|\Sigma_t^{vu}\|_{\mathrm{F}}^2 \le \operatorname{tr}((J\Sigma_0)^2) + 2M^2.$$

Split Σ_t into its block-diagonal and off-diagonal parts. Their operator norms are bounded by the square root of the last expression and by $\|\Sigma_t^{vu}\|_{\mathrm{op}} \le M$, respectively. Thus $\|\Sigma_t\|_{\mathrm{op}} \le B$. Finally, $\det\Sigma_0 \le \lambda_{\min}(\Sigma_t) B^{m-1}$ gives $\Sigma_t \succeq c\mathrm{Id}_m$, and the same lower bound holds on both principal blocks. □

Remark C.7 (Coarse ellipticity versus the spectral-margin rate)**.** *At $\Sigma_0 = \alpha\mathrm{Id}_m$, the determinant argument gives*

$$c_{\mathrm{coarse}} := \alpha\left(\frac{\alpha}{\sqrt{m\alpha^2 + 2M^2} + M}\right)^{m-1}.$$

Its rate $4\kappa c_{\mathrm{coarse}}$ can be extremely small in dimension, whereas the support gap $g(\Sigma_0) = 2\alpha$ gives $4\alpha\kappa$ directly. For general initial covariances, the margin argument likewise gives a rate determined by the signed spectrum, with no additional dimension factor. Trace and initial loss alone cannot provide a uniform positive rate as the support gap tends to zero (Example C.9).

Remark C.8 (Degenerate boundary leaves)**.** *In the 2×2 covariance setting of Section 2, write $\Sigma = \left(\begin{smallmatrix} p & r \\ r & q \end{smallmatrix}\right)$, $b_0 := (p_0 - q_0)/2$, and $\alpha^2 := p_0 q_0 - r_0^2$, with predictor loss $\ell(r) := (r-\beta)^2/2$. On the boundary $b_0 = \alpha = 0$, the scalar equation is $\dot r = -2|r|(r-\beta)$. If $\beta \ne 0$, the ray with $r_0\beta > 0$ converges to β, whereas $r_0\beta < 0$ converges to the nonoptimal apex $r = 0$; the apex itself is stationary. For example, when $\beta = 1$ and $p_0 = q_0 = 1$, $r_0 = -1$, one has $p_t = q_t = -r_t = (2e^{2t} - 1)^{-1}$.*

When $\beta = 0$, one has $r_t = r_0/(1 + 2|r_0|t)$: convergence to the optimal apex is algebraic for $r_0 \neq 0$. These boundary cases have $g(\Sigma_0) = 0$ and are not covered by the positive-support-gap convergence theorem.

Example C.9 (Fixed trace and predictor, arbitrarily slow escape)**.** *Let $d_u = d_v = 2$, $0 < \delta < 1$, and*

$$\Sigma_0 = \operatorname{diag}(\delta, 2 - \delta, \delta, 2 - \delta), \qquad \ell(S) := \tfrac{1}{2} \left\| S - e_1 e_1^\top \right\|_{\mathrm{F}}^2 .$$

Here $L = \kappa = 1$, $\operatorname{tr} \Sigma_0 = 4$, $\Sigma_0^{vu} = 0$, and $\Delta_0 = 1/2$, independently of δ. The active scalar block evolves by

$$r' = 2\sqrt{r^2 + \delta^2}(1 - r), \qquad r(0) = 0.$$

The instantaneous relative decay of the loss at zero is 4δ. Until $r = 1/2$, $r' \leq 2(r + \delta)$, so $r(t) \leq \delta(e^{2t} - 1)$. The time to reach $1/2$ is at least $\frac{1}{2}\log(1 + 1/(2\delta))$, which diverges as $\delta \downarrow 0$. There can therefore be no uniform positive exponential rate with a fixed finite prefactor determined only by the displayed trace, initial loss gap, L, and κ. At $\delta = 0$ the active block is stationary and nonoptimal. Theorem 3.1 records exactly the missing scale: $g(\Sigma_0) = 2\delta$. The slow escape occurs even for a strongly convex loss with a unique predictor minimizer.

C.4 Finite Width Convergence

Write $r_0 := \operatorname{rank} \widetilde{\Sigma}_0$ for the population covariance rank.

Theorem C.10 (Relative initialization stability)**.** *Under Assumption 1.1, let $\widetilde{\Sigma}_0 \succeq 0$ with $g(\widetilde{\Sigma}_0) > 0$, and put $\widetilde{B}_0 := \left\| \widetilde{\Sigma}_0 \right\|_{\mathrm{op}}$ and $\widetilde{\Delta}_0 := \ell(\widetilde{\Sigma}_0^{vu}) - \ell_\star$. If an empirical covariance satisfies*

$$(1 - \varepsilon)\widetilde{\Sigma}_0 \preceq \Sigma_0 \preceq (1 + \varepsilon)\widetilde{\Sigma}_0, \qquad 0 < \varepsilon < 1, \tag{33}$$

then its first-order factor flow is global, converges to an optimal predictor, and its loss gap $\Delta_t = \ell(\Sigma_t^{vu}) - \ell_\star$ satisfies

$$\Delta_t \leq \Delta_0 e^{-2(1-\varepsilon)\kappa g(\widetilde{\Sigma}_0)t}. \tag{34}$$

Only the lower inequality in (33) is needed for this claim. With both inequalities, $d_\wedge := \min(d_u, d_v)$ gives the deterministic prefactor bound

$$\Delta_0 \leq \left(\sqrt{\widetilde{\Delta}_0} + \varepsilon \widetilde{B}_0 \sqrt{L d_\wedge / 2} \right)^2 . \tag{35}$$

For the sampling law in Corollary 3.4, the width condition (7) guarantees (33) with probability at least $1 - \delta$.

Proof of Theorem C.10. From $\widetilde{\Sigma}_0 - a_+(\widetilde{\Sigma}_0)J \succeq 0$ and the lower relative inequality, $\Sigma_0 - (1 - \varepsilon)a_+(\widetilde{\Sigma}_0)J \succeq 0$, and similarly for $a_-(\widetilde{\Sigma}_0)$. Transporting these inequalities along the empirical fundamental matrix allows us to apply the proof of Theorem 3.1 with margin sum $(1 - \varepsilon)g(\widetilde{\Sigma}_0)$ and initial empirical loss gap. The two-sided relative comparison gives $\left\| \Sigma_0 - \widetilde{\Sigma}_0 \right\|_{\mathrm{op}} \leq \varepsilon \widetilde{B}_0$ and hence $\left\| \Sigma_0^{vu} - \widetilde{\Sigma}_0^{vu} \right\|_{\mathrm{F}} \leq \sqrt{d_\wedge}\varepsilon \widetilde{B}_0$. Equation (35) follows from (22). For the probability bound, choose an orthonormal basis $T \in \mathbb{R}^{m \times r_0}$ of the range of $\widetilde{\Sigma}_0$ and write $\widetilde{\Sigma}_0 = T\Lambda T^\top$, where $\Lambda \succ 0$. The vectors $\eta_i := T^\top \xi_i \in \mathbb{R}^{r_0}$ have identity second moment and subgaussian norm at most K, and $w_i = T\Lambda^{1/2}\eta_i$. The covariance tail bound (Vershynin, 2018, Exercise 4.7.3) applied in this range gives

$$\left\| n^{-1} \textstyle\sum_i \eta_i \eta_i^\top - \mathrm{Id}_{r_0} \right\|_{\mathrm{op}} \leq CK^2 \left(\sqrt{\frac{r_0 + u}{n}} + \frac{r_0 + u}{n} \right)$$

with failure probability at most $2e^{-u}$. Set $u := \log(2/\delta)$ and adjust C in (7). Multiplication on the left by $T\Lambda^{1/2}$ and on the right by its transpose gives (33). □

Proof of Corollary 3.4. Theorem C.10 gives the relative covariance event and the loss rate. Its lower bound gives $g(\Sigma_0) \geq (1 - \varepsilon)g(\widetilde{\Sigma}_0) > 0$, so Theorem 3.1 gives global existence and factor convergence. The two-sided event preserves the kernel of $\widetilde{\Sigma}_0$, so the empirical initial covariance is positive definite exactly when $\widetilde{\Sigma}_0$ is. □

Small width for full-white sampling The rank improvement in Remark 3.5 concerns the chosen population law. A positive population support gap alone does not give a small-width support gap guarantee for full-white samples.

Example C.11 (Full-white sampling can lose the support gap)**.** *Let $d_u = D$, $d_v = 1$, and sample from $N(0, \widetilde{\Sigma}_0)$ with $\widetilde{\Sigma}_0 = \sigma^2 \mathrm{Id}_{D+1}$, $\sigma > 0$. There is a universal $c > 0$ such that, if $1 \le n \le D/16$, then $\mathbb{P}\{g(\Sigma_0) = 0\} \ge 1 - 4e^{-cD}$, although $g(\widetilde{\Sigma}_0) = 2\sigma^2$. Thus a width bound depending only on $\min(d_u, d_v)$ cannot uniformly preserve a positive support gap for full-white sampling. Convergence of these trajectories is not decided by this example.*

Proof. Write $U_0 = \sigma X$ and $V_0 = \sigma y^\top$, where $X \in \mathbb{R}^{D\times n}$ and $y \in \mathbb{R}^n$ have independent standard Gaussian entries. Gaussian singular-value and norm concentration (Vershynin, 2018, Section 7.3), with deviation $\sqrt{D}/8$, gives

$$\sigma_{\min}(X) \ge \sqrt{D} - \sqrt{n} - \sqrt{D}/8 \ge 5\sqrt{D}/8, \qquad \|y\| \le \sqrt{n} + \sqrt{D}/8 \le 3\sqrt{D}/8$$

with probability at least $1 - 4e^{-cD}$, for a universal $c > 0$. Hence $X^\top X - yy^\top \succeq (D/4)\mathrm{Id}_n$. The nonzero spectrum of $J\Sigma_0$ is the spectrum of $\sigma^2(X^\top X - yy^\top)/n$, so it has n positive eigenvalues and $D + 1 - n \ge 2$ zero eigenvalues. In (5), $d_v = 1$ therefore gives $g(\Sigma_0) = \lambda_2 - \lambda_1 = 0$. □

A separate initial-accuracy budget A relative covariance bound preserves the rate; an absolute bound also controls the initial loss. In the sampling setting of Corollary 3.4, write $\widetilde{B}_0 := \left\|\widetilde{\Sigma}_0\right\|_{\mathrm{op}}$, $\widetilde{\Delta}_0 := \ell(\widetilde{\Sigma}_0^{vu}) - \ell_\star$, and $d_\wedge := \min(d_u, d_v)$. Besides $\varepsilon, \delta \in (0,1)$, choose an absolute tolerance $\zeta > 0$ and set

$$\beta := \min\{\varepsilon, \zeta/\widetilde{B}_0\}, \qquad \overline{\Delta} := \left(\sqrt{\widetilde{\Delta}_0} + \zeta\sqrt{Ld_\wedge/2}\right)^2. \tag{36}$$

Theorem C.12 (High-probability finite-width convergence)**.** *Under Assumption 1.1, let $\widetilde{\Sigma}_0 \succeq 0$ with $g(\widetilde{\Sigma}_0) > 0$ and use the sampling law and tolerances above. There is a universal $C > 0$ such that, if*

$$n \ge CK^4\beta^{-2}\big(r_0 + \log(2/\delta)\big), \tag{37}$$

then with probability at least $1 - \delta$ the empirical covariance has positive support gap, the factor flow is global, and

$$\Delta_t \le \Delta_0 e^{-2\kappa(1-\varepsilon)g(\widetilde{\Sigma}_0)t} \le \overline{\Delta} e^{-2\kappa(1-\varepsilon)g(\widetilde{\Sigma}_0)t}.$$

The factors converge, their limiting predictor belongs to $\mathcal{S}_\star$, and $\mathrm{dist}(\Sigma_t^{vu}, \mathcal{S}_\star)^2 \le 2\Delta_t/\kappa$.

Proof of Theorem C.12. Apply the relative concentration event with tolerance β. It implies the relative bounds with tolerance ε and $\left\|\Sigma_0 - \widetilde{\Sigma}_0\right\|_{\mathrm{op}} \le \beta\widetilde{B}_0 \le \zeta$. The two initial supporting inequalities give the margin bounds, and Lemma C.1 gives the loss gap bound. More precisely, putting $\underline{g} := (1-\varepsilon)g(\widetilde{\Sigma}_0)$ and $\overline{B}_0 := (1+\varepsilon)\widetilde{B}_0$, the same event gives

$$a_\pm(\Sigma_0) \ge (1-\varepsilon)a_\pm(\widetilde{\Sigma}_0), \quad \|\Sigma_0\|_{\mathrm{op}} \le \overline{B}_0, \quad \Delta_0 \le \overline{\Delta}. \tag{38}$$

Theorem 3.1 then applies to the empirical covariance. The distance estimate follows from Lemma C.1. More generally, any valid quadratic-growth constant λ gives $\mathrm{dist}(\Sigma_t^{vu}, \mathcal{S}_\star)^2 \le 2\Delta_t/\lambda$. □

Proposition C.13 (The Gaussian initial-error cost)**.** *For independent $u_i \sim N(0, \sigma^2\mathrm{Id}_{d_u})$ and $v_i \sim N(0, \sigma^2\mathrm{Id}_{d_v})$, and the quadratic loss $\ell(S) := \|S - \Sigma_\star^{vu}\|_{\mathrm{F}}^2/2$, one has exactly*

$$\mathbb{E}\Delta_0 = \tfrac{1}{2}\|\Sigma_\star^{vu}\|_{\mathrm{F}}^2 + \frac{\sigma^4 d_u d_v}{2n}. \tag{39}$$

For balanced dimensions and fixed σ, width n proportional to m therefore gives an expected initial sampling error of order m, despite the dimension-free exponent $4(1-\varepsilon)\sigma^2\kappa$ at any fixed relative tolerance $\varepsilon \in (0,1)$. Keeping that error bounded in expectation requires n of order $d_u d_v$, when the teacher norm is fixed. Thus width proportional to m can preserve the convergence rate while allowing the expected initial error to grow with dimension.

Proof of Proposition C.13. Each entry of $n^{-1}\sum_i v_i u_i^\top$ has mean zero and variance σ^4/n. Sum these variances; the cross term with the teacher has zero expectation. □

C.5 Discrete-Time Convergence

The factor update transports the initial supporting inequalities even though it no longer preserves J exactly. To track this error, define

$$\Phi_0 := \mathrm{Id}, \quad \Phi_{k+1} := M_k \Phi_k, \quad \mathcal{J}_k := \Phi_k J \Phi_k^\top, \quad \mathcal{D}_k := \mathcal{J}_k - J. \tag{40}$$

The cross block of (9) also gives

$$\Sigma_{k+1}^{vu} - \Sigma_k^{vu} = -\tau(E_k \Sigma_k^u + \Sigma_k^v E_k) + \tau^2 E_k (\Sigma_k^{vu})^\top E_k. \tag{41}$$

The next two lemmas control the accumulated defect $\mathcal{D}_k$ and the quadratic correction in the predictor update.

Lemma C.14 (Transported support and accumulated defect)**.** *If $\Sigma_0 \succeq 0$, with initial margins $a_+(\Sigma_0), a_-(\Sigma_0)$ as in (4), then every finite iterate satisfies*

$$\Sigma_k - a_+(\Sigma_0)\mathcal{J}_k \succeq 0, \qquad \Sigma_k + a_-(\Sigma_0)\mathcal{J}_k \succeq 0.$$

If $\|\mathcal{D}_k\|_{\mathrm{op}} \le \eta$ for some $\eta \in [0,1)$, then $\Sigma_k^u \succeq (1-\eta)a_+(\Sigma_0)\mathrm{Id}$ and $\Sigma_k^v \succeq (1-\eta)a_-(\Sigma_0)\mathrm{Id}$. With $e_k := \|E_k\|_{\mathrm{F}}$,

$$\mathcal{D}_{k+1} = M_k \mathcal{D}_k M_k + \tau^2 G_k J G_k, \qquad \|\mathcal{D}_N\|_{\mathrm{op}} \le \tau \exp\left(2 \sum_{j<N} \tau e_j\right) \sum_{j<N} \tau e_j^2. \tag{42}$$

Proof of Lemma C.14. The proof of Proposition 2.3 gives the initial supporting inequalities. Congruence by Φ_k transports them, regardless of invertibility. The uu and vv blocks of $\mathcal{J}_k$ are respectively at least $(1-\eta)\mathrm{Id}$ and at most $-(1-\eta)\mathrm{Id}$, proving the hidden lower bounds. Since $G_k J + J G_k = 0$, one has $M_k J M_k = J + \tau^2 G_k J G_k$, giving the recursion. Its norm is at most $(1+\tau e_k)^2 \|\mathcal{D}_k\|_{\mathrm{op}} + \tau^2 e_k^2$. Unroll from zero and bound the amplification products by $\exp(2\sum_{j<N} \tau e_j)$. □

Descent and length. The bootstrap needs two estimates: loss descent and a finite total gradient length. The next lemma obtains both from the same step condition. For the proof, abbreviate the descent scale, initial gradient bound, and covariance bound as

$$\alpha_\star := \frac{g(\Sigma_0)}{2\sqrt{2}}, \qquad M_\nabla := \sqrt{2L\Delta_0}, \qquad B_\star := \|\Sigma_0\|_{\mathrm{op}} e^{2I_\star}. \tag{43}$$

Then $I_\star = \sqrt{2\Delta_0/\kappa}/\alpha_\star$ and $q = 1 - 2\alpha_\star \kappa \tau$. Condition (10) is equivalent to requiring both $\tau \le \alpha_\star/(B_\star M_\nabla + \frac{9}{2} L B_\star^2)$ and $\tau \le (1 - 1/\sqrt{2})\alpha_\star/(e^{2I_\star}\Delta_0)$.

Lemma C.15 (Descent with unequal hidden lower bounds)**.** *Suppose $\|\Sigma_k\|_{\mathrm{op}} \le B$, $\Delta_k \le \Delta_0$, $\Sigma_k^u \succeq b_+ \mathrm{Id}$, and $\Sigma_k^v \succeq b_- \mathrm{Id}$. Write $h := b_+ + b_- > c > 0$. If $\tau M_\nabla \le 1$ and $\tau(B M_\nabla + \frac{9}{2} L B^2) \le h - c$, then*

$$\Delta_{k+1} \le \Delta_k - c\tau e_k^2 \le (1 - 2c\kappa\tau)\Delta_k. \tag{44}$$

If this descent holds for $j = r, \dots, N-1$, then

$$\sum_{j=r}^{N-1} \tau e_j^2 \le \Delta_r / c, \qquad \sum_{j=r}^{N-1} \tau e_j \le \frac{1}{c}\sqrt{\frac{2}{\kappa}}(\sqrt{\Delta_r} - \sqrt{\Delta_N}). \tag{45}$$

Proof of Lemma C.15. Let $Z := \Sigma_{k+1}^{vu} - \Sigma_k^{vu}$. By (9) and smoothness, $\|Z\|_{\mathrm{F}} \le 3\tau B e_k$ and $\langle E_k, Z \rangle \le -\tau h e_k^2 + \tau^2 B M_\nabla e_k^2$. The descent lemma therefore gives the first assertion. Telescope it for the squared-gradient sum. At positive loss gap, PL gives

$$\sqrt{\Delta_j} - \sqrt{\Delta_{j+1}} \ge \frac{c\tau e_j^2}{2\sqrt{\Delta_j}} \ge c\tau\sqrt{\kappa/2}\, e_j.$$

At zero loss gap $e_j = 0$. Summation proves the length bound. □

The proof below also establishes the following explicit tails in Theorem 3.6:

$$\frac{\|W_k - W_\infty\|_{\mathrm{F}}}{\sqrt{n}} \le \sqrt{2B_\star}\, I_\star q^{k/2}, \qquad \|\Sigma_k - \Sigma_\infty\|_{\mathrm{op}} \le 2B_\star I_\star q^{k/2}. \tag{46}$$

Proof of Theorem 3.6. Let $\eta := 1 - 1/\sqrt{2}$. Since $g(\Sigma_0) \le 2B_0$ for $B_0 := \|\Sigma_0\|_{\mathrm{op}}$, one has $\alpha_\star \le B_0/\sqrt{2} \le B_\star/\sqrt{2}$. The first restriction in (10) implies

$$\tau M_\nabla < 1/\sqrt{2}, \qquad 0 < 2\alpha_\star \kappa \tau \le \frac{4\kappa\alpha_\star^2}{9LB_\star^2} \le \frac{2}{9}.$$

In particular $q \in [7/9, 1)$. Inductively suppose, through iterate N, that $\Delta_k \le q^k \Delta_0$, $\|\Sigma_k\|_{\mathrm{op}} \le B_\star$, and $\|\mathcal{D}_k\|_{\mathrm{op}} \le \eta$, with descent proved for $k < N$. These assertions start at $N = 0$ since $\mathcal{D}_0 = 0$. At N, Lemma C.14 supplies hidden lower bounds with sum $(1-\eta)g(\Sigma_0) = 2\alpha_\star$. Lemma C.15 applies with $c = \alpha_\star$, $h = 2\alpha_\star$, and $B = B_\star$, proving the next loss-gap estimate. Its finite-horizon sums now give

$$\sum_{j=0}^{N} \tau e_j \le I_\star, \qquad \sum_{j=0}^{N} \tau e_j^2 \le \Delta_0/\alpha_\star.$$

Thus $\|\Phi_{N+1}\|_{\mathrm{op}} \le e^{I_\star}$ implies the next covariance upper bound, and (42) gives $\|\mathcal{D}_{N+1}\|_{\mathrm{op}} \le \tau e^{2I_\star}\Delta_0/\alpha_\star \le \eta$. This closes the induction.

Taking tails in (45) yields $\sum_{j\ge k} \tau e_j \le I_\star q^{k/2}$. Factor increments have norm

$$\|W_{j+1} - W_j\|_{\mathrm{F}} / \sqrt{n} = \tau \sqrt{\mathrm{tr}(G_j \Sigma_j G_j)} \le \tau \sqrt{2B_\star}\, e_j.$$

This proves factor convergence and its tail estimate. Similarly $\|\Phi_k - \Phi_\infty\|_{\mathrm{op}} \le e^{I_\star} I_\star q^{k/2}$ gives the covariance tail bound by subtracting the two congruences. The limit predictor is optimal by continuity and the loss-gap bound. Finally, if $r := \tau M_\nabla < 1$, every factor update is invertible and

$$\sigma_{\min}(\Phi_k) \ge \exp\left(-\frac{I_\star}{1-r}\right), \qquad \lambda_{\min}^{+}(\Sigma_k) \ge \lambda_{\min}^{+}(\Sigma_0) \exp\left(-\frac{2I_\star}{1-r}\right).$$

Indeed $\sigma_{\min}(M_j) \ge 1 - \tau e_j > 0$ and $\log(1-x) \ge -x/(1-r)$ for $0 \le x \le r$. The bound makes Φ_∞ invertible, proving rank preservation also at the limit. □

The factor-increment estimate also applies to population discrete pushforward updates. Coupling each particle with its limiting position gives a Wasserstein tail at most $\sqrt{2B_\star} I_\star q^{k/2}$.

White initialization and local stability For exact white empirical initialization $\Sigma_0 = W_0 W_0^\top/n = \sigma^2 \mathrm{Id}_m$ and $\Delta_0 = \ell(0) - \ell_\star > 0$, $I_\star = 2\sqrt{\Delta_0}/(\sigma^2\sqrt{\kappa})$ and the explicit condition is

$$\tau \le \min\left\{ \frac{1}{2e^{2I_\star}\sqrt{L\Delta_0} + \frac{9}{\sqrt{2}} L\sigma^2 e^{4I_\star}}, \quad \frac{(\sqrt{2}-1)\sigma^2}{2e^{2I_\star}\Delta_0} \right\}.$$

The contraction factor is $1 - \sqrt{2}\kappa\sigma^2\tau$. Here the empirical covariance is exactly $\sigma^2\mathrm{Id}_m$. For iid Gaussian initialization, its fluctuations are handled by Corollary C.18. As $\Delta_0 \downarrow 0$ with σ, L, κ fixed, the sufficient threshold tends to $\sqrt{2}/(9L\sigma^2)$.

Remark C.16 (Local stability and the global stepsize)**.** *Suppose ℓ is C^2 near a minimizer at zero, and linearize at $\Sigma = \sigma^2\mathrm{Id}$. The predictor perturbation evolves to first order by $\delta\Sigma^{vu} \mapsto (\mathrm{Id} - 2\tau\sigma^2\nabla^2\ell(0))[\delta\Sigma^{vu}]$. Thus $0 < \tau < 1/(L\sigma^2)$ contracts its nonzero-curvature directions; kernel directions are neutral at this order. Theorem 3.6 additionally controls covariance growth and the accumulated signature defect along the entire trajectory, explaining its more restrictive global threshold.*

Adjustable rate budgets. Smaller steps permit a smaller loss in the conserved conditioning, bringing the discrete exponent closer to its continuous-time value.

Proposition C.17 (Adjustable rate budgets)**.** *Under the loss and initialization assumptions of Theorem 3.6, choose $0 < c < h < g(\Sigma_0)$ and set $I_c := \sqrt{2\Delta_0/\kappa}/c$, $B_c := B_0 e^{2I_c}$. The convergence and rank conclusions of Theorem 3.6 hold with contraction factor $q_c := 1 - 2c\kappa\tau$, and the tail bounds (46) hold with $(I_\star, B_\star, q)$ replaced by (I_c, B_c, q_c), if*

$$0 < \tau \le \min\left\{ \frac{1}{2M_\nabla}, \quad \frac{h-c}{B_c M_\nabla + \frac{9}{2} L B_c^2}, \quad \frac{(1 - h/g(\Sigma_0))c}{e^{2I_c}\Delta_0} \right\}.$$

Taking $c = (1-\epsilon)g(\Sigma_0)$, $h = (1-\epsilon/2)g(\Sigma_0)$ for any $\epsilon \in (0,1)$ certifies exponent $2\kappa(1-\epsilon)g(\Sigma_0)$ in time $k\tau$, at an explicit sufficiently small step. This is arbitrarily close to the flow exponent.

Proof of Proposition C.17. Use $\eta = 1 - h/g(\Sigma_0)$ in the preceding induction and Lemma C.15 with c, h. The positivity guard follows from $c(h-c) \le g(\Sigma_0)^2/4 \le B_0^2$: $2c\kappa\tau \le 4/9$. The remaining estimates use I_c, B_c. □

Corollary C.18 (A stepsize fixed before sampling)**.** *Use the population parameters and width condition of Theorem C.12, with the bounds in (38), and define before sampling*

$$\overline{\alpha}_\star := \frac{g}{2\sqrt{2}}, \quad \overline{M}_\nabla := \sqrt{2L\overline{\Delta}}, \quad \overline{I}_\star := \frac{1}{\overline{\alpha}_\star}\sqrt{\frac{2\overline{\Delta}}{\kappa}}, \quad \overline{B}_\star := \overline{B}_0 e^{2\overline{I}_\star}.$$

Choose a deterministic positive stepsize satisfying

$$\tau \le \min\left\{ \frac{\overline{\alpha}_\star}{\overline{B}_\star \overline{M}_\nabla + \frac{9}{2}L\overline{B}_\star^2}, \quad \frac{(1-1/\sqrt{2})\overline{\alpha}_\star}{e^{2\overline{I}_\star}\overline{\Delta}} \right\}. \tag{47}$$

With probability at least $1-\delta$, factor GD converges to an optimal predictor, and

$$\Delta_k \le (1 - 2\overline{\alpha}_\star\kappa\tau)^k \Delta_0 \le (1 - 2\overline{\alpha}_\star\kappa\tau)^k \overline{\Delta}.$$

Its factor/covariance tail estimates hold with the corresponding barred constants and contraction factor.

At fixed relative tolerance, the sampled stepsize can deteriorate with the dimension through $\overline{\Delta}$. The independent absolute accuracy budget controls this initial-loss contribution.

Proof of Corollary C.18. On (38), the empirical descent scale is at least $\overline{\alpha}_\star$, whereas the empirical loss gap, force bound, length bound, and covariance upper bound are at most their barred counterparts. Both admissible stepsize terms increase with the descent scale and decrease with the loss gap and initial covariance norm. Hence the chosen step satisfies Theorem 3.6 for the empirical initialization. Its contraction factor is no larger than the stated one. If the realized initial loss gap vanishes, the empirical factors are stationary, and the conclusions hold directly. □

Boundary cases. For the scalar loss $\ell(s) := (s+1)^2/2$, the singular initialization $U_0 = V_0 \neq 0$ preserves equality of the factors and keeps $\Sigma_k^{vu} = \|U_k\|^2/n \ge 0$, so its loss gap is at least $1/2$. Conversely, even white initialization does not justify arbitrary steps: for $\ell(s) := (s-1)^2/2$, $\Sigma_0 = \mathrm{Id}_2$, and $\tau = 1$, one gets $\Sigma_1 = \left(\begin{smallmatrix} 2 & 2 \\ 2 & 2 \end{smallmatrix}\right)$ and $\Sigma_2 = 0$, after which training is trapped.

C.6 Deep Linear Residual Networks

Write $P_{b:a} := A_b \cdots A_a$ when $a \le b$, and $P_{b:a} := \mathrm{Id}$ otherwise. Thus $P_{>s} = P_{S-1:s+1}$ and $P_{<s} = P_{s-1:0}$. Differentiating $P_S = P_{>s} A_s P_{<s}$ gives

$$\nabla_{\Sigma_s^{vu}} \ell(P_S) = h\widetilde{E}_s, \qquad \widetilde{E}_s = P_{>s}^\top E P_{<s}^\top. \tag{48}$$

In residual mean-field time the particle equations are $(\dot{u}_{s,i}, \dot{v}_{s,i}) = -(\widetilde{E}_s^\top v_{s,i}, \widetilde{E}_s u_{s,i})$. The cross block of (11) is therefore $\dot{\Sigma^{vu}}_s = -\widetilde{E}_s \Sigma_s^u - \Sigma_s^v \widetilde{E}_s$.

In the finite-depth setting of Theorem 4.1, put $\overline{\sigma}_S := (1 + B/S)^{S-1}$ and $I_{\mathrm{res}} := \overline{\sigma}_S(\overline{g}_S \underline{\sigma}_S^2)^{-1}\sqrt{2\Delta_0/\kappa}$. The proof gives the common layer-force tail

$$\int_t^\infty \left\| \widetilde{E}_s(r) \right\|_{\mathrm{F}} \mathrm{d}r \le I_{\mathrm{res}} e^{-\kappa \overline{g}_S \underline{\sigma}_S^2 t}. \tag{49}$$

Proof of Theorem 4.1. Write $g_s := g(\Sigma_s(0))$ within this proof. The endpoint derivative and layer covariance equations give

$$-\dot{\Delta} = \frac{1}{S}\sum_{s=0}^{S-1}\left[\mathrm{tr}(\widetilde{E}_s^\top \widetilde{E}_s \Sigma_s^u) + \mathrm{tr}(\widetilde{E}_s \widetilde{E}_s^\top \Sigma_s^v)\right] \ge \frac{1}{S}\sum_{s=0}^{S-1} g_s \left\|\widetilde{E}_s\right\|_{\mathrm{F}}^2.$$

Both the lower and upper singular-value bounds of the residual factors multiply through the prefix and suffix, so $\underline{\sigma}_S \|E\|_{\mathrm{F}} \le \left\|\widetilde{E}_s\right\|_{\mathrm{F}} \le \overline{\sigma}_S \|E\|_{\mathrm{F}}$. Therefore

$$-\dot{\Delta} \ge \overline{g}_S \underline{\sigma}_S^2 \|E\|_{\mathrm{F}}^2 \ge 2\kappa \overline{g}_S \underline{\sigma}_S^2 \Delta.$$

This proves the loss bound on each finite existence interval using the average layer margin. The square-root PL argument used in Theorem 3.1, followed by the upper force bound, gives (49) first with any finite endpoint. Thus all layer fundamental matrices and their inverses have norm at most $e^{I_{\mathrm{res}}}$. Their covariances remain bounded and positive definite. The finite-dimensional vector field is locally Lipschitz; these bounds prevent a finite maximal existence time. The force-tail bound now holds on $[t, \infty)$ and implies layer convergence by the argument of Theorem 3.1. Continuity of the finite residual product proves optimality of its limit. □

Proof of Theorem 4.2. The L^1 training equation in Section 4 means

$$\Sigma_s(t) = \Sigma_s(0) - \int_0^t (\widetilde{G}_s(r)\Sigma_s(r) + \Sigma_s(r)\widetilde{G}_s(r))\,\mathrm{d}r, \qquad \widetilde{G}_s = G_{\widetilde{E}_s}. \tag{50}$$

This equality in L^1 gives an absolutely continuous representative for almost every layer. The path bound makes the forces uniformly bounded over depth on finite training intervals, justifying differentiation of the endpoint and integration over depth. An L^1 residual field has invertible propagators. Gronwall applied to each propagator and its inverse gives $e^{-B}\|E\|_{\mathrm{F}} \le \left\|\widetilde{E}_s\right\|_{\mathrm{F}} \le e^{B}\|E\|_{\mathrm{F}}$. The depth integrals of the conserved supporting inequalities yield

$$-\dot{\Delta} = \int_0^1 \left[\mathrm{tr}(\widetilde{E}_s^\top \widetilde{E}_s \Sigma_s^u) + \mathrm{tr}(\widetilde{E}_s \widetilde{E}_s^\top \Sigma_s^v)\right] \mathrm{d}s \ge \overline{g} e^{-2B} \|E\|_{\mathrm{F}}^2 .$$

These integrals are finite by the L^1 covariance assumption and the uniform-in-depth force bounds. PL proves (12). As $g(\Sigma_s(0)) > 0$ almost everywhere and $g(\Sigma_s(0)) \le 2\|\Sigma_s(0)\|_{\mathrm{op}}$, its integral is finite and positive. The square-root argument gives the common layer-force tail bound

$$\int_t^\infty \left\|\widetilde{E}_s(r)\right\|_{\mathrm{F}} \mathrm{d}r \le I_{\mathrm{res}} e^{-\kappa \overline{g} e^{-2B} t}, \qquad I_{\mathrm{res}} := \frac{e^{3B}}{\overline{g}} \sqrt{\frac{2\Delta_0}{\kappa}}.$$

The fundamental-matrix argument proves almost-everywhere convergence and positive definiteness. The covariance tails are bounded by $2e^{2I_{\mathrm{res}}}\|\Sigma_s(0)\|_{\mathrm{op}} I_{\mathrm{res}} e^{-\kappa \overline{g} e^{-2B} t}$, an integrable bound over depth. This proves L^1 convergence. Continuity of the endpoint map on bounded L^1 sets, obtained by variation of constants and Gronwall, proves optimality of the limiting endpoint. □

Remark C.19 (Sharpness of the boundedness hypothesis)**.** *Unlike the two-layer flow of Theorem 3.1, positive conserved layer margins do not rule out unbounded weight growth as $S \to \infty$ or along continuous-depth training. A linear neural ODE has positive determinant, so it cannot fit an orientation-reversing target; training may instead approach a singular endpoint through unbounded residuals. For a concrete example, take the invertible scalar target $p_\star := -1$ and loss $\ell(p) := (p+1)^2/2$, with $L = \kappa = 1$. Initialize every layer with $\Sigma_s(0) = \mathrm{Id}_2$, realizable at width $n = 2$; every conserved support gap equals 2. At each finite depth $S \ge 2$, all residuals decrease to $-S$, so $\sup_{t\ge 0} \max_s |\Sigma_s^{vu}(t)| = S$ and no uniform $B < S$ exists. In continuous depth they decrease without bound, giving $\sup_{t\ge 0} \int_0^1 |\Sigma_s^{vu}(t)|\,\mathrm{d}s = \infty$. Both flows are global, but their endpoints tend to zero and their losses to $1/2 > \ell_\star = 0$. At finite depth the target is representable: failure occurs along this symmetric trajectory, whereas the neural ODE excludes the target altogether. In both cases, positive conserved margins therefore fail to ensure the residual bounds required by the theorems.*

Verification of Remark C.19. The scalar target -1 is invertible, and $|\ell'(p)|^2 = (p+1)^2 = 2\ell(p)$, so $L = \kappa = 1$ and $\ell_\star = 0$. The width-two initialization $U_s(0) = (\sqrt{2}, 0)$, $V_s(0) = (0, \sqrt{2})$ gives $\Sigma_s(0) = \mathrm{Id}_2$ and initial endpoint 1 at every depth. For scalar layers write $a_s := \Sigma_s^u = \Sigma_s^v$ and $r_s := \Sigma_s^{vu}$. Equality of the diagonal blocks is preserved, and the layer equations reduce to

$$\dot{a}_s = -2\widetilde{E}_s r_s, \qquad \dot{r}_s = -2\widetilde{E}_s a_s.$$

Thus $a_s^2 - r_s^2 = 1$, and positivity gives

$$\Sigma_s(t) = \begin{pmatrix} \sqrt{1 + r_s(t)^2} & r_s(t) \\ r_s(t) & \sqrt{1 + r_s(t)^2} \end{pmatrix}.$$

The eigenvalues of $J\Sigma_s$ are ± 1, whereas those of Σ_s are $\sqrt{1 + r_s^2} \pm |r_s| > 0$. Uniqueness and symmetry preserve identical layers. Denote their common residual by r_S. At finite depth, $\widetilde{E}_s = (p_S + 1)(1 + r_S/S)^{S-1}$, giving

$$\begin{aligned} p_S &= (1 + r_S/S)^S, \qquad r_S(0) = 0, \\ \dot{r}_S &= -2\sqrt{1 + r_S^2}\big((1 + r_S/S)^S + 1\big)(1 + r_S/S)^{S-1}. \end{aligned} \tag{51}$$

For $S \geq 2$, its vector field is locally Lipschitz and vanishes at $r_S = -S$. Starting at zero, the solution decreases and cannot reach this equilibrium in finite time. It stays in the compact interval $[-S, 0]$, so it is global. Any limit strictly above $-S$ would have strictly negative velocity; hence its limit is $-S$, $p_S \to 0$, and $\ell(p_S) \to 1/2$. In particular, every candidate bound $B < S$ is exceeded at a finite training time, although $|r_S(t)| < S$ at every finite time.

The finite-depth architecture can represent -1 by taking one residual factor equal to -1 and the others equal to 1. The obstruction here is the preserved symmetry of the initialized trajectory: all its residual factors stay positive and approach zero. At $S = 1$, this obstruction disappears; $\dot{r}_1 = -2\sqrt{1 + r_1^2}(r_1 + 2)$ instead gives $r_1 \to -2$ and the optimal endpoint -1.

For a depth-constant field, the continuous propagators are $P_{b,a} = e^{(b-a)r}$, so $\widetilde{E}_s = (e^r + 1)e^r$ and

$$p = e^r, \qquad \dot{r} = -2\sqrt{1 + r^2}(e^r + 1)e^r, \qquad r(0) = 0. \tag{52}$$

For $r \leq 0$ its velocity is bounded, strictly negative, and has no finite zero. The solution is therefore global and tends to $-\infty$. Since $\int_0^1 |\Sigma_s^{vu}(t)|\,\mathrm{d}s = |r(t)|$, every finite integral bound is eventually exceeded as well. The endpoint tends to zero. The limiting loss $1/2$ is the infimum over positive scalar endpoints and exceeds the ambient minimum $\ell_\star = 0$. The associated depth-constant covariance is a global locally absolutely continuous L^1 trajectory, positive definite at every finite time. Moreover, for $S \geq 2$ and $-S < r_S \leq 0$, $(1+r_S/S)^{S-1} \leq e^{r_S/2}$ and $(1+r_S/S)^S + 1 \leq 2$. Since $\sqrt{1 + r^2}e^{r/2} \leq 1$ for $r \leq 0$, both $|\dot{r}_S|$ and $|\dot{r}|$ are at most 4. On $[0, T]$, all residuals therefore lie in $[-4T, 0]$. The finite-depth vector fields converge uniformly on this compact interval, and continuous dependence gives $r_S \to r$ uniformly on every bounded training interval $[0, T]$. □

Remark C.20 (Zero-target variant)**.** *The same path bounds can fail while the loss tends to zero. With $\ell(p) := p^2/2$ and the same initialization, the equations are $\dot{r}_S = -2\sqrt{1 + r_S^2}(1 + r_S/S)^{2S-1}$ and $\dot{r} = -2\sqrt{1 + r^2}e^{2r}$. The preceding argument gives $r_S \to -S$ for every $S \geq 1$, $r \to -\infty$ in continuous depth, and endpoints tending to the zero target. A uniform residual-path bound would keep the endpoints bounded away from zero instead.*

Remark C.21 (Residual-path control and orientation)**.** *Positive-definite layer covariances can coexist with singular residual products. For a scalar depth-two network, take both layer predictors equal to -2, so $A_0 = A_1 = 0$, and the endpoint loss $\ell(p) := (p-1)^2/2$. Both backpropagated gradients vanish although the endpoint is not optimal. These predictors admit positive-definite layer covariances $\left(\begin{smallmatrix} \sqrt{5} & -2 \\ -2 & \sqrt{5} \end{smallmatrix}\right)$, whose signed eigenvalues are $+1, -1$. Extending the earlier discretization and sampling arguments to this setting requires control of the residual products along the corresponding training trajectory; the two-layer GD theorem does not apply verbatim.*

In continuous depth, $\det P(t) = \exp(\int_0^1 \operatorname{tr} \Sigma_s^{vu}(t)\,\mathrm{d}s) > 0$. The uniform path bound also keeps $\sigma_{\min}(P(t)) \geq e^{-B}$, so any convergent endpoint remains in the orientation-preserving invertible component. If every global minimizer has negative determinant, these path bounds are incompatible with convergence to a global minimizer.

C.7 Heavy-ball Momentum

We first expand the closed phase equation, then establish global existence, position-deformation estimates, and the energy bootstrap. For positions $x := (u, v)$ and velocities $y := (p, q)$, let

$\nu_t \in \mathcal{P}_2(\mathbb{R}^{2m})$ be their joint distribution. Its covariance is $\Sigma_t := \int \binom{x}{y}\binom{x}{y}^\top \mathrm{d}\nu_t(x,y)$, with mixed block $\Xi_t := \int xy^\top \mathrm{d}\nu_t(x,y)$. Finite-width factors correspond to $\nu_t := n^{-1}\sum_i \delta_{(x_i(t),y_i(t))}$. The characteristics satisfy

$$\dot{x} = y, \qquad \dot{y} = -G(\Sigma^{vu})x - \gamma y. \tag{53}$$

Taking its second moments gives the block system

$$\begin{aligned} \dot{\Sigma}_t^{(u,v)} &= \Xi_t + \Xi_t^\top, \qquad \dot{\Xi}_t = \Sigma_t^{(p,q)} - \Sigma_t^{(u,v)} G_t - \gamma \Xi_t, \\ \dot{\Sigma}_t^{(p,q)} &= -G_t \Xi_t - \Xi_t^\top G_t - 2\gamma \Sigma_t^{(p,q)}. \end{aligned} \tag{54}$$

In particular, the predictor velocity and acceleration are

$$\dot{\Sigma^{vu}} = \Sigma^{qu} + \Sigma^{vp}, \qquad \ddot{\Sigma^{vu}} + \gamma \dot{\Sigma^{vu}} = 2\Sigma^{qp} - E\Sigma^u - \Sigma^v E. \tag{55}$$

Proof of Proposition 5.1. Since G_t is symmetric, $\mathsf{M}_t^\top J_{\mathrm{c}} + J_{\mathrm{c}}\mathsf{M}_t = -\gamma J_{\mathrm{c}}$. Differentiate the two-characteristic pairing and the fundamental-matrix identity. Invertibility also gives $\Psi_t J_{\mathrm{c}} (\Psi_t)^\top = e^{-\gamma t} J_{\mathrm{c}}$, whence $J_{\mathrm{c}}\Sigma_t = e^{-\gamma t}(\Psi_t)^{-\top}(J_{\mathrm{c}}\Sigma_0)(\Psi_t)^\top$. □

Proposition C.22 (Global full phase flow). *If ℓ is C^1, has locally Lipschitz gradient, and is bounded below, every $\Sigma_0 \succeq 0$ has a global positive semidefinite solution of (54). This assertion alone gives neither uniform position ellipticity nor convergence to a minimizer.*

Proof of Proposition C.22. Local well-posedness follows from local Lipschitz continuity. The linear phase characteristic lift preserves positivity. The energy identity gives $\operatorname{tr}\Sigma_t^{(p,q)} \le 2\mathcal{E}_0$. Positivity and Cauchy–Schwarz yield

$$\sqrt{\operatorname{tr}\Sigma_t^{(u,v)}} \le \sqrt{\operatorname{tr}\Sigma_0^{(u,v)}} + t\sqrt{2\mathcal{E}_0}, \qquad \|\Xi_t\|_{\mathrm{F}}^2 \le \operatorname{tr}\Sigma_t^{(u,v)} \operatorname{tr}\Sigma_t^{(p,q)}.$$

The first inequality also follows directly from the L^2 triangle inequality along characteristics, including times when the trace is zero. Every block is bounded on each finite interval; continuation proves global existence. □

For the matrix calculations in the momentum sections, abbreviate $A := \Sigma^{(u,v)}$ and $B := \Sigma^{(p,q)}$. Whenever $A \succ 0$, define

$$D := \Xi^\top A^{-1}, \qquad Q := B - DAD^\top \succeq 0. \tag{56}$$

Here D is a regression matrix, not assumed symmetric, and Q is its positive semidefinite residual covariance (the Schur complement). They encode all phase blocks by $\Xi = AD^\top$ and $B = DAD^\top + Q$.

Proposition C.23 (General covariance coordinates). *On the chart $A \succ 0$, (54) is equivalent to*

$$\begin{aligned} \dot{A} &= DA + AD^\top, \\ \dot{D} &= -G - \gamma D - D^2 + QA^{-1}, \\ \dot{Q} &= -2\gamma Q - DQ - QD^\top. \end{aligned} \tag{57}$$

Moreover, positive semidefiniteness controls the position deformation:

$$\|D\|_{\mathrm{op}}^2 \le \frac{2\mathcal{K}}{\lambda_{\min}(A)}. \tag{58}$$

Proof of Proposition C.23. Differentiate $D = \Xi^\top A^{-1}$ and $Q = B - DAD^\top$, using (54). This gives the three displayed identities. Since $DAD^\top \preceq B$, its trace is at most $2\mathcal{K}$, whereas $\lambda_{\min}(A)\,\|D\|_{\mathrm{op}}^2 \le \operatorname{tr}(DAD^\top)$. □

Lemma C.24 (Conditional phase-energy decay). *Under Assumption 1.1, let a positive semidefinite solution of (54) satisfy $a\mathrm{Id}_m \preceq \Sigma_t^{(u,v)} \preceq b\mathrm{Id}_m$ for $0 \le t \le T$, where $0 < a \le b < \infty$. Put $\mathfrak{K} := 4a\kappa$, $\mathfrak{L} := 4bL$, and $\rho := \mathfrak{L} + 2\sqrt{2L\mathcal{E}_0}$. Then*

$$\mathcal{E}_t \le 3\mathcal{E}_0 e^{-\lambda t}, \qquad \lambda := \frac{a\kappa\gamma}{\rho + 2\gamma^2}, \qquad 0 \le t \le T. \tag{59}$$

Proof. Set $\mathcal{N} := \operatorname{tr}(G\Sigma^{(u,v)}G)$, the mean squared particle force. The hidden principal blocks are between $a\mathrm{Id}$ and $b\mathrm{Id}$, hence

$$2a\,\|E\|_{\mathrm{F}}^2 \le \mathcal{N} \le 2b\,\|E\|_{\mathrm{F}}^2\,.$$

Combining PL with the smooth-gradient upper bound in Lemma C.1, and applying Cauchy–Schwarz to the phase law, gives

$$\mathfrak{K}\Delta \le \mathcal{N} \le \mathfrak{L}\Delta, \qquad |\dot{\Delta}| = |\operatorname{tr}(G\Xi)| \le \sqrt{2\mathcal{K}\mathcal{N}}. \tag{60}$$

All second derivatives below are understood almost everywhere. Lipschitz continuity of $\nabla\ell$ implies that E_t is locally absolutely continuous and $\left\|\dot{E}_t\right\|_{\mathrm{F}} \le L\left\|\dot{\Sigma^{vu}}_t\right\|_{\mathrm{F}}$. Differentiating (55) gives

$$\ddot{\Delta} = H - \mathcal{N} - \gamma\dot{\Delta}, \qquad H := \left\langle \dot{E}, \dot{\Sigma^{vu}} \right\rangle + 2\left\langle E, \Sigma^{qp}\right\rangle \le \rho\mathcal{K}.$$

For the last bound, positive phase blocks give $\|\Sigma^{qu}\|_{\mathrm{F}}^2 \le b\operatorname{tr}\Sigma^q$ and $\|\Sigma^{vp}\|_{\mathrm{F}}^2 \le b\operatorname{tr}\Sigma^p$, hence $\left\|\dot{\Sigma^{vu}}\right\|_{\mathrm{F}}^2 \le 4b\mathcal{K}$. Also $\|\Sigma^{qp}\|_{\mathrm{F}} \le \sqrt{\operatorname{tr}\Sigma^q \operatorname{tr}\Sigma^p} \le \mathcal{K}$ and $\|E\|_{\mathrm{F}} \le \sqrt{2L\mathcal{E}_0}$.

Choose $\epsilon := \gamma/[2(\rho + 2\gamma^2)]$ and $\mathcal{V} := \mathcal{E} + \epsilon\dot{\Delta}$. Since

$$\epsilon \le \frac{1}{4\sqrt{2\rho}} \le \frac{1}{\sqrt{2\mathfrak{L}}}, \qquad |\dot{\Delta}| \le \sqrt{\mathfrak{L}/2}\,\mathcal{E},$$

one has $\mathcal{E}/2 \le \mathcal{V} \le 3\mathcal{E}/2$. The inequality $\gamma\sqrt{2\mathcal{K}\mathcal{N}} \le \mathcal{N}/4 + 2\gamma^2\mathcal{K}$ yields

$$\dot{\mathcal{V}} \le -\tfrac{3}{4}\epsilon\mathcal{N} - \big(2\gamma - \epsilon(\rho + 2\gamma^2)\big)\mathcal{K} \le -\tfrac{3}{4}\epsilon\mathfrak{K}\Delta - \gamma\mathcal{K}.$$

Now $\frac{3}{4}\epsilon\mathfrak{K} \le \gamma$, since $a\kappa \le bL$ and $\rho \ge 4bL$. It follows that $\dot{\mathcal{V}} \le -\frac{2}{3}(\frac{3}{4}\epsilon\mathfrak{K})\mathcal{V} = -\lambda\mathcal{V}$. Integrating the a.e. inequality proves (59).

□

Under Theorem 5.2, the position-covariance tail is

$$\left\|\Sigma_t^{(u,v)} - \Sigma_\infty^{(u,v)}\right\|_{\mathrm{F}} \le \frac{4\sqrt{6b_s\mathcal{E}_0}}{\lambda_s(\gamma)}e^{-\lambda_s(\gamma)t/2}. \tag{61}$$

For the position factors W or the position marginal of the characteristic phase law, the normalized Frobenius or W_2 tail has prefactor $2\sqrt{6\mathcal{E}_0}/\lambda_s(\gamma)$ at the same exponent.

Proof of Theorem 5.2. The full phase flow is global by Proposition C.22. On the connected position-positive chart containing time zero, let $h(t) := \int_0^t \|D_r\|_{\mathrm{op}}\,\mathrm{d}r$. The equation $\dot{A} = DA + AD^\top$ yields $A_t = Z_tA_0Z_t^\top$ with $\dot{Z} = DZ$, $Z_0 = \mathrm{Id}_m$, and thus

$$a_0e^{-2h(t)}\mathrm{Id}_m \preceq A_t \preceq b_0e^{2h(t)}\mathrm{Id}_m.$$

As long as $h \le s$, Lemma C.24 and Proposition C.23 imply

$$h(t) \le \int_0^t \sqrt{2\mathcal{K}_r/a_s}\,\mathrm{d}r \le \frac{2}{\lambda_s}\sqrt{\frac{6\mathcal{E}_0}{a_s}} = \mathfrak{S}_s(\gamma) < s.$$

There is no first time at which $h = s$. Nor can the chart terminate earlier: on each such interval $A \succeq a_s\mathrm{Id}$ and the globally defined full phase blocks stay bounded on finite horizons. This proves the global tube and energy estimate. Furthermore the positive-block factorization gives $\|\Xi\|_{\mathrm{F}}^2 \le b_s\operatorname{tr}B \le 6b_s\mathcal{E}_0e^{-\lambda_st}$. Integrate $\dot{A} = \Xi + \Xi^\top$ to get (61). For particles or a characteristic phase law, the remaining L^2 position displacement is at most $2\sqrt{6\mathcal{E}_0}\,e^{-\lambda_st/2}/\lambda_s$. The limit is therefore attained, with zero loss gap and full phase covariance $\operatorname{diag}(A_\infty, 0)$. □

If $\mathcal{E}_0 = 0$, the nonnegative kinetic and potential energies both vanish, the force vanishes, and the momentum trajectory is stationary. Positive definiteness of the *phase* covariance is never required; in particular zero-speed initialization is covered.

C.8 An exact decomposition of hidden kinetic energy

Prescribing the position covariance and its derivative does not determine the kinetic energy: velocity components that leave the covariance unchanged can still contribute to it.

Proposition C.25 (Minimum-energy lift of a covariance velocity)**.** *Fix $A \succ 0$ and a symmetric matrix H as the desired value of $\dot{A}$. Let $C = C^\top$ be the unique solution of $CA + AC = H$. Every positive semidefinite phase covariance with this position moment and derivative has regression coordinates*

$$D = C + \Omega, \qquad \Omega A + A\Omega^\top = 0, \qquad Q \succeq 0,$$

and its kinetic energy decomposes as

$$\mathcal{K} = \tfrac{1}{2}\operatorname{tr}(CAC) + \tfrac{1}{2}\operatorname{tr}(\Omega A\Omega^\top) + \tfrac{1}{2}\operatorname{tr} Q. \tag{62}$$

Thus the graph lift $D = C$, $Q = 0$ has minimum kinetic energy. An additional positive semidefinite Schur complement contributes exactly $\operatorname{tr} Q/2$.

Proof. In an eigenbasis of A, the Sylvester equation divides H_{ij} by the strictly positive sum of eigenvalues, proving existence and uniqueness. Equation $DA + AD^\top = H$ is equivalent to the stated condition on Ω. Since $B = DAD^\top + Q$, expand its trace. The two cross terms vanish in sum: ΩA is skew-symmetric and C is symmetric. Each term in (62) is nonnegative. □

Example C.26 (Identical position covariance and covariance derivative, arbitrarily large energy)**.** *Let $A = \sigma^2 \mathrm{Id}_m$, $C = 0$, choose any skew-symmetric Ω, and set $D = \Omega$, $Q = 0$. The initial predictor is zero because $A = \sigma^2 \mathrm{Id}_m$, and $\dot{A} = 0$ makes its derivative zero as well, independently of Ω. Nevertheless $\mathcal{K} = \sigma^2 \|\Omega\|_{\mathrm{F}}^2 /2$ can be arbitrarily large. The phase covariance is positive semidefinite; nonzero Ω gives a nonsymmetric regression matrix. For fixed a_0, b_0, L, κ, the best feasibility ratio (84) tends to zero as $\mathcal{E}_0 \to \infty$. Thus the initial position covariance and predictor speed do not determine whether the energy-damping certificate holds.*

D Tightened Conditional Momentum Rates

This appendix sharpens Lemma C.24. It gives the exact optimizer within the same Lyapunov estimates, the simpler parameter choice used in the numerical figures, and the corresponding global bootstrap with its energy prefactor.

Proposition D.1 (Tightened conditional energy decay)**.** *Under the loss and position bounds of Lemma C.24 on $[0, T]$, fix $\eta \in (0, 1)$ and $\varepsilon > 0$. With its constants $\mathfrak{K}, \mathfrak{L}, \rho$, set*

$$q := \varepsilon\sqrt{\mathfrak{L}/2}, \tag{63}$$

$$r_{\varepsilon,\eta} := \min\{\varepsilon(1-\eta)\mathfrak{K},\ 2\gamma - \varepsilon(\rho + \gamma^2/(2\eta))\}, \tag{64}$$

$$\lambda_{\varepsilon,\eta} := r_{\varepsilon,\eta}/(1+q). \tag{65}$$

If $q < 1$ and $r_{\varepsilon,\eta} > 0$, the corrected energy $\mathcal{V}_t := \mathcal{E}_t + \varepsilon\dot{\Delta}_t$ satisfies, for $0 \le t \le T$,

$$(1-q)\mathcal{E}_t \le \mathcal{V}_t \le (1+q)\mathcal{E}_t, \tag{66}$$

$$\mathcal{V}_t \le \mathcal{V}_0 e^{-\lambda_{\varepsilon,\eta} t}, \tag{67}$$

$$\ell(\Sigma_t^{vu}) - \ell_\star \le \mathcal{E}_t \le \frac{\mathcal{V}_0}{1-q} e^{-\lambda_{\varepsilon,\eta} t} \le \frac{1+q}{1-q}\mathcal{E}_0 e^{-\lambda_{\varepsilon,\eta} t}. \tag{68}$$

If $\dot{\Delta}_0 = 0$, the first energy prefactor is $\mathcal{E}_0/(1-q)$.

Proof. The force bounds (60) give $|\dot{\Delta}| \le \sqrt{2\mathfrak{L}\Delta\mathcal{K}} \le \sqrt{\mathfrak{L}/2}\,\mathcal{E}$, proving the equivalence. For the derivative, use $\dot{\mathcal{E}} = -2\gamma\mathcal{K}$, $\ddot{\Delta} \le \rho\mathcal{K} - \mathcal{N} - \gamma\dot{\Delta}$, and Young's inequality:

$$\gamma|\dot{\Delta}| \le \gamma\sqrt{2\mathcal{K}\mathcal{N}} \le \eta\mathcal{N} + \frac{\gamma^2}{2\eta}\mathcal{K}. \tag{69}$$

Since $\mathcal{N} \ge \mathfrak{K}\Delta$, this yields

$$\dot{\mathcal{V}} \le -\varepsilon(1-\eta)\mathcal{N} - [2\gamma - \varepsilon(\rho + \gamma^2/(2\eta))]\mathcal{K} \le -r_{\varepsilon,\eta}\mathcal{E} \le -\lambda_{\varepsilon,\eta}\mathcal{V}.$$

Integrate on $[0, T]$ and use the equivalence once more. □

Proposition D.2 (Optimal rate within the Lyapunov bounds)**.** *Under Proposition D.1, fix* $\eta \in (0,1)$ *and write*

$$A_\eta := (1-\eta)\mathfrak{K}, \qquad B_\eta := \rho + \gamma^2/(2\eta). \tag{70}$$

The unique maximizing correction at fixed η *is admissible and gives*

$$\varepsilon_\eta^\star = \frac{2\gamma}{A_\eta + B_\eta}, \tag{71}$$

$$q_\eta^\star := \varepsilon_\eta^\star \sqrt{\mathfrak{L}/2} < 1, \tag{72}$$

$$\lambda_\eta := \frac{2\gamma A_\eta}{A_\eta + B_\eta + \gamma\sqrt{2\mathfrak{L}}}. \tag{73}$$

The unique maximizer over η *and its rate are*

$$\eta_\star = \frac{\gamma}{\gamma + \sqrt{\gamma^2 + 2\rho + 2\gamma\sqrt{2\mathfrak{L}}}} \in (0, 1/2), \tag{74}$$

$$\lambda_\star := \max_{0<\eta<1} \lambda_\eta = \frac{2\gamma\mathfrak{K}}{\mathfrak{K} + \frac{1}{2}(\gamma + \sqrt{\gamma^2 + 2\rho + 2\gamma\sqrt{2\mathfrak{L}}})^2}. \tag{75}$$

Proof. At fixed η, the rate is the minimum of $\varepsilon A_\eta/(1+\varepsilon\sqrt{\mathfrak{L}/2})$ and $(2\gamma - \varepsilon B_\eta)/(1+\varepsilon\sqrt{\mathfrak{L}/2})$. The first is strictly increasing and the second strictly decreasing, so their intersection gives the unique optimum. It is admissible because $B_\eta > \rho + \gamma^2/2 \geq \gamma\sqrt{2\rho} \geq \gamma\sqrt{2\mathfrak{L}}$, hence $q_\eta^\star < 1$; both dissipation terms there equal $\varepsilon_\eta^\star A_\eta > 0$.

For the remaining optimization, set $\widetilde{\rho} := \rho + \gamma\sqrt{2\mathfrak{L}}$ and divide numerator and denominator of (73) by $1-\eta$. Maximizing the rate is equivalent to minimizing $(\widetilde{\rho} + \gamma^2/(2\eta))/(1-\eta)$. Its derivative has the sign of $\widetilde{\rho}\eta^2 + \gamma^2\eta - \gamma^2/2$, which increases strictly and vanishes only at (74). The quotient there is $\frac{1}{2}(\gamma + \sqrt{\gamma^2 + 2\widetilde{\rho}})^2$, proving the formula. □

The numerical figures use the simpler admissible choice

$$\eta_{\text{approx}} := \frac{\gamma}{\gamma + \sqrt{\gamma^2 + 2\rho}} \in (0, 1/2). \tag{76}$$

It minimizes B_η/A_η: differentiation gives $\rho\eta^2 + \gamma^2\eta - \gamma^2/2 = 0$. Thus it optimizes the dissipation rate before accounting for the equivalence factor $1 + q_\eta^\star$. That factor is retained in the following rigorous bound.

Corollary D.3 (Explicit rate used in the figures)**.** *Under Proposition D.1, use* η_{approx} *from* (76) *and set*

$$A_{\text{approx}} := (1-\eta_{\text{approx}})\mathfrak{K}, \qquad B_{\text{approx}} := \rho + \frac{\gamma^2}{2\eta_{\text{approx}}}, \tag{77}$$

$$\varepsilon_{\text{approx}} := \frac{2\gamma}{A_{\text{approx}} + B_{\text{approx}}}, \qquad q_{\text{approx}} := \varepsilon_{\text{approx}}\sqrt{\mathfrak{L}/2}. \tag{78}$$

These parameters are admissible. With $\mathcal{V}_0 = \mathcal{E}_0 + \varepsilon_{\text{approx}}\dot{\Delta}_0$,

$$\lambda_{\text{approx}} := \frac{\varepsilon_{\text{approx}} A_{\text{approx}}}{1 + q_{\text{approx}}}, \tag{79}$$

$$\ell(\Sigma_t^{vu}) - \ell_\star \leq \mathcal{E}_t \leq \frac{\mathcal{V}_0}{1 - q_{\text{approx}}} e^{-\lambda_{\text{approx}} t}, \qquad 0 \leq t \leq T. \tag{80}$$

If $\dot{\Delta}_0 = 0$*, the prefactor is* $\mathcal{E}_0/(1 - q_{\text{approx}})$.

Proof. The fixed-η optimum in Proposition D.2 is admissible for every $\eta \in (0,1)$. Apply Proposition D.1 with $\eta = \eta_{\text{approx}}$. □

The approximate choice concerns only optimization: its bound is exact, and $\lambda_{\text{approx}} \leq \lambda_\star$. Neither exponent is claimed to be the optimal rate of the underlying training dynamics.

Corollary D.4 (Global bootstrap with the correct prefactor)**.** *Under the loss and initialization hypotheses of Theorem 5.2, choose a tube $a_s := a_0 e^{-2s}$, $b_s := b_0 e^{2s}$ and an admissible (ε, η) for its constants. Put*

$$\mathfrak{C}_0 := \frac{\mathcal{E}_0 + \varepsilon \dot{\Delta}_0}{(1-q)\mathcal{E}_0}.$$

If $\mathcal{E}_0 > 0$ and

$$\frac{2}{\lambda_{\varepsilon,\eta}} \sqrt{\frac{2\mathfrak{C}_0 \mathcal{E}_0}{a_s}} < s, \tag{81}$$

then the tube and (68) hold globally. A sufficient condition using only initial position bounds and total energy replaces $\mathfrak{C}_0$ by $(1+q)/(1-q)$.

Proof. Within the initial connected positive-position chart, $\int_0^t \|D\|_{\mathrm{op}} \le \int_0^t \sqrt{2\mathcal{E}_r/a_s}\,\mathrm{d}r$. Proposition D.1 bounds this by the left side of (81). The strict first-exit proof of Theorem 5.2 then applies unchanged. Initial positivity and the Lyapunov equivalence imply $1 \le \mathfrak{C}_0 \le (1+q)/(1-q)$. □

The role of the prefactor. A faster conditional exponent must be accompanied by its own prefactor when testing the global deformation budget. The sharper state-dependent prefactor uses the initial loss derivative $\dot{\Delta}_0 = \langle E_0, \Sigma_0^{qu} + \Sigma_0^{vp} \rangle$ as well as total energy. In particular, zero predictor speed gives $\mathfrak{C}_0 = 1/(1-q)$. The parameter-only prefactor uses no extra initial derivative. Maximizing the conditional exponent alone need not minimize the global budget: the same choice of parameters also changes q and hence the prefactor. The closed-form optimizer should therefore be tested in (81), not substituted there with an unchanged prefactor.

E Large-Position-Scale Certificate Asymptotics

Increasing position variance can open a damping window without reducing the initial loss gap. We quantify that tradeoff and the best rates certified by the same bootstrap. For the two-scale white initialization of Corollary 5.3, the initial mechanical excess energy is

$$\mathcal{E}_0 = \ell(0) - \ell_\star + \frac{m\sigma_{qp}^2}{2}.$$

This quantity is fixed when σ_{qp} is fixed and $\sigma_{vu} \to +\infty$. It is useful to name the bootstrap gap

$$\mathfrak{G}_s(\gamma) := \mathfrak{S}_s(\gamma) - s, \qquad s_{\mathrm{cert}}(\gamma) := \inf\{s > 0 : \mathfrak{G}_s(\gamma) < 0\}.$$

Whenever this set is nonempty, define the corresponding boundary rate by

$$\lambda_{\mathrm{cert}}(\gamma) := \lambda_{s_{\mathrm{cert}}(\gamma)}(\gamma),$$

using continuity in s. The strict bootstrap condition need not hold at the boundary itself: the theorem applies at interior budgets $s > s_{\mathrm{cert}}$ with $\mathfrak{G}_s(\gamma) < 0$, whose rates approach this boundary value. Thus λ_{cert} is a limiting supremal certificate, rather than a rate asserted at a possibly nonadmissible boundary point. Define also the largest damping for which the bootstrap window is nonempty:

$$\gamma_{\max}^{\mathrm{cert}}(\sigma_{vu}) := \sup\{\gamma > 0 : \text{ there exists } s > 0 \text{ such that } \mathfrak{G}_s(\gamma) < 0\}.$$

The assumption $\mathcal{E}_0 > 0$ below excludes only the trivial zero-excess case, in which both the predictor loss gap and kinetic energy vanish and the trajectory is stationary.

Explicit damping interval and feasibility. At a fixed deformation budget, admissibility reduces to a quadratic inequality, making both damping boundaries explicit. For $\mathcal{E}_0 > 0$ and fixed deformation budget $s > 0$, write

$$c_s := \frac{s\kappa a_s^{3/2}}{4\sqrt{3\mathcal{E}_0/2}}. \tag{82}$$

The theorem's strict certificate is exactly $2\gamma^2 - c_s\gamma + \rho_s < 0$, or

$$c_s^2 > 8\rho_s, \qquad \frac{c_s - \sqrt{c_s^2 - 8\rho_s}}{4} < \gamma < \frac{c_s + \sqrt{c_s^2 - 8\rho_s}}{4}. \tag{83}$$

These are sufficient boundaries, not instability thresholds.

Optimizing feasibility. To make this interval nonempty as easily as possible, maximize

$$\frac{c_s^2}{8\rho_s} = \frac{s^2\kappa^2 a_0^3 e^{-6s}}{192\mathcal{E}_0(4b_0Le^{2s} + 2\sqrt{2L\mathcal{E}_0})}. \tag{84}$$

Its unique maximizer s_{feas} belongs to $(1/4, 1/3)$ and satisfies

$$\frac{1}{s_{\mathrm{feas}}} = 3 + \frac{4b_0Le^{2s_{\mathrm{feas}}}}{4b_0Le^{2s_{\mathrm{feas}}} + 2\sqrt{2L\mathcal{E}_0}}. \tag{85}$$

For a proof, write the denominator term as $Ae^{2s} + B$, with $A, B > 0$. The logarithmic derivative is $2/s - 6 - 2Ae^{2s}/(Ae^{2s} + B)$, and its derivative is $-2/s^2 - 4ABe^{2s}/(Ae^{2s} + B)^2 < 0$. It changes sign between $1/4$ and $1/3$. The resulting scalar equation is explicit and uniquely solvable; no trajectory has to be computed. This budget optimizes *nonemptiness*, not the upper damping endpoint or the decay exponent. Those are different optimization questions.

Asymptotic certificate bounds. The large-position limit distinguishes fixed-damping behavior from the largest damping that the certificate can accommodate.

Proposition E.1 (Large-position-scale certificate asymptotics)**.** *For the two-scale white initialization above, fix $\sigma_{qp} \geq 0$, assume $\mathcal{E}_0 > 0$, and let $\sigma_{vu} \to +\infty$. For each fixed $\gamma > 0$ and fixed $s > 0$,*

$$\mathfrak{S}_s(\gamma) = \frac{16Le^{5s}}{\kappa\gamma\,\sigma_{vu}}\sqrt{\frac{3\mathcal{E}_0}{2}} + O(\sigma_{vu}^{-3}).$$

The expansion is uniform for s in bounded sets. Consequently,

$$s_{\mathrm{cert}}(\gamma) = \frac{16L}{\kappa\gamma\,\sigma_{vu}}\sqrt{\frac{3\mathcal{E}_0}{2}} + o(\sigma_{vu}^{-1}),$$

and

$$\lambda_{\mathrm{cert}}(\gamma) = \frac{\kappa\gamma}{4L} - \frac{16}{\sigma_{vu}}\sqrt{\frac{3\mathcal{E}_0}{2}} + o(\sigma_{vu}^{-1}).$$

Moreover, the upper damping boundary satisfies

$$\gamma_{\max}^{\mathrm{cert}}(\sigma_{vu}) = \frac{\kappa\sigma_{vu}^3}{24e\sqrt{3\mathcal{E}_0/2}}\big(1 + o(1)\big).$$

Proof. For fixed $\gamma > 0$, the exact expression (15) can be expanded directly. Since

$$\rho_s = 4L\sigma_{vu}^2e^{2s} + 2\sqrt{2L\mathcal{E}_0},$$

we have

$$\lambda_s(\gamma) = \frac{\kappa\gamma}{4Le^{4s}} + O(\sigma_{vu}^{-2}).$$

Substituting this into the definition of $\mathfrak{S}_s(\gamma)$, with $a_s = \sigma_{vu}^2e^{-2s}$, gives the first expansion. Write

$$c_0 := \frac{16L}{\kappa\gamma}\sqrt{\frac{3\mathcal{E}_0}{2}}.$$

The expansion is uniform near $s = 0$. Thus every $s = c/\sigma_{vu}$, with $c > 0$ fixed, satisfies

$$\mathfrak{G}_s(\gamma) = \frac{c_0 - c}{\sigma_{vu}} + o(\sigma_{vu}^{-1}),$$

uniformly when c ranges over a compact set. Hence budgets whose leading constant is below c_0 are inadmissible, whereas those whose leading constant is strictly above c_0 are admissible for all sufficiently large σ_{vu}. This proves the displayed expansion for $s_{\mathrm{cert}}(\gamma)$. Evaluating

$$\lambda_s(\gamma) = \frac{\kappa\gamma}{4L}e^{-4s} + O(\sigma_{vu}^{-2})$$

at $s = s_{\mathrm{cert}}(\gamma)$ gives the expansion for $\lambda_{\mathrm{cert}}(\gamma)$.

It remains to identify the high-damping boundary. First observe the exact upper bound

$$\lambda_s(\gamma) = \frac{a_s\kappa\gamma}{\rho_s + 2\gamma^2} \leq \frac{\kappa\sigma_{vu}^2e^{-2s}}{2\gamma}.$$

Consequently every admissible pair satisfies

$$\mathfrak{S}_s(\gamma) \geq \frac{8\gamma e^{3s}}{\kappa\sigma_{vu}^3}\sqrt{\frac{3\mathcal{E}_0}{2}},$$

and hence

$$\gamma < \frac{\kappa\sigma_{vu}^3}{8\sqrt{3\mathcal{E}_0/2}}\, se^{-3s} \leq \frac{\kappa\sigma_{vu}^3}{24e\sqrt{3\mathcal{E}_0/2}}.$$

This proves the required asymptotic upper bound uniformly over all budgets, not only over bounded s.

For the matching lower bound, take γ of order σ_{vu}^3 and keep s in a compact subset of $(0, +\infty)$. Then $2\gamma^2$ dominates ρ_s in (15), and

$$\lambda_s(\gamma) = \frac{\kappa\sigma_{vu}^2 e^{-2s}}{2\gamma}\big(1 + o(1)\big).$$

Consequently,

$$\mathfrak{S}_s(\gamma) = \frac{8\gamma e^{3s}}{\kappa\sigma_{vu}^3}\sqrt{\frac{3\mathcal{E}_0}{2}}\big(1 + o(1)\big),$$

so the condition $\mathfrak{S}_s(\gamma) < s$ is asymptotically equivalent to

$$\gamma < \frac{\kappa\sigma_{vu}^3}{8\sqrt{3\mathcal{E}_0/2}}\, se^{-3s}.$$

The right-hand side is maximized at $s = 1/3$. Choosing this budget and any fixed multiplicative factor below the maximum gives an admissible pair for all sufficiently large σ_{vu}, proving the matching lower bound and the claimed asymptotic formula. □

Proposition E.1 has two complementary messages. At fixed damping, increasing the initial position scale makes the certified ellipticity deformation vanish and the certified rate converges to $\kappa\gamma/(4L)$. At fixed σ_{vu}, however, this particular bootstrap theorem cannot certify arbitrarily large damping: the admissible region in (s, γ) eventually disappears, with the upper edge scaling like σ_{vu}^3. This is a limitation of the certificate, not a claim that the underlying damped dynamics fail in the large-damping regime.

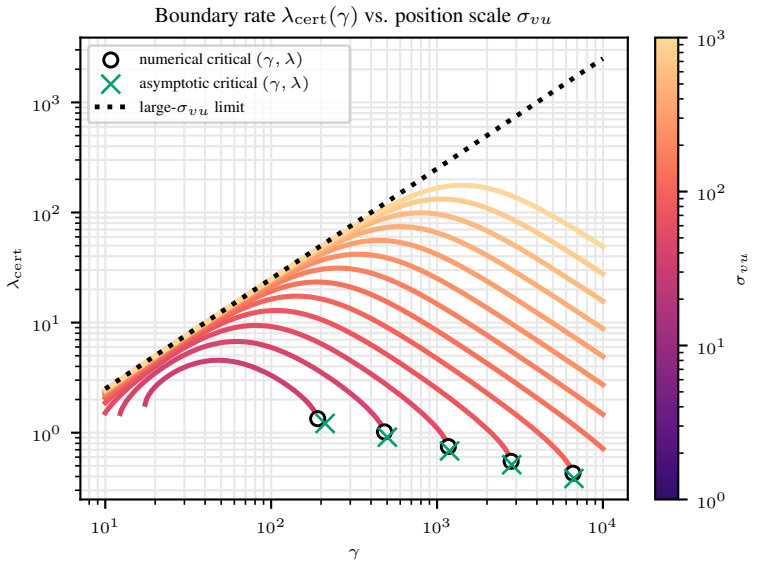

Figure 5: Limiting bootstrap rate at the admissibility boundary as a function of γ for several initial position scales σ_{vu}. Solid curves are obtained by grid-searching the bootstrap certificate for the boundary value $\lambda_{s_{\mathrm{cert}}}(\gamma)$; rates at strictly admissible nearby budgets approach these curves. The dotted black curve is the large-σ_{vu} limiting rate $\kappa\gamma/(4L)$. The plot uses $(d_u, d_v) = (3, 2)$, $\kappa = L = 1$, $\sigma_{qp} = 1$, and $\Delta_0 = \ell(0) - \ell_\star = 1$. For scales whose upper damping boundary lies within the scanned range, markers show numerical and asymptotic approximations to the maximal admissible damping and its corresponding boundary rate.

F Counting Covariance-Level Mean-Field Laws

This appendix records the Lie-algebra computation behind the covariance-level completeness statement used in Section 2. It is a continuum, covariance-level analogue of the local/generic finite-width completeness theorem of Marcotte–Gribonval–Peyré Marcotte et al. (2023).

Width-consistent laws and their Wasserstein interpretation are defined in Appendix B. We work here directly with functions of the covariance and the vector fields that must annihilate them.

F.1 Split Bures vector fields

To characterize the laws shared by the architecture, we allow every smooth split loss:

$$F(\Sigma) := \ell(\Sigma^{vu}), \qquad \Sigma = \begin{pmatrix} \Sigma^u & \Sigma^{uv} \\ \Sigma^{vu} & \Sigma^v \end{pmatrix}.$$

If $E := \nabla\ell(\Sigma^{vu})$, set

$$G_E := \begin{pmatrix} 0 & E^\top \\ E & 0 \end{pmatrix} \in \mathfrak{p}_J, \qquad \mathfrak{p}_J := \{G \in \mathbb{S}^m : GJ + JG = 0\}.$$

The corresponding Bures vector field is

$$X_\ell(\Sigma) := -(G_{\nabla\ell(\Sigma^{vu})}\Sigma + \Sigma G_{\nabla\ell(\Sigma^{vu})}). \tag{86}$$

We denote the resulting space of split Bures vector fields by

$$\mathcal{X}_{\mathrm{split}} := \{X_\ell : \ell \in C^\infty(\mathbb{R}^{d_v \times d_u})\}.$$

It is a real vector space. Its constant-gradient subspace is obtained from linear losses $\ell_E(\Sigma^{vu}) := \langle E, \Sigma^{vu}\rangle$:

$$\mathcal{X}_{\mathrm{lin}} := \{X_E : E \in \mathbb{R}^{d_v \times d_u}\}, \qquad X_E(\Sigma) := -(G_E\Sigma + \Sigma G_E).$$

F.2 Generated Lie algebra

Lie brackets reveal the directions generated by all split losses, which determine how many independent common laws can remain. Let

$$\mathfrak{o}(J) := \{A \in \mathbb{R}^{m\times m} : A^\top J + JA = 0\}$$

be the Lie algebra of $O(d_u, d_v)$. For $A \in \mathfrak{o}(J)$, define the infinitesimal congruence field

$$\xi_A(\Sigma) := -(A\Sigma + \Sigma A^\top).$$

Its flow is $\Sigma(t) = e^{-tA}\Sigma(0)e^{-tA^\top}$, so it is tangent to the $O(d_u, d_v)$-congruence orbit of Σ. When $A = G_E \in \mathfrak{p}_J$, $\xi_A = X_E$.

Proposition F.1 (Split fields generate $\mathfrak{o}(d_u, d_v)$)**.** *With the bracket convention* $[X, Y] := DY[X] - DX[Y]$*, the map*

$$A \longmapsto \xi_A$$

is a Lie-algebra homomorphism:

$$[\xi_A, \xi_B] = \xi_{[A,B]}.$$

Moreover,

$$\mathrm{Lie}(\mathfrak{p}_J) = \mathfrak{o}(J).$$

Consequently the Lie algebra generated by $\mathcal{X}_{\mathrm{lin}}$*, the constant-gradient split Bures fields, is*

$$\{\xi_A : A \in \mathfrak{o}(J)\}.$$

Proof. The homomorphism identity follows by differentiating the linear vector fields ξ_A. It remains to compute the matrix Lie algebra generated by $\mathfrak{p}_J$. For

$$G_E = \begin{pmatrix} 0 & E^\top \\ E & 0 \end{pmatrix}, \qquad G_F = \begin{pmatrix} 0 & F^\top \\ F & 0 \end{pmatrix},$$

one has

$$[G_E, G_F] = \begin{pmatrix} E^\top F - F^\top E & 0 \\ 0 & EF^\top - FE^\top \end{pmatrix}.$$

These brackets span the block-diagonal skew-symmetric algebra

$$\mathfrak{k}_J := \left\{ \begin{pmatrix} A & 0 \\ 0 & D \end{pmatrix} : A^\top = -A,\ D^\top = -D \right\},$$

by choosing E, F supported on one common row or one common column. Since

$$\mathfrak{o}(J) = \mathfrak{k}_J \oplus \mathfrak{p}_J,$$

the generated Lie algebra is all of $\mathfrak{o}(J)$. □

Fields versus their generated distribution. The full space $\mathcal{X}_{\mathrm{split}}$ contains nonlinear fields and is not identified with a finite-dimensional matrix Lie algebra. Each of its fields is a smooth, state-dependent linear combination of a basis of the X_E's. Consequently all such fields and their brackets remain tangent to the same congruence orbits, while the constant-gradient fields already generate the whole orbit tangent space. It is this distribution that determines the common conservation laws.

F.3 Completeness of the spectral laws

The generated directions coincide with the tangent directions of covariance congruence orbits; their codimension gives the local law count. For a constant-gradient split Bures field, $B := J\Sigma$ satisfies

$$\dot{B} = [G_E, B].$$

The trace powers $\mathcal{H}_k$ in (2) are therefore conserved. The next corollary states that, locally and generically, these are the only smooth covariance-level laws common to all split losses.

Let

$$\Omega_{\mathrm{simp}} := \{\Sigma \in \mathbb{S}^m_{++} : J\Sigma \text{ has simple spectrum}\}.$$

Corollary F.2 (Marcotte-style covariance-level completeness)**.** *Let h be a smooth function on an open subset of Ω_{simp}. If h is conserved by every split Bures vector field X_ℓ of (86), then h is locally a smooth function of the eigenvalues of $J\Sigma$, equivalently of*

$$\mathcal{H}_1(\Sigma), \dots, \mathcal{H}_m(\Sigma).$$

These m laws are functionally independent on Ω_{simp}. Thus the characteristic polynomial of $J\Sigma$ gives a complete local/generic list of covariance-level mean-field conservation laws for the split two-layer linear architecture.

Proof. It is enough to use the linear split losses. If $X_E h = 0$ for every E, then h is also annihilated by all iterated Lie brackets of the X_E's. By Proposition F.1,

$$\mathrm{d}h(\Sigma)[\xi_A(\Sigma)] = 0 \qquad \text{for all } A \in \mathfrak{o}(J).$$

Thus h is constant on the connected pieces of the $O(d_u, d_v)$-congruence orbits.

The orbit invariants are exactly the eigenvalues of $J\Sigma$ on Ω_{simp}. Indeed,

$$J(Q\Sigma Q^\top) = Q^{-\top}(J\Sigma)Q^\top \qquad (Q^\top J Q = J),$$

so the spectrum is preserved. Conversely, for $\Sigma \succ 0$, the matrix $J\Sigma$ is similar to the symmetric matrix $\Sigma^{1/2} J \Sigma^{1/2}$, hence has real nonzero eigenvalues with d_u positive and d_v negative signs. On the simple-spectrum stratum its J-orthogonal eigenbasis can be locally normalized to a J-orthonormal basis, so equal locally ordered eigenvalues imply membership in the same local $O(d_u, d_v)$-orbit. Therefore any smooth orbit-constant function factors locally through the eigenvalues. Newton identities replace the eigenvalues by $\mathcal{H}_1, \dots, \mathcal{H}_m$. More explicitly, for the locally ordered eigenvalues λ_j,

$$\frac{\partial \mathcal{H}_k}{\partial \lambda_j} = k\lambda_j^{k-1}, \qquad 1 \le k, j \le m.$$

This scaled Vandermonde matrix is invertible at distinct eigenvalues. The eigenvalues themselves vary independently by varying the diagonal covariance in a fixed normalized eigenbasis. Hence the spectral map has rank m, and its local level sets have dimension $m(m+1)/2 - m = m(m-1)/2$, exactly the orbit dimension. □

For empirical covariances $\Sigma = WW^\top/n$ with $W = \binom{U}{V}$, the finite-width Gram imbalance is

$$\Gamma := V^\top V - U^\top U = -W^\top J W.$$

The nonzero spectra of $J\Sigma = JWW^\top/n$ and $-\Gamma/n = W^\top J W/n$ coincide. Corollary F.2 is therefore the covariance quotient of the Marcotte–Gribonval–Peyré completeness theorem: the full labeled finite-width conservation data include the matrix Γ, while the covariance-level law keeps only its spectrum.

F.4 A ReLU contrast

Homogeneity preserves a scalar balance beyond linear networks, but it does not give the same covariance closure or spectral family. For a two-layer ReLU or more generally one-homogeneous network, the characteristic equations associated with a potential $v^\top \psi(u)$, with ψ one-homogeneous and possibly time-dependent, have the form

$$\dot{u} = -D\psi(u)^\top v, \qquad \dot{v} = -\psi(u).$$

Euler's identity $D\psi(u)u = \psi(u)$ gives

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\|u\|^2 - \|v\|^2\right) = 0.$$

For ReLU this calculation is interpreted almost everywhere along the characteristics, with an activation-boundary convention satisfying Euler's identity (for instance derivative zero at a zero preactivation). Thus the full mean-field flow preserves the pushforward distribution of $\|u\|^2 - \|v\|^2$, and its covariance-level shadow is the total imbalance

$$\int \left(\|u\|^2 - \|v\|^2\right) \mathrm{d}\mu(u, v).$$

Unlike the linear split case above, this ReLU statement is not obtained from a closed finite-dimensional Bures covariance equation; it is included only to highlight how much smaller the covariance-level conservation-law problem can be than the full particle-level one. No completeness claim for ReLU laws is made here.

G INTRINSIC DYNAMICS

Special conserved spectral values express some covariance blocks in terms of fewer variables. This appendix derives these exact reductions.

G.1 A TWO-POINT SIGNED SPECTRUM: HYPERBOLIC ISOTROPY

A special two-point spectrum determines the hidden covariance blocks from the predictor and therefore closes a smaller equation.

Definition G.1 (Hyperbolic isotropy). *For $\alpha > 0$, the hyperbolically isotropic (HI) leaf is*

$$\mathcal{M}_\alpha := \{\Sigma \succeq 0 : \Sigma J \Sigma = \alpha^2 J\}. \tag{87}$$

The identity forces $\Sigma \succ 0$ and is equivalent to $(J\Sigma)^2 = \alpha^2 \mathrm{Id}_m$. The symmetric representative of $J\Sigma$ then has d_u eigenvalues $+\alpha$ and d_v eigenvalues $-\alpha$; equivalently, $\mathcal{H}_k = \alpha^k(d_u + (-1)^k d_v)$ for $k = 1, \dots, m$. Spectral conservation proves invariance of $\mathcal{M}_\alpha$ despite the repeated eigenvalues. Directly, the derivative of $\Sigma J \Sigma$ is $-G(\Sigma J\Sigma) - (\Sigma J\Sigma)G$, with constant solution $\alpha^2 J$.

Proposition G.2 (Predictor-only intrinsic equation). *Every predictor $\Sigma^{vu} \in \mathbb{R}^{d_v \times d_u}$ has a unique covariance on $\mathcal{M}_\alpha$, given by*

$$\Sigma = \begin{pmatrix} ((\Sigma^{vu})^\top \Sigma^{vu} + \alpha^2 \mathrm{Id}_{d_u})^{1/2} & (\Sigma^{vu})^\top \\ \Sigma^{vu} & (\Sigma^{vu}(\Sigma^{vu})^\top + \alpha^2 \mathrm{Id}_{d_v})^{1/2} \end{pmatrix}. \tag{88}$$

On this leaf, the Bures dynamics close on the predictor:

$$\dot{\Sigma^{vu}} = -E((\Sigma^{vu})^\top \Sigma^{vu} + \alpha^2 \mathrm{Id})^{1/2} - (\Sigma^{vu}(\Sigma^{vu})^\top + \alpha^2 \mathrm{Id})^{1/2} E, \qquad E := \nabla \ell(\Sigma^{vu}). \tag{89}$$

For $d_u = d_v = 1$, this is $\dot{\Sigma^{vu}} = -2\sqrt{(\Sigma^{vu})^2 + \alpha^2}\, \ell'(\Sigma^{vu})$.

Proof. Expanding (87) gives

$$(\Sigma^u)^2 - (\Sigma^{vu})^\top \Sigma^{vu} = \alpha^2 \mathrm{Id}, \quad (\Sigma^v)^2 - \Sigma^{vu}(\Sigma^{vu})^\top = \alpha^2 \mathrm{Id}, \quad \Sigma^{vu}\Sigma^u = \Sigma^v \Sigma^{vu}.$$

Positivity selects the positive square roots. Conversely, an SVD proves the intertwining identity and gives eigenvalues α and $\sqrt{\sigma_i^2 + \alpha^2} \pm \sigma_i > 0$, proving membership in $\mathcal{M}_\alpha$. Substituting the diagonal blocks into (24) gives the intrinsic equation. □

The same eigenvalue calculation gives the exact extremal envelopes

$$\lambda_{\max}(\Sigma) = \sqrt{\|\Sigma^{vu}\|_{\mathrm{op}}^2 + \alpha^2} + \|\Sigma^{vu}\|_{\mathrm{op}}, \qquad \lambda_{\min}(\Sigma) = \frac{\alpha^2}{\lambda_{\max}(\Sigma)}.$$

In the general convergence theorems this special spectral level has $a_+(\Sigma_0) = a_-(\Sigma_0) = \alpha$. Thus its loss exponent is $4\alpha\kappa$, and the GD contraction factor is $1 - \sqrt{2}\alpha\kappa\tau$.

Finite-width form. The covariance constraint has an exact finite-width form that depends only on the product $WW^\top$. For $\Sigma = WW^\top/n$, condition (87) is $WW^\top JWW^\top = n^2\alpha^2 J$. In terms of U, V, it reads

$$(UU^\top)^2 - UV^\top VU^\top = n^2\alpha^2 \mathrm{Id}_{d_u}, \qquad (VV^\top)^2 - VU^\top UV^\top = n^2\alpha^2 \mathrm{Id}_{d_v},$$
$$VU^\top UU^\top = VV^\top VU^\top.$$

Any covariance in (88) admits empirical factors at every width $n \geq m$, for example $W = \sqrt{n}\,\Sigma^{1/2} Q_0$ with $Q_0 Q_0^\top = \mathrm{Id}_m$.

Comparison with relaxed balance. Marcotte et al. (2026, Theorem 3.8) derive intrinsic dynamics under a different condition, $\Gamma_{\mathrm{rb}} := (V^\top V - U^\top U)/n = \beta \mathrm{Id}_n$, on a full-column-rank stratum. It implies

$$(\Sigma^u)^2 + \beta\Sigma^u = (\Sigma^{vu})^\top \Sigma^{vu}, \qquad (\Sigma^v)^2 - \beta\Sigma^v = \Sigma^{vu}(\Sigma^{vu})^\top.$$

For $\beta \neq 0$, zero singular directions have ambiguous branches. Where both factors have full column rank, the range projectors Π_u, Π_v of $(\Sigma^{vu})^\top, \Sigma^{vu}$ select

$$\Sigma^u = \Pi_u[-\tfrac{\beta}{2}\mathrm{Id} + ((\Sigma^{vu})^\top \Sigma^{vu} + \tfrac{\beta^2}{4}\mathrm{Id})^{1/2}],$$
$$\Sigma^v = \Pi_v[\tfrac{\beta}{2}\mathrm{Id} + (\Sigma^{vu}(\Sigma^{vu})^\top + \tfrac{\beta^2}{4}\mathrm{Id})^{1/2}].$$

If $n = d_u = d_v$ and both factors are invertible, the projectors are identities and the predictor equation coincides with (89) for $\alpha = |\beta|/2$. The leaves remain distinct: for $\beta \neq 0$, relaxed balance is equivalent to $\Sigma J \Sigma = -\beta\Sigma$, $\operatorname{rank}\Sigma = n$. Indeed a left inverse of $W/\sqrt{n}$ proves the converse. It forces $n \leq d_v$ if $\beta > 0$, and $n \leq d_u$ if $\beta < 0$; positive HI instead has rank m. For $\beta = 0$, positivity selects the unshifted square roots without a rank assumption, giving the boundary equation $\alpha = 0$, not a positive-definite initialization. Without the rank qualifications, relaxed balance does not ensure predictor closure: $n = 1, d_u = d_v = 2, U = 0, V = e_1$ or e_2 gives the same predictor and $\beta = 1$, but different predictor velocities for $\ell(S) := \|S + \mathrm{Id}_2\|_{\mathrm{F}}^2/2$.

Remark G.3 (Conserved values do not generally close the predictor). *For $d_u = d_v = 2$, take $\Sigma^{vu} = 0$, $\Sigma^v = \operatorname{diag}(3,4)$, and $\Sigma^u = \operatorname{diag}(1,2)$ or $\operatorname{diag}(2,1)$. Both states have signed spectrum $\{1,2,-3,-4\}$, hence the same spectral laws. For $\ell(S) := \|S + e_1 e_1^\top\|_{\mathrm{F}}^2/2$, their predictor velocities are $-4e_1e_1^\top$ and $-5e_1e_1^\top$. The special two-point spectrum gives an algebraic graph unavailable at these general spectral values.*

G.2 A single phase spectral scale: conformal symplectic isotropy

A special conformal spectral identity eliminates the velocity covariance within a positive-position chart of the phase flow.

Definition G.4 (Conformal symplectic isotropy). *A phase covariance is conformally symplectically isotropic (CSI) at scale $\theta \geq 0$ if*

$$\Sigma J_{\mathrm{c}} \Sigma = \theta^2 J_{\mathrm{c}}. \tag{90}$$

For $\theta > 0$, this forces $\Sigma \succ 0$ and is equivalent to $\Sigma^{1/2} J_{\mathrm{c}} \Sigma^{1/2}$ having m copies of each eigenvalue $+i\theta, -i\theta$. The renormalized trace powers in Proposition 5.1 are zero at odd orders and $2m(-1)^k\theta_0^{2k}$ at order $2k$. Conformal transport preserves (90) with $\theta_t = e^{-\gamma t}\theta_0$, including the singular boundary $\theta_0 = 0$, which is specified algebraically, without diagonalizability.

Proposition G.5 (Reduced phase dynamics). *On the chart $A := \Sigma^{(u,v)} \succ 0$, condition (90) is equivalent to*

$$D = C = C^\top, \qquad Q = \theta^2 A^{-1}, \qquad \Xi = AC, \quad \Sigma^{(p,q)} = CAC + \theta^2 A^{-1}.$$

The reduced equations are

$$\dot{A} = CA + AC, \qquad \dot{C} = -G - \gamma C - C^2 + \theta_t^2 A^{-2}, \qquad \dot{\theta}_t = -\gamma\theta_t. \tag{91}$$

They use only the two symmetric matrices A, C, with known scale θ_t, instead of all blocks of the phase covariance.

Proof. Substitute $\Xi = AD^\top$, $\Sigma^{(p,q)} = DAD^\top + Q$. The upper-left block of (90) is $A(D - D^\top)A = 0$, forcing symmetry. The upper-right block then becomes $AQ = \theta^2\mathrm{Id}$; the other blocks follow. Substitution in Proposition C.23 gives (91). This calculation includes $\theta = 0$. □

White position and velocity blocks with $\Xi_0 = 0$ give $\theta_0 = \sigma_{vu}\sigma_{qp}$. Zero velocity gives $C_0 = 0, \theta_0 = 0$, allowing any $A_0 \succ 0$. Neither choice is needed for convergence: the general theorem controls D by energy even when it is nonsymmetric and Q is independent.

Remark G.6 (What the reduced state still contains). *Even on CSI, the two cross blocks Σ^{vu}, Σ^{qp} do not generally form a closed state. With $\theta = 1, C = 0$, choose $A = \mathrm{Id}_2$ or $2\mathrm{Id}_2$ and $\Sigma^{(p,q)} = A^{-1}$. Both have $\Sigma^{vu} = \Sigma^{qp} = \dot{\Sigma}^{vu} = 0$, yet $\ddot{\Sigma}^{vu} = -2E$ or $-4E$ for the same nonzero scalar loss gradient. Thus the full position block remains part of the intrinsic variables.*

H A NONQUADRATIC EXAMPLE: SMOOTHED LASSO

To illustrate the smooth PL loss class beyond quadratic objectives, we repeat the three experiments of Figure 2 with a smoothed entrywise ℓ_1 penalty on the predictor:

$$\ell(S) = \frac{1}{2}\|S - S_{\text{teach}}\|_{\text{F}}^2 + \lambda \sum_{i,j} \left(\sqrt{S_{ij}^2 + \delta^2} - \delta\right), \qquad \lambda = 0.1, \quad \delta = 0.01. \tag{92}$$

The Hessian is diagonal, with entries $1 + \lambda\delta^2/(S_{ij}^2 + \delta^2)^{3/2} \in [1, 1 + \lambda/\delta]$. Thus ℓ is globally 1-strongly convex and satisfies Assumption 1.1 with $\kappa = 1$ and $L = 11$. The penalty acts on S, not separately on the factors, so the covariance closure and spectral conservation laws apply without modification. Its unique minimizer S_{reg} is generally different from the teacher; we compute it entrywise from

$$(S_{\text{reg}})_{ij} - (S_{\text{teach}})_{ij} + \lambda \frac{(S_{\text{reg}})_{ij}}{\sqrt{(S_{\text{reg}})_{ij}^2 + \delta^2}} = 0.$$

All loss gaps below are relative to $\ell_\star = \ell(S_{\text{reg}})$. Positive smoothing does not generally yield exact sparsity; the nonsmooth limit $\delta = 0$ is outside Assumption 1.1.

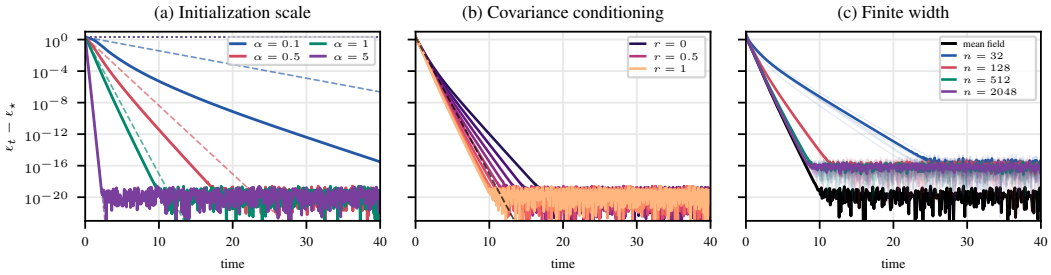

Figure 6: Smoothed-Lasso counterpart of Figure 2, using the same dimensions $(d_u, d_v) = (16, 9)$, teacher, initialization scales, and sampling seeds. **(a)** White covariance $\Sigma_0 = \alpha \text{Id}_m$ for $\alpha = 0.1, 0.5, 1, 5$. Dashed curves are the support-gap envelopes $\Delta_0 e^{-4\alpha\kappa t}$; dotted curves use the internal coarse determinant bound of Remark C.7. **(b)** One input-output diagonal pair is scaled by $r \in \{0, 1/8, \ldots, 1\}$. Here $g(\Sigma_0) = 2r$; the dashed curve shows the white-reference envelope. **(c)** The population covariance flow from Id_m and finite-width flows at widths $n = 32, 128, 512, 2048$, with five independent trials each.