THEORY OF FEATURE FORMATION, EXPRESSION AND REORGANIZATION IN QUADRATIC NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

A neural network learns a task by forming useful features and learning how strongly to use them in its predictions. Classical exact mode solutions for deep linear networks describe how expression grows from weights already aligned with the task, leaving the formation of those features unaddressed. We derive how feature directions form from unaligned weights in a quadratic network and how growing strengths affect further direction learning. Under population gradient flow, we solve the early direction dynamics in the limit of small initialization across a broad class of input distributions. The network acquires a hierarchy of directions set by the task and data while its output remains small, with alignment timescales governed by differences in growth rates. For isotropic Gaussian inputs, the direction law is exact throughout training and does not require small initialization. We quantify silent alignment feature by feature, showing that smaller initialization delays substantial use of each feature without slowing its formation. For more general inputs, growing feature strengths can feed back on direction learning through the changing prediction error and reorganize the features formed early in training. Exact solutions for structured input distributions show how anisotropy and fourth moments govern this feedback-induced reorganization: the network can replace a feature acquired early or mix it with another direction, or features can promote or suppress one another's growth. The resulting theory connects the formation of a representation to its subsequent development as the network learns to use it.

1 INTRODUCTION

A neural network learns a task by forming useful task-relevant features and learning how strongly to use them in its predictions (Bengio et al., 2013). Understanding feature learning therefore requires studying both feature formation and expression, and the interplay between them. How are the two organized in time during learning, and how does each shape the other? For example, it has been observed that learning can produce a "silent alignment" (Atanasov et al., 2022), in which a useful feature need not produce an immediate substantial improvement in prediction, because the network may still make little use of it. In turn, as the network makes greater use of the formed features during training, its predictions change the error that drives further learning, which could feed back and reorganize the formed features. In this work, we aim to follow the life cycle of a feature through its formation, expression and its subsequent reorganization as the network learns to use it.

Classic works have solved the learning dynamics of deep linear networks mode by mode, explaining learning plateaus and stage-like transitions from task-aligned initial weights (Saxe et al., 2014; 2019; Lampinen & Ganguli, 2019), and more generally as matrix dynamics from generic initialization in two-layer networks (Fukumizu, 1998; Braun et al., 2022; Dominé et al., 2025). But we still lack an exact, feature-resolved account of learning dynamics that tells us which features the network represents at each moment, in what order they form, how long each takes to form, and how long before each is used by the network. Furthermore, these exactly solvable theories explicitly work in the regime of linear networks, which have a highly simplified function class and are insensitive to higher-order statistics in the learning problem, which are known to shape learned features (Ingrosso & Goldt, 2022; Lufkin et al., 2024).

Here, we study quadratic neural networks, which have recently emerged as a compelling nonlinear solvable model of learning and have been used to study signal recovery and overparameterization (Li

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.

et al., 2018; Sarao Mannelli et al., 2020; Martin et al., 2024; Erba et al., 2025), high-dimensional gradient flow (Martin et al., 2026), scaling laws (Ben Arous et al., 2025; Defilippis et al., 2026), loss landscapes (Arjevani et al., 2026), and normal form approximations to small-initialization dynamics across architectures (Ziyin et al., 2026). The quadratic network is the simplest nonlinear model whose learning is sensitive to higher-order statistics of the inputs, while remaining analytically tractable. Our results are as follows.

- We show that for generic initialization and isotropic Gaussian inputs, the dynamics of the subspace spanned by the learned features can be solved exactly. Features form in a hierarchy ordered by their strength in the target, each at a time set by an eigenvalue gap and independent of initialization scale, leading to a feature-specific window of silent alignment at small initialization (Section 3).
- We show that at small initialization, early feature formation for general inputs follows the same dynamics as in the isotropic Gaussian case, with the target reshaped by the input statistics. The input covariance and fourth moments, together with the target, therefore decide which features form first and in what order (Section 4).
- We show that as the network starts to use its features, its own predictions change what it learns next and can reorganize the features formed early (Section 5). In exactly solvable cases, unequal input variances make a later feature replace an early one, kurtosis mixes two feature directions into one, and spherical fourth moments make feature strengths compete or cooperate while preserving their directions. As a whole, the features that form first need not be the ones stabilized at convergence.

2 Model and its direction–amplitude decomposition

We first define the model and derive its population gradient flow, then split the dynamics into feature formation and expression. Suppose the input to the model $x \in \mathbb{R}^d$ follows an arbitrary mean-zero distribution with covariance $\Sigma \succ 0$ and finite fourth moments. We study a teacher–student setup (Seung et al., 1992; Saad & Solla, 1995): a student network learns a target generated by a teacher network of the same form. The student is a two-layer network of width $k \leq d$ with quadratic activation and unit output weights, centered to have zero mean output (Li et al., 2018; Sarao Mannelli et al., 2020; Martin et al., 2024; Ben Arous et al., 2025; Martin et al., 2026),

$$f_W(x) = \sum_{j=1}^{k} \left[(w_j^\mathsf{T} x)^2 - w_j^\mathsf{T} \Sigma w_j\right] = x^\mathsf{T} M x - \mathrm{tr}(M\Sigma), \tag{1}$$

with weights $W = [w_1, \ldots, w_k] \in \mathbb{R}^{d\times k}$ and the learned effective matrix $M = WW^\mathsf{T} \succeq 0$. The teacher is $f_\star(x) = x^\mathsf{T} M_\star x - \mathrm{tr}(M_\star \Sigma)$, with $M_\star \succeq 0$ of rank r, which need not equal k. We will study the dynamics of gradient flow to minimize the population squared loss $\mathcal{L}(W) = \frac{1}{2}\mathbb{E}[(f_W(x) - f_\star(x))^2]$.

To derive the dynamics, let $R_t = M_\star - M_t$ be the current residual matrix. The prediction error can be written in the same centered quadratic form: $f_\star(x) - f_{W_t}(x) = x^\mathsf{T} R_t x - \mathrm{tr}(R_t\Sigma)$. We define the time-dependent growth matrix, which drives learning, as the average of the prediction error against the centered input outer product:

$$H_t = \frac{1}{2}\mathbb{E}\left[\left(f_\star(x) - f_{W_t}(x)\right)\left(xx^\mathsf{T} - \Sigma\right)\right] = H_\star - H_{M_t}. \tag{2}$$

For isotropic Gaussian inputs it is simply the residual, $H_t = R_t$, the part of the teacher not yet learned; in general, it is this residual reshaped by the input covariance and fourth moments (Section 4). The teacher contribution $H_\star$, which is defined as the same average as above but with $f_\star$ in place of the error, is fixed, while the student contribution H_{M_t}, the average with f_{W_t}, changes as the network learns to predict. After absorbing the constant factor in the gradient into time, the exact population flow of the learning dynamics is given by the closed matrix differential equation (Theorem A.1; see also Martin et al., 2024; Ben Arous et al., 2025)

$$\dot{W}_t = H_t W_t, \qquad \dot{M}_t = H_t M_t + M_t H_t. \tag{3}$$

The matrix H_t is symmetric. So, its eigenvectors give the instantaneous preferred directions, and its eigenvalues give the growth rates of weights along them, hence its name. We would like to decompose and follow feature formation and expression separately, that is, which input directions the network has learned and how strongly it uses them (Ben Arous et al., 2025). But orthogonal

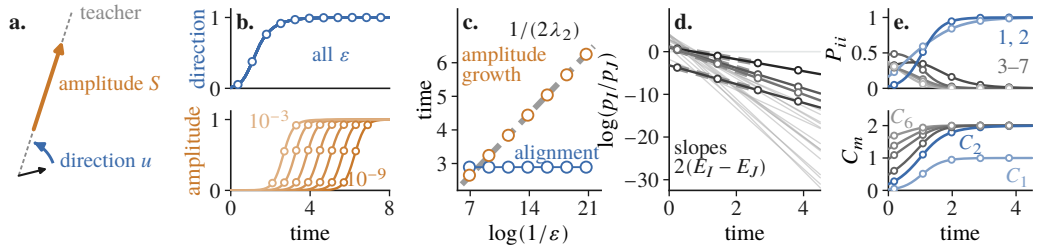


Figure 1: **Feature directions form before they are expressed, in a hierarchy set by the teacher's eigenvalues.** (a) The direction u turns toward the teacher before the amplitude S grows. (b) Smaller initialization delays amplitude growth; the direction curves coincide. (c) The alignment time is constant; the time to half the weakest active amplitude grows with slope $1/(2\lambda_2)$. (d) Log odds of every overlap against $J = \{1, 2\}$ are straight, with predicted slopes (dashes). (e) Individual captures P_{ii} can rise and fall; cumulative captures C_m only rise. Lines: gradient flow; circles: exact solution.

rotation of the hidden-unit columns leaves M and the learned function unchanged. We therefore describe the subspace spanned by the weights and the function expressed within it as follows. For linearly independent columns of W at initialization, write $M = USU^\mathsf{T}$ with $U \in \mathbb{R}^{d\times k}$, $U^\mathsf{T}U = I_k$ and $S \in \mathbb{R}^{k\times k}$, $S \succ 0$. The columns of U form an orthonormal basis for the student subspace, and $P = UU^\mathsf{T}$ is its projector. The positive matrix S, which we call the amplitude matrix, specifies the function expressed within that subspace (Figure 1a). We choose $U^\mathsf{T}\dot{U} = 0$ to remove arbitrary rotation among the basis vectors, so $\dot{U}$ records only motion of the subspace itself. This decomposition separates motion of the student subspace from changes in the function expressed inside it and leads to the following direction and amplitude dynamics (Lemma A.2):

$$\dot{U} = (I - P)H_tU, \qquad \dot{S} = (U^\mathsf{T}H_tU)S + S(U^\mathsf{T}H_tU). \tag{4}$$

We first focus on direction learning. Only the component of H_tU perpendicular to the current subspace can rotate it. Using the split $H_t = H_\star - H_{M_t}$, we can see separately how the teacher and the student's own prediction contribute to this motion:

$$\dot{U} = \underbrace{(I - P)H_\star U}_{\text{teacher drive}} - \underbrace{(I - P)H_{M_t}U}_{\text{student feedback}}. \tag{5}$$

The teacher contribution to equation (3) is linear in M_t. Since f_{W_t}, and hence H_{M_t}, is linear in M_t, the student contribution is quadratic. At small amplitude the output f_{W_t} is near zero, so the teacher drive dominates. As the amplitudes grow, the student feedback becomes important, and its perpendicular component can change the directions acquired early. The split is exact (Lemma A.2): we solve both parts exactly for isotropic inputs (Section 3), then treat the teacher drive (Section 4) and the student feedback (Section 5) for general inputs.

3 Exact feature formation with isotropic inputs

For isotropic Gaussian inputs, the full learning dynamics, along with the feature formation dynamics, can be solved exactly. So, assume $x \sim \mathcal{N}(0, I)$. Then $H_t = R_t = M_\star - M$, so $\dot{M} = M_\star M + MM_\star - 2M^2$. This is the well-known matrix Riccati ODE, and its solution appears in earlier analyses of quadratic networks (Martin et al., 2024; Ben Arous et al., 2025; Ziyin et al., 2026). The matrix solution alone, however, does not show which directions the student acquires, in what order and how quickly, and how long before each is expressed. We therefore follow the student subspace through its projector. The student matrix acts within its own range, so $(I - P)MU = 0$. Its contribution therefore disappears from the direction equation, leaving

$$\dot{U} = (I - P)M_\star U, \qquad \dot{P} = (I - P)M_\star P + PM_\star(I - P). \tag{6}$$

The student subspace thus moves as if driven by the teacher alone, whatever its amplitudes.

3.1 THE PROJECTOR IS A NORMALIZED LINEAR FLOW

To follow the student subspace, we study its projector $P(t)$. Take $M_0 = U_0 S_0 U_0^\mathsf{T}$ with $U_0^\mathsf{T} U_0 = I_k$ and $S_0 \succ 0$. Since motion inside the subspace changes only its basis and not its span, the subspace moves exactly as the span of the linear flow $\dot{Y} = M_\star Y$ from $Y(0) = U_0$. The columns of its solution $Y(t) = e^{tM_\star} U_0$ remain linearly independent, and their span gives the exact student subspace (Theorem B.1), whose projector is

$$P(t) = Y(t)(Y(t)^\mathsf{T} Y(t))^{-1} Y(t)^\mathsf{T}. \tag{7}$$

Equation 7 can be read as a linear evolution of the initial basis, followed by a normalization that removes its changing lengths and mutual overlaps. The result describes only the span and contains no initial amplitude matrix S_0 (Corollary B.2; Figure 1b,c). We note in passing that the projector dynamics in equation (6) are the projector form of Brockett's double-bracket flow (Brockett, 1991). The amplitude dynamics within this subspace have been calculated in concurrent work (Ziyin et al., 2026), and the full matrix solution can be written as $M(t) = Y(t)N(t)^{-1}Y(t)^\mathsf{T}$ with $N(t) = S_0^{-1} + 2\int_0^t Y(s)^\mathsf{T} Y(s)\, ds$ (Theorem B.1). Since $N(t) \succ 0$ (the sum of $S_0^{-1} \succ 0$ and a positive semidefinite integral), $M(t)$ and $Y(t)$ describe the same subspace: the amplitudes change what the network expresses within the subspace, but not the subspace itself.

3.2 SUBSPACE OVERLAPS FOLLOW A REPLICATOR EQUATION

Next, we show that the student's overlaps with teacher subspaces exactly follow a replicator equation, which in evolutionary biology describes selection among competing species (Hofbauer & Sigmund, 1998): teacher subspaces compete like species whose fitness is set by their eigenvalues, and the student's overlap shifts toward the fittest, so teacher directions are acquired in order of their eigenvalues. To see this, we use the projector solution in equation (7) to follow how the student subspace aligns with the teacher's fixed eigenvectors $v_1, \ldots, v_d$, where $M_\star = V \Lambda V^\mathsf{T}$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$. We measure the overlap by projecting a unit volume from the student subspace onto a chosen teacher subspace; at rank two, this is the area of a projected unit square. More generally, choose an index set $I = \{i_1, \ldots, i_k\}$, and let V_I collect the corresponding orthonormal teacher eigenvectors. The squared projected volume is $p_I(t) = \det(V_I^\mathsf{T} U(t))^2$. This overlap equals one when the subspaces coincide and zero when the student subspace contains a direction orthogonal to all the chosen teacher directions. The overlaps sum to one across all choices of I (Appendix C), so they are normalized weights describing the student's alignment. To calculate how these overlaps change, we use $Y(t) = e^{tM_\star} U_0$, which spans the same student subspace but is not orthonormal. Along teacher direction v_i, its coordinates grow as $e^{\lambda_i t}$. The volume projected onto the directions in I therefore grows as $e^{E_I t}$, where $E_I = \sum_{i \in I} \lambda_i$. Dividing these squared volumes by their sum removes the overall volume growth of Y and recovers the same normalized overlaps, $p_I(t) \propto p_I(0) e^{2E_I t}$. Every pairwise log odds is therefore linear in time, and differentiating gives the replicator equation for the overlap weights (Theorem C.1):

$$\log \frac{p_I(t)}{p_J(t)} = \log \frac{p_I(0)}{p_J(0)} + 2(E_I - E_J)t, \qquad \dot{p}_I = 2(E_I - \bar{E}) p_I, \tag{8}$$

where $\bar{E}(t) = \sum_J p_J(t) E_J$ is the overlap-weighted mean score. The differences between growth scores set the slopes of the log odds (Figure 1d). The student's overlap increases with teacher eigensubspaces whose scores exceed the current mean and decreases with those whose scores fall below it. When $\lambda_k > \lambda_{k+1}$, the leading k-dimensional teacher subspace has the largest score. From a generic initial subspace, its overlap therefore tends to one.

3.3 FORMATION VERSUS EXPRESSION TIMESCALES AND SILENT ALIGNMENT

We now ask how long the student takes to form the leading teacher directions and to express them, and how the two timescales compare. First, a simple way to measure formation would be the capture of each teacher direction v_i, its squared projection onto the student subspace, $P_{ii}(t) = \|U(t)^\mathsf{T} v_i\|^2 = \sum_{I \ni i} p_I(t)$. But an individual capture can transiently rise and later fall as the student turns toward stronger teacher directions, so it does not define a formation time. Instead, we show that the cumulative capture $C_m(t) = \sum_{i=1}^m P_{ii}(t)$ of the first m directions cannot decrease (Proposition C.2; Figure 1e), which gives a well-defined timescale. Take the student rank k below the teacher rank, with

$\lambda_k > \lambda_{k+1}$, and consider the first $m \leq k$ directions. Their remaining capture error is $m - C_m(t)$. The closest competing subspace that misses one of them replaces direction m with the strongest excluded direction, $k+1$. Its score is lower by $\lambda_m - \lambda_{k+1}$, which sets the decay rate of the error, $m - C_m(t) \sim B_m e^{-2(\lambda_m - \lambda_{k+1})t}$. Here $B_m > 0$ is set by the initial subspace overlaps and is independent of the initialization scale (Appendix C.1). The error reaches a small tolerance η at the alignment time $t_m^{\text{align}}(\eta) \approx \log(B_m/\eta)/(2(\lambda_m - \lambda_{k+1}))$, which sets the formation timescale of these directions. The final approach to the leading k-dimensional subspace is therefore controlled by $\lambda_k - \lambda_{k+1}$. A small gap makes this last part of alignment slow.

Unlike the formation timescale, the expression timescale depends on the initialization scale. To see this, write $S_0 = \varepsilon \bar{S}_0$ with fixed $\bar{S}_0 \succ 0$. For a learned matrix diagonal in the teacher basis (for any initial subspace, $v_j^\mathsf{T} M v_j$ grows at most as $e^{2\lambda_j t}$; Appendix C.2), the amplitude s_j of each mode j with $\lambda_j > 0$ grows logistically, as in the classical mode solutions for deep linear networks (Saxe et al., 2014), and reaches a fixed fraction $\alpha \in (0,1)$ of its final value λ_j at a time $t_j(\alpha)$:

$$\dot{s}_j = 2s_j(\lambda_j - s_j), \qquad s_j(t) = \frac{\lambda_j}{1 + (\lambda_j/s_j(0) - 1)e^{-2\lambda_j t}}, \qquad t_j(\alpha) \sim \frac{\log(1/\varepsilon)}{2\lambda_j}. \tag{9}$$

The estimate of $t_j(\alpha)$ holds for $s_j(0) = \Theta(\varepsilon)$ as $\varepsilon \to 0$. Decreasing the initialization therefore delays appreciable amplitude growth by a time proportional to $\log(1/\varepsilon)$, while the alignment time remains fixed (Corollary B.2; see also Atanasov et al., 2022). Figures 1b and 1c show the same separation in the generic matrix solution.

We can now compare the two timescales. A well-separated interval of silent alignment for mode m occurs when

$$t_m^{\text{align}}(\eta) \approx \frac{\log(B_m/\eta)}{2(\lambda_m - \lambda_{k+1})} \quad \ll \quad \frac{\log(1/\varepsilon)}{2\lambda_m} \sim t_m(\alpha). \tag{10}$$

For any fixed positive eigenvalue gap, sufficiently small initialization therefore guarantees this separation. The student has then acquired the leading m directions before amplitude m becomes appreciable, even if stronger modes are already expressed (Appendix C.2). For the whole subspace to align while the entire output remains small, the last alignment time must precede growth of mode 1 (Appendix C.3); the loss then stays nearly flat while the subspace aligns (Figure A1).

4 Feature formation under data anisotropy and higher moments

With isotropic Gaussian inputs, the growing amplitudes are decoupled from the student subspace dynamics. We now restore general covariance and fourth moments, which change what the student learns first while the output is small and can turn the subspace and reorganize features once the amplitudes grow.

Covariance and fourth moments determine the growth matrix. For general inputs, the growth matrix depends on the input covariance and, for non-Gaussian inputs, also on the input fourth moments (see also Ingrosso & Goldt, 2022; Lufkin et al., 2024). For Gaussian inputs, by Wick's theorem, these moments are sums of products of covariances (Isserlis, 1918). The fourth cumulant, then, is the difference from that Gaussian value, $\kappa_{ijab} = \mathbb{E}[x_i x_j x_a x_b] - \Sigma_{ij}\Sigma_{ab} - \Sigma_{ia}\Sigma_{jb} - \Sigma_{ib}\Sigma_{ja}$. To calculate its contribution to growth, we weight the cumulant by the current residual entries and write the resulting matrix as K_{R_t}, with $(K_{R_t})_{ij} = \sum_{a,b} \kappa_{ijab}(R_t)_{ab}$. Expanding the growth-matrix average then gives the following exact growth matrix

$$H_t = \Sigma R_t \Sigma + \frac{1}{2} K_{R_t}. \tag{11}$$

The first term is the Gaussian contribution with covariance Σ, and the second accounts for the difference in fourth moments. The teacher and student contributions defined in Section 2 are therefore $H_\star = \Sigma M_\star \Sigma + \frac{1}{2} K_{M_\star}$ and $H_{M_t} = \Sigma M_t \Sigma + \frac{1}{2} K_{M_t}$. The loss depends on the same moments, $\mathcal{L}(M_t) = \operatorname{tr}(R_t \Sigma R_t \Sigma) + \frac{1}{2}\operatorname{tr}(R_t K_{R_t})$. Since we study a quadratic network, no moment above order four enters either the loss or the gradient flow (Theorem A.1; see also Arjevani et al., 2026). Input distributions with matching covariance and fourth moments therefore give the same population dynamics from the same initialization.

Universal early-time replicator dynamics. While the student's amplitudes remain small, its contribution H_{M_t} is small, and the direction equation reduces to its teacher drive, $\dot{U} \simeq (I - P) H_\star U$,

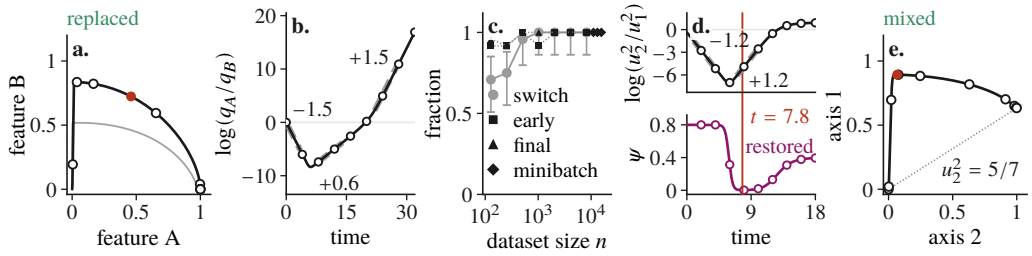


Figure 2: **Growing amplitudes replace or mix the first feature, and the replacement survives finite data.** (a) The covariance weight path turns from feature B to feature A; gray, one finite dataset. (b) Its log odds follow the three predicted slopes. (c) The switch becomes more frequent as the dataset grows; minibatch runs also switch. (d) Kurtosis log odds over the first-axis residual ψ, which vanishes and is then restored. (e) The kurtosis weight path ends at a mixed direction. Red: the switch and the zero residual; circles: independent closures.

with a projector error that we prove is at most of the order of the amplitudes, uniformly for times up to order $\log(1/\varepsilon)$ (Appendix D). The matrix $H_\star$ thus takes the place of $M_\star$ in the isotropic direction equation and acts as an effective teacher for early direction learning. It is the correlation of the target with the input outer product: it equals $M_\star$ for isotropic Gaussian inputs, and the input covariance and fourth moments reshape it otherwise. To obtain the student subspace dynamics, we evolve the initial basis with $e^{tH_\star}$ and take its span, as in Section 3. The subspace-overlap solution from Section 3 now describes the early hierarchy using the eigenvectors and eigenvalues of $H_\star$.

5 Finite-amplitude feedback reorganizes features

Later in the learning dynamics, as the student's amplitudes grow, its output changes the prediction error that drives learning. The student contribution to the growth matrix can then no longer be neglected. Restoring it gives the direction equation $\dot{U} = (I-P)H_\star U - (I-P)\big(\Sigma M \Sigma + \frac{1}{2}K_M\big)U$. The first term is the familiar early teacher-driven motion. The second, however, describes how the function already expressed by the student changes that motion. For isotropic Gaussian inputs, the student contribution is M, and MU lies within the student subspace, so it produces no perpendicular motion. But with general covariance, $\Sigma M \Sigma U$ need not lie within that subspace, and neither need the fourth-moment term unless the fourth moments are spherical (Appendix E). The resulting perpendicular contribution can then reorganize the representation. We now follow the student-feedback term of equation (5) through the full trajectory in three exactly solvable input families.

5.1 Feature replacement due to anisotropic covariance

For Gaussian inputs, unequal variances alone can make a feature learned early give way to another (see also Atanasov et al., 2022). Take $\Sigma = \mathrm{diag}(d_1, \ldots, d_d)$ and $M_\star = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ in the same basis, and consider a rank-one student $M = ww^\mathsf{T}$. Let $q_i = d_i w_i^2$ be its squared weight on axis i, weighted by the input variance, and let $r = \sum_i q_i$ be their total. The weight dynamics then simplify to:

$$\dot{q}_i = 2d_i(d_i\lambda_i - r)q_i. \tag{12}$$

Initially, at small r, q_i grows at the fractional rate $2g_i$, where $g_i = d_i^2\lambda_i$. But as the total r grows, it slows growth on every axis, more strongly where the input variance is larger. Growth along axis i stops when r reaches $\widetilde{\lambda}_i = d_i\lambda_i$, so unequal variances can favor different axes early and late.

Among the initially present features, suppose B has the largest g_i and A has the largest $\widetilde{\lambda}_i$. From sufficiently small initial squared weights of comparable size, B grows first. As B approaches its plateau at $r = \widetilde{\lambda}_B$, A still grows because $\widetilde{\lambda}_A > \widetilde{\lambda}_B$. Then, A's growth raises r above B's threshold, and B starts to shrink. The stable surviving feature is A, with $r = \widetilde{\lambda}_A$ (Appendix F). The log ratio of q_A to q_B obeys $\frac{d}{dt}\log(q_A/q_B) = 2(g_A - g_B) - 2(d_A - d_B)r$, which is linear in r. At $r \simeq 0$, $r = \widetilde{\lambda}_B$ and $r = \widetilde{\lambda}_A$ it gives the predicted slopes $2(g_A - g_B)$, $2d_A(\widetilde{\lambda}_A - \widetilde{\lambda}_B)$ and $2d_B(\widetilde{\lambda}_A - \widetilde{\lambda}_B)$.

The slope changes from negative to positive and then steepens as A replaces B. This explains the weight path in Figure 2a, and the log odds follow all three slopes in Figure 2b. Moreover, the full trajectory can be solved exactly through a closed scalar equation (Proposition F.1), whose solution gives the open circles in Figure 2a,b. Empirically, we also observe this switch in finite datasets and under minibatch training (Figure 2c).

5.2 Directional kurtosis produces a mixed feature

Directional fourth moments can turn an initially acquired feature into a mixture of teacher directions, even when the covariance is the identity. As before, consider a rank-one student learning two teacher components, with the second stronger than the first, $\lambda_2 > \lambda_1$. Set $\Sigma = I$ and give only the first teacher axis v_1 positive excess kurtosis κ. The larger fourth moment increases the loss penalty for leaving this component unfitted. The student's weight along the first axis initially grows at rate $(1+\kappa/2)\lambda_1$, compared with λ_2 along the second. When the first rate is larger, the student approaches the first axis from sufficiently small, comparable initial components. But the preference for the first axis weakens as the student learns that component. To see that, write $M = muu^\mathsf{T}$, where m is the amplitude and u is a unit direction with components $u_i = v_i^\mathsf{T} u$ along the teacher axes. The first-axis residual is $\psi(t) = \lambda_1 - mu_1^2$, and one can show that the direction changes according to

$$\frac{d}{dt}\log\frac{u_2^2}{u_1^2} = 2(\lambda_2 - \lambda_1) - \kappa\psi(t). \tag{13}$$

A positive slope means that the student turns toward the second axis. The difference $\lambda_2 - \lambda_1$ favors this motion, while the first-axis residual opposes it through the kurtosis term. Initially $\psi \simeq \lambda_1$, giving the negative slope $2[\lambda_2 - (1+\kappa/2)\lambda_1]$. Learning the first component reduces ψ and weakens its advantage. With well-separated learning stages, the residual temporarily approaches zero and the slope approaches $2(\lambda_2 - \lambda_1) > 0$ (Figure 2d). But abandoning the first component would restore the error that made it favorable. The student therefore retains both components, settling at a mixed direction where their growth rates balance:

$$m^\star = \lambda_2, \qquad (u_1^\star)^2 = \frac{(1+\kappa/2)\lambda_1 - \lambda_2}{(\kappa/2)\lambda_2}. \tag{14}$$

The student thus represents one feature with components along both teacher axes (Figure 2e; Appendix F.1). The full trajectory is solvable for any student rank, because the teacher axes remain fixed and the isotropic matrix solution carries over (Proposition F.2; Figure 2d,e).

5.3 Cooperative and competitive amplitude dynamics via spherical moments

Here, we give a different solvable example with spherical fourth moments, where the direction dynamics are preserved after rescaling time while the amplitudes become coupled. Consider inputs with identity covariance and fourth moments that are the same in every direction. Write their common fourth moment as $\mathbb{E}[(a^\mathsf{T}x)^4] = 3(1+\bar\kappa)$ for any unit vector a. The growth matrix and, after rescaling time, the direction equation become

$$H_t = (1+\bar\kappa)R_t + \tfrac{\bar\kappa}{2}\operatorname{tr}(R_t)I, \qquad \frac{dU}{d\tau} = (I-P)M_\star U, \qquad \tau = (1+\bar\kappa)t, \tag{15}$$

because the identity term cannot turn the subspace. The student therefore follows the same sequence of subspaces as under Gaussian inputs, at a different speed (Figure 3a).

The amplitudes, however, no longer grow independently: each feature's growth depends on the total amplitude still unlearned across all features. Write this shared residual as $\theta = \operatorname{tr}(M_\star - M)$, the total teacher amplitude minus the total student amplitude. For a student diagonal in the teacher basis, the amplitude dynamics are

$$\frac{ds_j}{d\tau} = 2s_j(\lambda_j - s_j + c\theta), \qquad c = \frac{\bar\kappa}{2(1+\bar\kappa)}. \tag{16}$$

Growth of any feature reduces θ, so learning one feature changes the growth of the others (Figure 3b). For $c > 0$, a growing feature reduces the shared positive contribution to growth. The features therefore compete. In Figure 3c, early amplitudes rise above their final values, then decrease as later

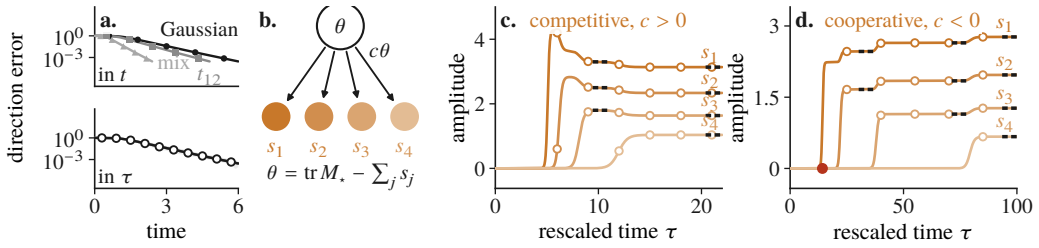


Figure 3: **Spherical fourth moments preserve the rescaled direction path while making amplitudes compete or cooperate.** (a) Gaussian, t_{12} and scale-mixture inputs separate in time t and coincide in rescaled time τ; circles: Gaussian solution; gray bands: minibatch runs. (b) The shared residual θ enters every amplitude's growth rate. (c) For $c > 0$, each new mode lowers the earlier amplitudes. (d) For $c < 0$, each raises them; s_4 starts to grow where $\lambda_4 + c\theta$ crosses zero (red). Dashes: stage plateaus; circles: independent closure.

features grow. For $c < 0$, growth instead weakens a shared inhibition, so the features cooperate. Earlier features help later ones grow in Figure 3d. A small mode suppressed by $\lambda_j + c\theta < 0$ begins growing when θ falls below $\lambda_j/|c|$.

We can calculate the equilibria underlying the successive amplitude plateaus exactly. Suppose the first m modes have reached appreciable amplitude while the others remain small, and let $R_m = \sum_{i>m} \lambda_i$ be the remaining teacher trace. Holding those other modes at zero gives the equilibrium values $\theta^{(m)} = R_m/(1 + mc)$ and $s_j^{(m)} = \lambda_j + c\,\theta^{(m)}$ for $j \le m$. With small initialization and separated learning stages, the full trajectory passes near these equilibria (as in the incremental learning of diagonal linear networks; Berthier, 2023; Pesme & Flammarion, 2023). Each newly expressed mode lowers the earlier plateaus for positive c and raises them for negative c, as the dashed lines in Figure 3c,d show. The full amplitude trajectories are self-consistently solvable (Martin et al., 2024) through a scalar closure (Theorem G.1). Solving this closure independently gives the circles in Figure 3c,d. Spherical fourth moments thus change how strongly features are expressed and introduce nonlinear interaction between them, but do not affect the directions formed during learning.

5.4 Dissociation of early feature formation and eventual reorganization

A feature that grows rapidly early need not be the one that would reduce the loss most, as the covariance example showed. We can understand these differences more generally by comparing how the data weight each feature's contribution to the loss. The comparison can be made explicit for independent input coordinates in the teacher basis, with variances d_a and excess kurtoses $\kappa_a > -2$. For a student diagonal in this basis, write the teacher and student amplitudes as λ_a and m_a, respectively. The loss then separates into

$$\mathcal{L}_{\mathrm{diag}} = \sum_a f_a^2 (\lambda_a - m_a)^2, \qquad f_a = d_a \sqrt{1 + \kappa_a/2}. \tag{17}$$

The factor f_a^2 gives the prediction cost of an amplitude error on coordinate a. A larger variance or excess kurtosis makes the same amplitude error more costly. Omitting feature a costs $f_a^2 \lambda_a^2$, so a student restricted to k features fits best by keeping the k largest $f_a \lambda_a$. Early growth, however, uses a different ranking scheme. In our setting, the initial growth matrix $H_\star$ has diagonal entries $f_a^2 \lambda_a$, giving the initial weight-growth rates. The overall comparison then is

$$\lambda_a \text{ (fitted amplitude)}, \qquad f_a \lambda_a \text{ (loss reduction)}, \qquad f_a^2 \lambda_a \text{ (initial growth)}. \tag{18}$$

The data factor enters with powers zero, one, and two, so the same features can have different rankings (Figure A3). For Gaussian inputs, $f_a = d_a$, recovering the covariance example: feature B grows first, but retaining A removes more loss.

This dissociation has been observed empirically in nonlinear deep networks, which favor features that are easy to extract, such as those with larger input amplitude, over more predictive ones (Hermann & Lampinen, 2020; Hermann et al., 2024). Our ranking gives an exact account of this preference in the

feature-learning regime. Here f_a acts as the availability of feature a and $f_a\lambda_a$ as its predictivity. The loss saved by keeping a feature depends on its predictivity alone, whereas its early growth depends on availability times predictivity, so a more available feature can form first even when it is less predictive. Heavy-tailed directions, whose f_a is larger, are also more available, and in the covariance example feedback later reverses the preference.

6 Related work

Exact learning dynamics and representation alignment. Work on linear networks, whose dynamics from generic initialization follow a solvable matrix Riccati ODE (Fukumizu, 1998), resolves how initialization, balance, and prior knowledge affect both alignment and strength (Woodworth et al., 2020; Braun et al., 2022; Kunin et al., 2024; Dominé et al., 2025). In nonlinear networks, small weights first align with directions correlated with the target, as shown for ReLU networks on finite or orthogonal data and, in the small-initialization limit, for homogeneous networks (Maennel et al., 2018; Boursier et al., 2022; Boursier & Flammarion, 2025; Kumar & Haupt, 2024). Silent alignment and effective representation dynamics describe such changes before substantial loss reduction (Atanasov et al., 2022; Van Rossem & Saxe, 2024), and plateaus can also arise from different learning speeds of the two layers (Berthier et al., 2025). In contrast, we solve alignment exactly in a nonlinear network and feature by feature: each teacher direction has its own alignment time, set by an eigenvalue gap and independent of the initialization scale, and its own silent-alignment window (Section 3).

Subspace learning and matrix factorization. The isotropic projector equation is related to eigenvector sorting and principal-subspace learning under a fixed matrix (Oja, 1982; Brockett, 1991; Yan et al., 1994). The factorization $M = WW^\intercal$ also connects to implicit regularization and incremental low-rank learning in matrix factorization and matrix sensing (Gunasekar et al., 2017; Li et al., 2018; Arora et al., 2019; Li et al., 2020; Stöger & Soltanolkotabi, 2021; Jin et al., 2023), part of a broader picture of staged, saddle-to-saddle learning across architectures (Zhang et al., 2026). In our network, the student subspace follows the projector flow exactly for isotropic inputs, whatever the amplitudes, and at small initialization for general inputs; this separates when each feature forms from the incremental growth of the fit.

Teacher-student theory and data structure. Statistical mechanics analyses of teacher–student learning describe high-dimensional stochastic dynamics through overlaps and other order parameters (Saad & Solla, 1995; Goldt et al., 2019; Arnaboldi et al., 2023b; Ben Arous et al., 2024; Collins-Woodfin et al., 2024), including for quadratic networks (Arnaboldi et al., 2023a; Ben Arous et al., 2025) and for curricula and structured data (Saglietti et al., 2022; Goldt et al., 2022; Braun et al., 2026). In contrast, at finite input dimension our population flow closes exactly on the learned matrix, follows the formation and expression of each feature separately, and depends on the input distribution only through its covariance and fourth moments.

7 Conclusion

In a quadratic network, we have derived exact results that follow each feature from its formation out of unaligned weights, through its expression, to its later reorganization. For isotropic Gaussian inputs, feature directions do not depend on the initial amplitudes, while small initialization delays their growth, so each feature forms before it is used. The same direction law describes the early feature hierarchy across a broad class of input distributions at small initialization, with the target reshaped by the input statistics. But as the amplitudes grow, the learned function changes the prediction error that drives further learning and can reorganize the features formed early. Our analysis is limited to a shallow network with a single trainable layer, and extending it to deeper networks is an important next step. For anisotropic covariance and general fourth moments, our full-trajectory solutions cover specific solvable cases, and the general case remains open. Nevertheless, taken together, our results trace a representation through the stages of learning, explaining why representational features form in a hierarchy, why this can happen while the loss barely changes, and how they can later be reorganized.

AI USE STATEMENT

We used generative AI tools to assist with checking the theory and its proofs, simplifying and improving the derivation, and writing code for simulations that test the theory, as well as editing the text and finding related work. We have reviewed all AI-assisted work and take responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

The appendix states the assumptions and gives complete derivations for the formal claims. Appendix I, together with the numerical checks placed beside the results they test, records the numerical details required to generate each figure and table. Scripts that regenerate the numerical results and figures will be shared with reviewers through an anonymous repository during the discussion period.

REFERENCES

Yossi Arjevani, Joan Bruna, Joe Kileel, Elzbieta Polak, and Matthew Trager. Geometry and optimization of shallow polynomial networks. *SIAM Journal on Applied Algebra and Geometry*, 10(2): 174–209, 2026. doi: 10.1137/25M1732994.

Luca Arnaboldi, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. Escaping mediocrity: How two-layer networks learn hard generalized linear models with SGD. *arXiv preprint arXiv:2305.18502*, 2023a. URL https://arxiv.org/abs/2305.18502.

Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional and mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layer networks. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 1199–1227, 2023b. URL https://proceedings.mlr.press/v195/arnaboldi23a.html.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://openreview.net/forum?id=B1eI3EHxUr.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=1NvflqAdoom.

Gérard Ben Arous, Reza Gheissari, and Aukosh Jagannath. High-dimensional limit theorems for SGD: Effective dynamics and critical scaling. *Communications on Pure and Applied Mathematics*, 77(3):2030–2080, 2024. doi: 10.1002/cpa.22169.

Gérard Ben Arous, Murat A. Erdogdu, Nuri Mert Vural, and Denny Wu. Learning quadratic neural networks in high dimensions: SGD dynamics and scaling laws. In *Advances in Neural Information Processing Systems*, volume 38, pp. 146725–146812, 2025. doi: 10.52202/085713-4910. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d7ce06e9293c3d8e6cb3f80b4157f875-Abstract-Conference.html.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013. doi: 10.1109/TPAMI.2013.50.

Raphaël Berthier. Incremental learning in diagonal linear networks. *Journal of Machine Learning Research*, 24(171):1–26, 2023. URL https://jmlr.org/papers/v24/22-1395.html.

Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *Foundations of Computational Mathematics*, 25(5):1627–1710, 2025. doi: 10.1007/s10208-024-09664-9.

Etienne Boursier and Nicolas Flammarion. Early alignment in two-layer networks training is a two-edged sword. *Journal of Machine Learning Research*, 26(183):1–75, 2025. URL https://jmlr.org/papers/v26/24-1523.html.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20105–20118, 2022. doi: 10.52202/068431-1462.

Guillaume Braun, Bruno Loureiro, Ha Quang Minh, and Masaaki Imaizumi. Fast escape, slow convergence: Learning dynamics of phase retrieval under power-law data. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Ae4eZpkXBX.

Lukas Braun, Clémentine C. J. Dominé, James E. Fitzgerald, and Andrew M. Saxe. Exact learning dynamics of deep linear networks with prior knowledge. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-0479. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/2b3bb2c95195130977a51b3bb251c40a-Abstract-Conference.html.

Roger W. Brockett. Dynamical systems that sort lists, diagonalize matrices, and solve linear programming problems. *Linear Algebra and its Applications*, 146:79–91, 1991. doi: 10.1016/0024-3795(91)90021-N.

Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. Hitting the high-dimensional notes: An ODE for SGD learning dynamics on GLMs and multi-index models. *Information and Inference: A Journal of the IMA*, 13(4):iaae028, 2024. doi: 10.1093/imaiai/iaae028.

Leonardo Defilippis, Yizhou Xu, Julius Girardin, Vittorio Erba, Emanuele Troiani, Lenka Zdeborová, Bruno Loureiro, and Florent Krzakala. Scaling laws and spectra of shallow neural networks in the feature learning regime. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Q3yLIIkt7z.

Clémentine Dominé, Nicolas Anguita, Alexandra Maria Proca, Lukas Braun, Daniel Kunin, Pedro A. M. Mediano, and Andrew M. Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ZXaocmXc6d.

Vittorio Erba, Emanuele Troiani, Lenka Zdeborová, and Florent Krzakala. The nuclear route: Sharp asymptotics of ERM in overparameterized quadratic networks. In *Advances in Neural Information Processing Systems*, volume 38, pp. 88862–88901, 2025. doi: 10.52202/085713-2971.

Kenji Fukumizu. Effect of batch learning in multilayer neural networks. In *Proceedings of the 5th International Conference on Neural Information Processing (ICONIP)*, pp. 67–70, 1998.

Sebastian Goldt, Madhu S. Advani, Andrew M. Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In *Advances in Neural Information Processing Systems*, volume 32, pp. 6979–6989, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/cab070d53bd0d200746fb852a922064a-Abstract.html.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of *Proceedings of Machine Learning Research*, pp. 426–471, 2022. URL https://proceedings.mlr.press/v145/goldt22a.html.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. *arXiv preprint arXiv:1705.09280*, 2017. URL https://arxiv.org/abs/1705.09280.

Katherine L. Hermann and Andrew K. Lampinen. What shapes feature representations? exploring datasets, architectures, and training. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9995–10006, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/71e9c6620d381d60196ebe694840aaaa-Abstract.html.

Katherine L. Hermann, Hossein Mobahi, Thomas Fel, and Michael C. Mozer. On the foundations of shortcut learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Tj3xLVuE9f.

Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998. doi: 10.1017/CBO9781139173179.

Alessandro Ingrosso and Sebastian Goldt. Data-driven emergence of convolutional structure in neural networks. *Proceedings of the National Academy of Sciences*, 119(40):e2201854119, 2022. doi: 10.1073/pnas.2201854119.

L. Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1–2):134–139, 1918. doi: 10.1093/biomet/12.1-2.134.

Jikai Jin, Zhiyuan Li, Kaifeng Lyu, Simon S. Du, and Jason D. Lee. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 15200–15238, 2023. URL https://proceedings.mlr.press/v202/jin23a.html.

Akshay Kumar and Jarvis Haupt. Directional convergence near small initializations and saddles in two-homogeneous neural networks. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2402.09226.

Daniel Kunin, Allan Raventós, Clémentine Dominé, Feng Chen, David Klindt, Andrew M. Saxe, and Surya Ganguli. Get rich quick: Exact solutions reveal how unbalanced initializations promote rapid feature learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://openreview.net/forum?id=eNM94i7R3A.

Andrew K. Lampinen and Surya Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ryfMLoCqtQ.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 2–47, 2018. URL https://proceedings.mlr.press/v75/li18a.html.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *arXiv preprint arXiv:2012.09839*, 2020. URL https://arxiv.org/abs/2012.09839.

Leon Lufkin, Andrew Saxe, and Erin Grant. Nonlinear dynamics of localization in neural receptive fields. In *Advances in Neural Information Processing Systems*, volume 37, pp. 25938–25960, 2024. doi: 10.52202/079017-0815. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2dab2f94544f9297d01a46a5453b93cd-Paper-Conference.pdf.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes ReLU network features. *arXiv preprint arXiv:1803.08367*, 2018. URL https://arxiv.org/abs/1803.08367.

Simon Martin, Francis Bach, and Giulio Biroli. On the impact of overparameterization on the training of a shallow neural network in high dimensions. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3655–3663, 2024. URL https://proceedings.mlr.press/v238/martin24a.html.

Simon Martin, Giulio Biroli, and Francis Bach. High-dimensional analysis of gradient flow for extensive-width quadratic neural networks. *Journal of Machine Learning Research*, 27(136):1–182, 2026. URL https://jmlr.org/papers/v27/26-0672.html.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15(3):267–273, 1982. doi: 10.1007/BF00275687.

Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 7475–7505, 2023. doi: 10.52202/075280-0329.

David Saad and Sara A. Solla. On-line learning in soft committee machines. *Physical Review E*, 52 (4):4225–4243, 1995. doi: 10.1103/PhysRevE.52.4225.

Luca Saglietti, Stefano Mannelli, and Andrew M. Saxe. An analytical theory of curriculum learning in teacher-student networks. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1535. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/84bad835faaf48f24d990072bb5b80ee-Abstract-Conference.html.

Stefano Sarao Mannelli, Eric Vanden-Eijnden, and Lenka Zdeborová. Optimization and generalization of shallow neural networks with quadratic activation functions. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13445–13455, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/9b8b50fb590c590ffbf1295ce92258dc-Paper.pdf.

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

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23): 11537–11546, 2019. doi: 10.1073/pnas.1820226116.

H. S. Seung, H. Sompolinsky, and N. Tishby. Statistical mechanics of learning from examples. *Physical Review A*, 45(8):6056–6091, 1992. doi: 10.1103/PhysRevA.45.6056.

Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23831–23843, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c82836ed448c41094025b4a872c5341e-Paper.pdf.

Loek Van Rossem and Andrew M. Saxe. When representations align: Universality in representation learning dynamics. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 49098–49121, 2024. URL https://proceedings.mlr.press/v235/van-rossem24a.html.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673, 2020. URL https://proceedings.mlr.press/v125/woodworth20a.html.

Wei-Yong Yan, Uwe Helmke, and John B. Moore. Global analysis of Oja's flow for neural networks. *IEEE Transactions on Neural Networks*, 5(5):674–683, 1994. doi: 10.1109/72.317720.

Yedi Zhang, Andrew Saxe, and Peter E. Latham. Saddle-to-saddle dynamics explains a simplicity bias across neural network architectures. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Vit5M0G5Gb.

Liu Ziyin, Yizhou Xu, Tomaso Poggio, and Isaac Chuang. Neural quadratic forms: A unified minimal model for sudden learning and scaling laws. *arXiv preprint arXiv:2608.13335*, 2026. URL https://arxiv.org/abs/2608.13335.

A POPULATION FLOW AND THE DIRECTION–AMPLITUDE SPLIT

We collect the formal statements and proofs here. The Frobenius inner product is $\langle A, B\rangle = \mathrm{tr}(A^\top B)$.

Theorem A.1 (Moment sufficiency and population flow)**.** *Let $x \in \mathbb{R}^d$ have mean zero, covariance $\Sigma \succ 0$, and finite fourth moment. Let*

$$f_M(x) = x^\top M x - \mathrm{tr}(M\Sigma), \qquad M = WW^\top.$$

For any symmetric A, set $(K_A)_{ij} = \sum_{a,b} \kappa_{ijab} A_{ab}$ and $H_A = \Sigma A \Sigma + \frac{1}{2} K_A$, where κ is the fourth cumulant,

$$\kappa_{ijab} = \mathbb{E}[x_i x_j x_a x_b] - \Sigma_{ij}\Sigma_{ab} - \Sigma_{ia}\Sigma_{jb} - \Sigma_{ib}\Sigma_{ja}. \tag{19}$$

Then the population loss depends on the input law only through (Σ, κ) and equals

$$\mathcal{L}(M) = \mathrm{tr}((M - M_\star)\Sigma(M - M_\star)\Sigma) + \frac{1}{2}\langle M - M_\star, K_{M-M_\star}\rangle.$$

After rescaling time by four, gradient flow is

$$\dot{W} = H_{M_\star - M} W, \qquad \dot{M} = H_{M_\star - M} M + M H_{M_\star - M}.$$

For any symmetric $d \times d$ matrices A and B,

$$\langle A, H_B\rangle = \langle H_A, B\rangle, \qquad \langle A, H_A\rangle = \frac{1}{2}\mathbb{E}\left[\left(x^\top A x - \mathrm{tr}(A\Sigma)\right)^2\right] \geq 0.$$

Proof. Set $A = M - M_\star$ and $e_A(x) = x^\top A x - \mathrm{tr}(A\Sigma)$. The residual derivative is

$$\nabla_W e_A(x) = 2(xx^\top - \Sigma)W.$$

Expanding the fourth moment into its three covariance pairings and the cumulant gives

$$\mathbb{E}[e_A(x)(xx^\top - \Sigma)] = 2\Sigma A \Sigma + K_A = 2H_A.$$

Therefore $\nabla_W \mathcal{L} = 4 H_A W$. The stated W equation follows after rescaling time, and differentiation of $M = WW^\top$ gives the matrix equation.

The same fourth-moment expansion yields

$$\tfrac{1}{2}\mathbb{E}[e_A(x)^2] = \mathrm{tr}(A\Sigma A\Sigma) + \tfrac{1}{2}\langle A, K_A\rangle.$$

The same expectation with two residual matrices gives the cross identity. Setting the two matrices equal gives

$$2\langle A, H_A\rangle = \mathbb{E}[e_A(x)^2] \geq 0.$$

This proves all claims. □

Lemma A.2 (Direction and amplitude split)**.** *Let $M(t) = U(t)S(t)U(t)^\top$ have rank k, with $U^\top U = I_k$ and $S \succ 0$. Choose the basis U so that $U^\top \dot{U} = 0$. If*

$$\dot{M} = HM + MH$$

for a symmetric matrix $H(t)$, then

$$\dot{U} = (I - P)HU, \qquad \dot{S} = (U^\top H U)S + S(U^\top H U), \qquad \dot{P} = (I - P)HP + PH(I - P),$$

where $P = UU^\top$. If the split holds at $t = 0$, it holds for all $t \geq 0$.

Proof. Differentiate $M = USU^\top$ and multiply by U on the right. The basis convention gives

$$\dot{M}U = \dot{U}S + U\dot{S}.$$

Projection with $U^\top$ gives the amplitude equation. Projection with $I - P$ gives

$$\dot{U} = (I - P)\dot{M}US^{-1} = (I - P)HU.$$

Differentiating $P = UU^\top$ then yields

$$\dot{P} = (I - P)HP + PH(I - P) = HP + PH - 2PHP.$$

Conversely, solve the direction and amplitude equations from a split of $M(0)$. Since $U^\top(I - P) = 0$, the direction equation gives $U^\top \dot{U} = 0$ and $\frac{d}{dt}(U^\top U) = 0$, so U stays orthonormal. The amplitude equation gives $S(t) = \Phi(t)S(0)\Phi(t)^\top$ with $\dot{\Phi} = (U^\top H U)\Phi$ and $\Phi(0) = I_k$, so S stays positive definite. Substituting shows that $USU^\top$ solves $\dot{M} = HM + MH$ with the same initial value, so it equals $M(t)$ for all $t \geq 0$. □

Taking $H = H_{M_\star - M} = H_\star - H_M$ and using linearity proves equation (5). We use P as the direction variable because it does not change when the basis rotates inside the student subspace.

B GLOBAL ISOTROPIC SOLUTION AND SEPARATED TIMESCALES

Theorem B.1 (Global isotropic matrix and projector solution). *Let $M_\star \succeq 0$ and $M_0 = U_0 S_0 U_0^\mathsf{T}$ with $U_0^\mathsf{T} U_0 = I_k$ and $S_0 \succ 0$. The unique solution of*

$$\dot{M} = M_\star M + M M_\star - 2M^2, \qquad M(0) = M_0,$$

is

$$M(t) = Y(t) N(t)^{-1} Y(t)^\mathsf{T}, \qquad Y(t) = e^{tM_\star} U_0, \qquad N(t) = S_0^{-1} + 2\int_0^t Y(s)^\mathsf{T} Y(s)\, ds. \tag{20}$$

Its range projector is equation (7). The solution exists for every $t \geq 0$.

Proof. The matrix $N(t)$ is positive definite because $S_0^{-1} \succ 0$ and its integral term is positive semidefinite. Differentiate $M = Y N^{-1} Y^\mathsf{T}$ using $\dot{Y} = M_\star Y$ and $\dot{N} = 2Y^\mathsf{T} Y$:

$$\begin{aligned}\dot{M} &= M_\star Y N^{-1} Y^\mathsf{T} + Y N^{-1} Y^\mathsf{T} M_\star - Y N^{-1} (2Y^\mathsf{T} Y) N^{-1} Y^\mathsf{T} \\ &= M_\star M + M M_\star - 2M^2.\end{aligned}$$

At $t = 0$, $N^{-1} = S_0$, so the initial condition holds. Because $N(t)$ stays positive definite, the solution has no finite-time singularity. Since $N^{-1} \succ 0$, M and Y have the same range, and the projector formula for a matrix of full column rank gives equation (7). □

Corollary B.2 (Initialization-scale separation). *Let $S_0 = \varepsilon \bar{S}_0$ for fixed $\bar{S}_0 \succ 0$. Then $P(t)$ is independent of ε. If the amplitude of an aligned mode j with $\lambda_j > 0$ starts at $s_j(0) = \Theta(\varepsilon)$, the time it takes to reach a fixed fraction of λ_j satisfies equation (9).*

Proof. The projector formula contains U_0 but not S_0. Solving the logistic equation in equation (9) for any threshold strictly between zero and λ_j gives the stated expansion. □

The logistic equation describes only modes that are already aligned. The matrix solution in theorem B.1 holds for any S_0, U_0 and $M_\star$, including when they do not share a basis.

Proposition B.3 (Best loss within the current subspace). *For Gaussian inputs with covariance I, let $M_\star \succeq 0$ and $M = USU^\mathsf{T}$, with $U^\mathsf{T} U = I_k$ and $S \succeq 0$. Then*

$$\mathcal{L}(M) = \|M_\star\|_F^2 - \|U^\mathsf{T} M_\star U\|_F^2 + \|S - U^\mathsf{T} M_\star U\|_F^2, \tag{21}$$

$$\min_{\substack{A \succeq 0 \\ A = PAP}} \mathcal{L}(A) = \|M_\star\|_F^2 - \|U^\mathsf{T} M_\star U\|_F^2, \qquad P = UU^\mathsf{T}. \tag{22}$$

The minimizing matrix is $PM_\star P$. Along the isotropic direction flow, the minimum in (22) cannot increase.

Proof. Write $\Omega = U^\mathsf{T} M_\star U$. Expanding the squared residual gives

$$\|M_\star - USU^\mathsf{T}\|_F^2 = \|M_\star\|_F^2 - 2\,\mathrm{tr}(S\Omega) + \|S\|_F^2.$$

Completing the square proves (21). Since $\Omega \succeq 0$, it is a feasible minimizer over matrices supported in P. All three terms are invariant under $U \mapsto UQ$, $S \mapsto Q^\mathsf{T} S Q$ for orthogonal Q.

Choose $U^\mathsf{T} \dot{U} = 0$. The direction equation gives

$$\dot{\Omega} = 2U^\mathsf{T} M_\star (I - P) M_\star U \succeq 0.$$

Consequently,

$$\frac{d}{dt}\left(\|M_\star\|_F^2 - \|\Omega\|_F^2\right) = -4\,\mathrm{tr}\left(\Omega\, U^\mathsf{T} M_\star (I - P) M_\star U\right) \leq 0.$$

Both matrices in the trace are positive semidefinite, proving the sign. □

The amplitude error in (21) can grow while the subspace improves and the output is still small.

Figure A1 uses the isotropic trajectory of Figure 1 with $\varepsilon = 10^{-7}$. We divide both losses by the loss of the zero predictor, $\|M_\star\|_F^2 = 11.7725$. The two sides of equation (21) agree to 2.3×10^{-14}. Along this trajectory, the best loss within the subspace decreases throughout, while the amplitude error first increases.

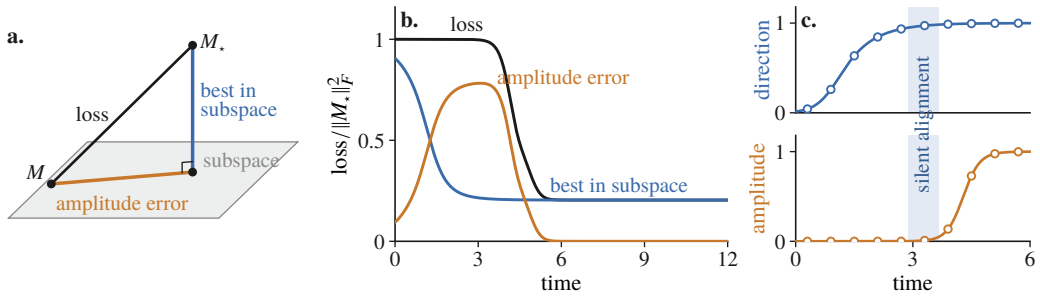


Figure A1: **The student subspace aligns before the loss decreases.** (a) Projecting the teacher onto the current subspace splits the loss, the squared distance from M to $M_\star$, into the best loss within the subspace and the amplitude error (equation (21)). (b) Along the isotropic trajectory the loss stays flat while the subspace aligns, because the best loss within the subspace falls and the amplitude error rises until the amplitudes grow. (c) On the same trajectory, the direction is acquired before the largest amplitude grows; shading marks silent alignment.

C Student subspace overlaps and the direction hierarchy

Theorem C.1 (Student subspace overlap replicator)**.** *Let $M_\star = V\Lambda V^\mathsf{T}$ and $U_0^\mathsf{T}U_0 = I_k$. For every k-subset I, define $p_I(t) = \det(V_I^\mathsf{T}U(t))^2$ and $E_I = \sum_{i\in I}\lambda_i$ as in the main text. Along the isotropic projector flow,*

$$p_I(t) = \frac{p_I(0)e^{2E_I t}}{\sum_J p_J(0)e^{2E_J t}}.$$

Consequently, $\dot{p}_I = 2(E_I - \bar{E})p_I$ and pairwise log odds obey equation (8).

Proof. In the teacher basis,

$$V^\mathsf{T}Y(t) = e^{t\Lambda}V^\mathsf{T}U_0.$$

Define the signed projected volume $a_I(t) = \det(V_I^\mathsf{T}Y(t))$. Then $a_I(t) = e^{E_I t}a_I(0)$ and $a_I(0)^2 = p_I(0)$. By the Cauchy–Binet identity, $\sum_J a_J(t)^2 = \det(Y(t)^\mathsf{T}Y(t))$. Since Y and U span the same subspace, write $Y = UB$ with B invertible. Thus $a_I(t)^2 = p_I(t)\det(B)^2$ and $\det(Y^\mathsf{T}Y) = \det(B)^2$, giving $p_I(t) = a_I(t)^2/\sum_J a_J(t)^2$. In particular, the overlaps sum to one. Substitution gives the stated overlap solution, and differentiation gives the replicator equation. □

Proposition C.2 (Monotone cumulative capture)**.** *Let Π_m project onto the first m teacher eigenvectors and let $C_m = \mathrm{tr}(\Pi_m P)$. Then*

$$\dot{C}_m = 2\sum_{i\le m}\sum_{\ell>m}(\lambda_i - \lambda_\ell)|P_{i\ell}|^2 \ge 0.$$

Proof. Work in the teacher basis. Substitute $\dot{P} = \Lambda P + P\Lambda - 2P\Lambda P$ into $\dot{C}_m$. Use $P^2 = P$ to expand the traces. Terms whose two indices are on the same side of the cutoff cancel, leaving the displayed cross terms. □

C.1 Alignment rates and threshold times

Let $1 \le k < d$, $\lambda_k > \lambda_{k+1}$, and $I_\star = \{1,\dots,k\}$. Assume $p_{I_\star}(0) > 0$. For an orthonormal frame, the Cauchy–Binet identity gives the capture $P_{ii} = \sum_{I\ni i} p_I$, hence, for $m \le k$,

$$m - C_m(t) = \sum_I \big(m - |I\cap\{1,\dots,m\}|\big)p_I(t). \tag{23}$$

Write $d_m(I) = m - |I\cap\{1,\dots,m\}|$, $\Delta_I = E_{I_\star} - E_I$, and $r_I = p_I(0)/p_{I_\star}(0)$. The exact replicator solution yields

$$m - C_m(t) = \frac{\sum_{I\ne I_\star} d_m(I) r_I e^{-2\Delta_I t}}{1 + \sum_{I\ne I_\star} r_I e^{-2\Delta_I t}}. \tag{24}$$

Among the competing subspaces with $d_m(I) > 0$, the one that replaces direction m by direction $k+1$ has the smallest score difference, $g_m = \lambda_m - \lambda_{k+1}$. Consequently, for

$$B_m = \sum_{\substack{I \neq I_\star \\ \Delta_I = g_m}} d_m(I) r_I > 0,$$

the remaining capture error is $B_m e^{-2g_m t}(1 + o(1))$ as $t \to \infty$. The condition $B_m > 0$ holds for generic initialization, and the sum includes all competing subspaces that tie at this rate. Monotonicity of C_m then gives

$$t_m^{\text{align}}(\eta) = \frac{\log(B_m/\eta)}{2g_m} + o(1), \qquad \eta \to 0. \tag{25}$$

These are asymptotics for small error, whereas the independence from the initialization scale ε holds for the entire projector trajectory. When the relevant eigenvalues are distinct, B_m is the ratio r_I of the single competing subspace.

The error at $m = k$ is the sum of squared sines of the principal angles to the leading subspace. Thus $\sin^2 \theta_{\max} \leq k - C_k \leq k \sin^2 \theta_{\max}$. The squared principal-angle error plotted in Figures 1 and A1c has the same asymptotic rate $2(\lambda_k - \lambda_{k+1})$, but its threshold time need not equal the cumulative-capture threshold time.

If $k = \operatorname{rank}(M_\star) < d$, the same proof applies with $\lambda_{k+1} = 0$; if $\operatorname{rank}(M_\star) < k < d$, a competing subspace missing a positive direction m loses at least λ_m in score, giving rate $2\lambda_m$ for generic initial overlaps.

C.2 Mode-specific expression and silent alignment

Under the assumptions of Appendix C.1, the alignment time for the first $m \leq k$ directions is given by (25) and is independent of ε. We now establish the amplitude-growth time for mode m under generic initialization, so that the two times can be compared as in (10). Let $S_0 = \varepsilon \bar{S}_0$, with fixed $\bar{S}_0 \succ 0$.

We measure the expression of mode m by the coefficient $a_m(t) = v_m^\mathsf{T} M(t) v_m$ rather than by an eigenvalue of S, whose eigenvector can change. Set $b_i = \bar{S}_0^{1/2} U_0^\mathsf{T} v_i$ and $F_i(t) = 2\int_0^t e^{2\lambda_i s}\, ds$. The exact matrix solution gives

$$a_m(t) = \varepsilon e^{2\lambda_m t} b_m^\mathsf{T} \left(I + \varepsilon \sum_i F_i(t) b_i b_i^\mathsf{T} \right)^{-1} b_m \leq \varepsilon \|b_m\|^2 e^{2\lambda_m t}.$$

The bound holds at all times, including after stronger modes have grown. In particular, for any fixed $0 < \rho < 1$, at $t = \rho \log(1/\varepsilon)/(2\lambda_m)$,

$$m - C_m(t) = O\left(\varepsilon^{\rho g_m/\lambda_m}\right), \qquad a_m(t) = O\left(\varepsilon^{1-\rho}\right).$$

Both the remaining capture error and a_m vanish as initialization shrinks, so the student acquires the leading m directions before it expresses mode m, even if stronger modes are already expressed.

The exact solution also determines when expression becomes appreciable. Put $T_m = \log(1/\varepsilon)/(2\lambda_m)$ and evaluate at $t = T_m + s$. In the matrix inverse above, the coefficients of $b_i b_i^\mathsf{T}$ diverge when $\lambda_i > \lambda_m$, tend to $e^{2\lambda_m s}/\lambda_m$ when $\lambda_i = \lambda_m$, and vanish when $\lambda_i < \lambda_m$. Let L_m project onto the orthogonal complement of the span of the b_i with $\lambda_i > \lambda_m$, and write $J_m = \{i : \lambda_i = \lambda_m\}$ and $Q_m = (b_i^\mathsf{T} L_m b_j)_{i,j \in J_m}$. The initial overlap with the leading k-dimensional teacher subspace is full rank, so $Q_m \succ 0$, and block inversion gives, uniformly for bounded s,

$$V_{J_m}^\mathsf{T} M(T_m + s) V_{J_m} \longrightarrow e^{2\lambda_m s} Q_m \left(I + \frac{e^{2\lambda_m s}}{\lambda_m} Q_m \right)^{-1},$$

whose diagonal entries increase from zero to λ_m. The all-time bound excludes crossings far before T_m, so the first time $t_m^{\text{expr}}(\alpha)$ at which a_m reaches $\alpha\lambda_m$, for fixed $0 < \alpha < 1$, satisfies

$$t_m^{\text{expr}}(\alpha) = \frac{\log(1/\varepsilon)}{2\lambda_m} + O(1), \qquad \varepsilon \to 0.$$

For fixed η and α, subtracting $t_m^{\text{align}}(\eta)$ gives the duration of the mode-specific silent interval when the difference is positive; substituting (25) gives the comparison in (10), which describes well-separated leading timescales rather than a sharp threshold.

C.3 First amplitude growth and the global silent interval

Let $S_0 = \varepsilon \bar{S}_0$, with fixed $\bar{S}_0 \succ 0$ and a generic initial subspace. For fixed $0 < \alpha < 1$, define the first output threshold by $t_{\text{out}} = \inf\{t \geq 0 : \|M(t)\|_2 \geq \alpha\lambda_1\}$, where $\lambda_1 > 0$. Since $M \succeq 0$, its spectral norm is its largest amplitude. The full matrix solution gives

$$t_{\text{out}} = \frac{\log(1/\varepsilon)}{2\lambda_1} + \frac{1}{2\lambda_1} \log \frac{\alpha\lambda_1}{(1-\alpha)q} + o(1), \qquad \varepsilon \to 0, \tag{26}$$

where $q > 0$ is the largest eigenvalue of $Q = V_+^\top U_0 \bar{S}_0 U_0^\top V_+$, and V_+ spans the teacher eigenspace with eigenvalue λ_1. For a simple top eigenvalue, $q = v_1^\top U_0 \bar{S}_0 U_0^\top v_1$.

To derive (26), set $T_\varepsilon = \log(1/\varepsilon)/(2\lambda_1)$ and evaluate equation (20) at $t = T_\varepsilon + s$. For bounded s, the rows of $\sqrt{\varepsilon}Y$ associated with eigenvalues below λ_1 vanish as $\varepsilon \to 0$. The remaining rows and the corresponding terms in εN give

$$M(T_\varepsilon + s) \longrightarrow V_+ \left[e^{2\lambda_1 s} Q \left(I + \frac{e^{2\lambda_1 s}}{\lambda_1} Q \right)^{-1} \right] V_+^\top, \tag{27}$$

uniformly on bounded intervals of s. Its largest eigenvalue is $qe^{2\lambda_1 s}/(1 + qe^{2\lambda_1 s}/\lambda_1)$. Solving for its crossing of $\alpha\lambda_1$ gives the offset in (26). This crossing is the first one, because earlier crossings are ruled out by the global bound

$$\operatorname{tr} M(t) \leq \varepsilon \operatorname{tr}(\bar{S}_0) e^{2\lambda_1 t}, \tag{28}$$

which follows from $N^{-1} \preceq S_0$ and $Y^\top Y \preceq e^{2\lambda_1 t} I$. The derivation holds for unaligned initialization and does not use the logistic equation.

Subtracting the ε-independent alignment time gives, for fixed η and α, $t_{\text{out}} - t_k^{\text{align}}(\eta) = \log(1/\varepsilon)/(2\lambda_1) - t_k^{\text{align}}(\eta) + O(1)$. A positive difference defines an interval in which the whole retained subspace is aligned to that accuracy and no amplitude exceeds the output threshold. Before t_{out}, $\|M\|_F \leq \sqrt{k}\alpha\lambda_1$ and $\mathbb{E}[f_W(x)^2] = 2\|M\|_F^2$, so the output and its effect on the loss stay small. The longer delay until the weakest retained amplitude reaches its threshold, plotted in Figure 1c, is not a loss-silent interval.

D The early projector for general inputs

Theorem D.1 (Early projector on a fixed time interval)**.** *Let H_A be the growth matrix of theorem A.1, set $H_\star = H_{M_\star}$, and let W_ε follow the population flow $\dot{W}_\varepsilon = (H_\star - H_{W_\varepsilon W_\varepsilon^\top}) W_\varepsilon$ from $W_\varepsilon(0) = \sqrt{\varepsilon} U_0 \bar{S}_0^{1/2}$, with $U_0^\top U_0 = I_k$ and $\bar{S}_0 \succ 0$. For every $T < \infty$, there is a constant C_T such that, for sufficiently small ε,*

$$\begin{aligned} \sup_{0 \leq t \leq T} \|W_\varepsilon(t) - e^{tH_\star} W_\varepsilon(0)\|_F &\leq C_T \varepsilon^{3/2}, \\ \sup_{0 \leq t \leq T} \|P_\varepsilon(t) - P_{\text{lin}}(t)\|_2 &\leq C_T \varepsilon, \end{aligned} \tag{29}$$

where, for $Y_{\text{lin}}(t) = e^{tH_\star} U_0$,

$$P_{\text{lin}}(t) = Y_{\text{lin}}(t) \big(Y_{\text{lin}}(t)^\top Y_{\text{lin}}(t) \big)^{-1} Y_{\text{lin}}(t)^\top.$$

Proof. Let $W_{\text{lin}}(t) = e^{tH_\star} W_\varepsilon(0)$ and $\delta = W_\varepsilon - W_{\text{lin}}$. Then

$$\dot{\delta} = H_\star \delta - H_{W_\varepsilon W_\varepsilon^\top} W_\varepsilon, \qquad \delta(0) = 0.$$

The growth matrix is linear in A and satisfies $\|H_A\|_2 \leq L_H \|A\|_F$ with $L_H = \frac{1}{2}\mathbb{E}\|xx^\top - \Sigma\|_F^2$. On a fixed interval, a standard bootstrap gives $\|W_\varepsilon\|_F = O(\sqrt{\varepsilon})$, so the forcing term is $O(\varepsilon^{3/2})$. Variation of constants gives the first bound.

The smallest singular value of $e^{tH_\star} U_0 \bar{S}_0^{1/2}$ is uniformly positive on $[0, T]$. The smallest singular values of the scaled frames are therefore bounded below by $b_T\sqrt{\varepsilon}$. Since the projector map is locally Lipschitz with constant $O(\varepsilon^{-1/2})$, the frame error of order $\varepsilon^{3/2}$ gives a projector error of order ε. □

D.1 ALIGNMENT ON A GROWING TIME INTERVAL

Theorem D.1 holds on a fixed time interval, so it does not cover times proportional to $\log(1/\varepsilon)$. The next theorem does. It uses $\operatorname{tr}(AH_A) \geq 0$ from theorem A.1, a spectral gap of $H_\star$, and an initial overlap.

Theorem D.2 (Alignment before the amplitudes grow)**.** *Let H_A be the growth matrix of theorem A.1 and $H_\star = H_{M_\star}$ with $M_\star \succeq 0$. Order the eigenvalues of $H_\star$ as $\gamma_1 \geq \cdots \geq \gamma_d$. Assume $\gamma_1 > 0$, $1 \leq k < d$, and $\delta = \gamma_k - \gamma_{k+1} > 0$. Let V_1 contain its first k orthonormal eigenvectors, let V_2 contain the rest, and set $P_H = V_1 V_1^\mathsf{T}$. Initialize $W_\varepsilon(0) = \sqrt{\varepsilon}\, U_0 \bar{S}_0^{1/2}$, where $U_0^\mathsf{T} U_0 = I_k$, $\bar{S}_0 \succ 0$, and $V_1^\mathsf{T} U_0$ is invertible. Write P_ε for the range projector of W_ε and P_{lin} for that of $e^{tH_\star} U_0$.*

For any fixed $0 < b < (2\gamma_1)^{-1}$ and sufficiently small ε, uniformly on $0 \leq t \leq b \log(1/\varepsilon)$,

$$\operatorname{tr} M_\varepsilon(t) \leq a_0 \varepsilon e^{2\gamma_1 t}, \tag{30}$$

$$\|P_\varepsilon(t) - P_{\mathrm{lin}}(t)\|_2 \leq D\varepsilon e^{2\gamma_1 t}, \tag{31}$$

$$\|P_\varepsilon(t) - P_H\|_2 \leq \|X_0\|_2 e^{-\delta t} + D\varepsilon e^{2\gamma_1 t}. \tag{32}$$

The constants can be chosen as

$$a_0 = \operatorname{tr} \bar{S}_0, \qquad L_H = \tfrac{1}{2}\mathbb{E}\|xx^\mathsf{T} - \Sigma\|_F^2,$$

$$X_0 = (V_2^\mathsf{T} U_0)(V_1^\mathsf{T} U_0)^{-1}, \qquad R = \|X_0\|_2 + 1, \qquad D = \frac{L_H a_0 (1+R)^2}{2\gamma_1 + \delta}.$$

In particular, at

$$t_\varepsilon = \frac{\log(1/\varepsilon)}{2\gamma_1 + \delta}, \qquad \alpha = \frac{\delta}{2\gamma_1 + \delta} > 0,$$

both $\|P_\varepsilon(t_\varepsilon) - P_H\|_2$ and $\operatorname{tr} M_\varepsilon(t_\varepsilon)$ are $O(\varepsilon^\alpha)$.

Proof. For every symmetric A, $\|H_A\|_2 \leq L_H \|A\|_F$ and $\operatorname{tr}(AH_A) = \frac{1}{2}\mathbb{E}[(x^\mathsf{T} A x - \operatorname{tr}(A\Sigma))^2] \geq 0$. Since $M_\varepsilon \succeq 0$, the matrix flow implies

$$\frac{d}{dt} \operatorname{tr} M_\varepsilon = 2 \operatorname{tr}(H_\star M_\varepsilon) - 2 \operatorname{tr}(M_\varepsilon H_{M_\varepsilon}) \leq 2\gamma_1 \operatorname{tr} M_\varepsilon.$$

This proves (30), prevents finite-time blowup of W_ε, and bounds its feedback by $\|H_{M_\varepsilon}\|_2 \leq L_H a_0 \varepsilon e^{2\gamma_1 t}$. The weight equation multiplies W_ε on the left by a continuous matrix, so W_ε keeps full column rank at every finite time.

Work in the basis $V = (V_1, V_2)$, with eigenvalue blocks Γ_1, Γ_2, and put $E = -V^\mathsf{T} H_{M_\varepsilon} V$. The current subspace is a graph over its first k axes whenever $A = V_1^\mathsf{T} W_\varepsilon$ is invertible. Its graph matrix $X = (V_2^\mathsf{T} W_\varepsilon) A^{-1}$ then obeys

$$\dot{X} = \Gamma_2 X - X\Gamma_1 + E_{21} + E_{22} X - X E_{11} - X E_{12} X. \tag{33}$$

The linear graph is $X_{\mathrm{lin}}(t) = e^{t\Gamma_2} X_0 e^{-t\Gamma_1}$, with $\|X_{\mathrm{lin}}(t)\|_2 \leq e^{-\delta t}\|X_0\|_2$. If $\|X\|_2 \leq R$, the last four terms of (33) have norm at most $(1+R)^2 L_H a_0 \varepsilon e^{2\gamma_1 t}$. Variation of constants therefore yields

$$\|X(t) - X_{\mathrm{lin}}(t)\|_2 \leq (1+R)^2 L_H a_0 \varepsilon \int_0^t e^{-\delta(t-s)} e^{2\gamma_1 s}\, ds \leq D\varepsilon e^{2\gamma_1 t}.$$

Choose ε so that $D\varepsilon^{1-2\gamma_1 b} \leq 1/2$. This estimate closes the bootstrap strictly inside $\|X\|_2 < R$. Also $\dot{A} = (\Gamma_1 + E_{11} + E_{12} X) A$ has bounded coefficients on the interval. Its fundamental matrix stays invertible, so A stays invertible and the graph description remains valid.

For two graphs X, Z, an orthonormal frame for the first is $Q_X = \binom{I}{X}(I + X^\mathsf{T} X)^{-1/2}$. If P_Z is the projector onto the second, then

$$(I - P_Z) Q_X = (I - P_Z) \begin{pmatrix} 0 \\ X - Z \end{pmatrix} (I + X^\mathsf{T} X)^{-1/2}.$$

The largest principal-angle sine is consequently at most $\|X - Z\|_2$. Equal-rank orthogonal projectors satisfy $\|P_X - P_Z\|_2 = \|(I - P_Z) Q_X\|_2$, which proves (31). Taking $Z = 0$ and adding the linear graph bound proves (32). The stated t_ε balances its two exponents. □

The student approaches the leading subspace of the effective teacher $H_\star$, which need not be that of $M_\star$, and the theorem needs neither a common basis for the teacher and the data moments nor $\gamma_k > 0$. Since $\|f_{M_\varepsilon}\|_{L^2} \le \sqrt{2L_H}\ \mathrm{tr}\, M_\varepsilon$, the student subspace approaches P_H while the output still vanishes. The exponent α comes from an upper bound and need not be sharp. The proof of theorem C.1 uses only $Y = e^{tM_\star}U_0$, so it applies to P_{lin} with $H_\star$ in place of $M_\star$. This gives the early overlap hierarchy of Section 4.

D.2 Numerical check

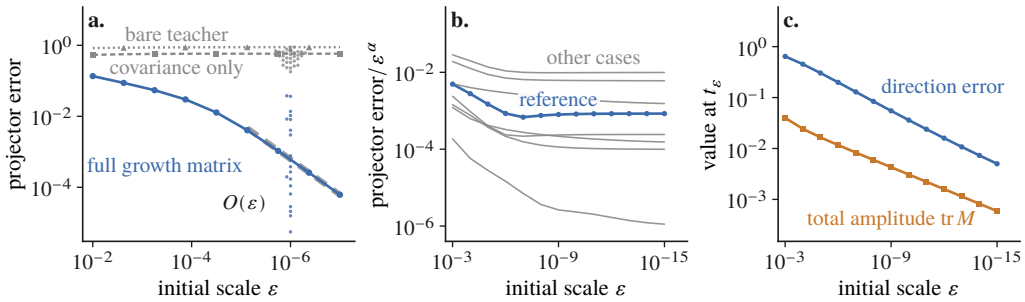


Figure A2: **The full growth matrix predicts early direction learning at fixed and growing times.** (a) Only the full growth matrix predicts the early projector with error proportional to ε when teacher and data moments have different eigenvectors; dots, twenty further rotations. (b) At t_ε, the error of the linear projector prediction over ε^α stays bounded; blue, reference problem; gray, seven other cases. (c) The direction error and the total amplitude both decrease. Observation time increases to the right in (b, c); these check an upper bound, not a sharp exponent.

The main rotated example has $d = 6$, $k = 2$, covariance eigenvalues

$$(0.60, 1.40, 0.85, 1.80, 1.05, 0.75),$$

coordinate excess kurtoses

$$(0, 3, -0.4, 1.5, 0, 0.8),$$

and teacher eigenvalues $(2.3, 1.8, 1.2, 0.7, 0.35, 0.1)$. The teacher basis and initial frame are fixed random rotations with seed 271828. The projector is evaluated at time 0.8 for nine scales between 10^{-2} and 10^{-7}. The log-log slope fitted to the four smallest scales is 0.971. At the smallest scale, the projector errors when predicting with the full growth matrix, with its covariance part only, and with the teacher alone are 6.16×10^{-5}, 0.581, and 0.887. Twenty further fixed rotations give median errors 4.83×10^{-4}, 0.421, and 0.710 at scale 10^{-6}.

The growing-time test in Figure A2b,c uses the same setup with teacher and initial-frame seeds 271828, 271829 and 271830, initial amplitude condition numbers 1 and 10^3 at fixed $\mathrm{tr}\, \bar{S}_0 = 2$, and, in the reference problem, two further initial subspaces with graph norms 10 and 100 relative to the leading eigenspace of $H_\star$. All eight cases are evaluated at thirteen scales $\varepsilon = 10^{-3}, \dots, 10^{-15}$, each at $t_\varepsilon = \log(1/\varepsilon)/(2\gamma_1 + \delta)$, integrating $W/\sqrt{\varepsilon}$ with DOP853 at relative tolerance 2×10^{-12}. The sufficient condition in the proof is conservative. Ten of the 104 trajectories satisfy it, and the others serve as numerical checks outside the theorem. In the reference problem ($\gamma_1 = 8.6698$, $\delta = 2.9089$, $\alpha = 0.14366$), the endpoint direction error falls from 0.6424 to 0.00500 and the total amplitude from 0.03993 to 0.000590 between the largest and smallest scales.

E When covariance or fourth moments preserve every subspace

Proposition E.1 (Covariance condition for direction preservation)**.** *Fix $1 \le k < d$. The covariance feedback*

$$(I - P)\Sigma(USU^\mathsf{T})\Sigma U$$

vanishes for every $U^\mathsf{T}U = I_k$ and every $S \succ 0$ if and only if $\Sigma = cI$ for some $c > 0$.

Proof. The forward direction is immediate. For the converse, linearity in S extends the condition from the positive cone to every symmetric S. Taking a rank-one S inside any k-frame shows that Σu lies in every k-dimensional subspace containing u. The intersection of those subspaces is $\mathrm{span}(u)$. Thus every unit vector is an eigenvector of Σ, which forces Σ to be scalar. □

Theorem E.2 (Fourth-moment condition for direction preservation)**.** *Fix $1 \le k < d$ and a fully symmetric fourth-cumulant tensor κ. The following statements are equivalent:*

(i) $(I - P)K_{USU^\top}U = 0$ *for every* $U^\top U = I_k$ *and every* $S \succ 0$*;*

(ii) $K_{uu^\top}u \in \mathrm{span}(u)$ *for every unit vector* u*;*

(iii) there is a scalar $\bar{\kappa}$ *such that* $K_A = \bar{\kappa}(\mathrm{tr}(A)I + 2A)$ *for every symmetric* A*.*

Proof. Statement (iii) makes both terms vanish after projection, so it implies (i). To obtain (ii) from (i), use a rank-one S selecting one column u and vary its orthonormal completion. Any component of $K_{uu^\top}u$ orthogonal to u can be excluded from one such completion, contradicting (i).

For (ii) to (iii), define the quartic form

$$q(u) = u^\top K_{uu^\top}u.$$

Full tensor symmetry gives $\nabla q(u) = 4K_{uu^\top}u$. Condition (ii) says that q has no tangential derivative on the unit sphere. It is therefore constant there. Homogeneity gives $q(u) = q_0\|u\|^4$. Four derivatives recover the isotropic fourth-order tensor, and the definition of K_A gives (iii) with $\bar{\kappa} = q_0/3$. □

For identity covariance and spherical fourth moments, a valid input distribution requires $\bar{\kappa} \ge -2/(d+2)$, which follows from $\mathrm{tr}(AH_A) \ge 0$ at $A = I$.

The two results treat the covariance and cumulant contributions separately, and the two contributions can cancel when combined. In $d = 2$, independent coordinates with $\Sigma = \mathrm{diag}(1, 2)$ and excess kurtoses $(2, -1)$ give $K_A = \mathrm{diag}(2A_{11}, -4A_{22})$ and $H_A = 2A$, so every subspace is preserved although neither condition holds.

F Exact solutions for the covariance and kurtosis examples

For the covariance example, we use the initial growth rates g_i and the values $\widetilde{\lambda}_i$ of the total r at which growth along axis i stops,

$$g_i = d_i^2\lambda_i, \qquad \widetilde{\lambda}_i = d_i\lambda_i. \tag{34}$$

Dividing the exact solutions in equation (36) and taking the logarithm gives

$$\log\frac{q_A(t)}{q_B(t)} = \log\frac{q_A(0)}{q_B(0)} + 2(g_A - g_B)t - 2(d_A - d_B)\rho(t). \tag{35}$$

Solving the scalar equation for ρ therefore gives both the weights q_i and their log ratios.

Proposition F.1 (Exact solution and feature replacement for the covariance example)**.** *Let the inputs be Gaussian with $\Sigma = \mathrm{diag}(d_i)$, let $M_\star = \mathrm{diag}(\lambda_i)$, and take $k = 1$. Define $q_i = d_i w_i^2$, $r = \sum_i q_i$, and $\widetilde{\lambda}_i, g_i$ as in equation (34). Then equation (12) holds, and with $\rho(t) = \int_0^t r(s)\,ds$,*

$$q_i(t) = q_i(0)e^{2g_i t - 2d_i\rho(t)}, \qquad \dot{\rho} = \sum_i q_i(0)e^{2g_i t - 2d_i\rho}, \qquad \rho(0) = 0. \tag{36}$$

If the values $\widetilde{\lambda}_i$ are distinct, the only stable surviving feature among the initially present axes is the axis with the largest $\widetilde{\lambda}_i$. For $q_i(0) = \varepsilon\bar{q}_i$ with fixed $\bar{q}_i > 0$, the axis B with the largest g_i grows first as $\varepsilon \to 0$. If the axis A with the largest $\widetilde{\lambda}_i$ differs from B, the feature that grows first is replaced.

Proof. Direct substitution into the weight flow gives

$$\dot{w}_i = d_i(\widetilde{\lambda}_i - r)w_i = (g_i - d_i r)w_i.$$

Differentiating $q_i = d_i w_i^2$ gives equation (12), and integration gives equation (36). A fixed point with $q_i > 0$ requires $r = \widetilde{\lambda}_i$. Distinct values of $\widetilde{\lambda}_i$ rule out coexistence. At a fixed point on any other axis, the axis with the largest $\widetilde{\lambda}_i$ has positive growth rate, so only that axis is stable. At comparable small initialization, q_i grows initially as $\varepsilon \bar{q}_i e^{2g_i t}$, so the axis with the largest g_i grows first. Subtracting the integrated equations gives equation (35). □

The three log-odds slopes of Section 5,

$$\underbrace{2(g_A - g_B)}_{\text{initial growth}}, \qquad \underbrace{2d_A(\widetilde{\lambda}_A - \widetilde{\lambda}_B)}_{\text{B plateau}}, \qquad \underbrace{2d_B(\widetilde{\lambda}_A - \widetilde{\lambda}_B)}_{\text{A endpoint}} \tag{37}$$

follow by evaluating

$$2(g_A - g_B) - 2(d_A - d_B)r$$

at $r \simeq 0$, $r \simeq \widetilde{\lambda}_B$, and $r \simeq \widetilde{\lambda}_A$.

For the one-axis kurtosis example, the residual $R_t = M_\star - M(t)$ gives the first-axis residual $\psi(t) = \lambda_1 - v_1^\mathsf{T} M(t) v_1$ and

$$K_{R_t} = \kappa \psi(t) v_1 v_1^\mathsf{T}. \tag{38}$$

In the paper's loss normalization, the additional term relative to isotropic Gaussian inputs is $(\kappa/2)\psi^2$, which increases the penalty for leaving the first teacher component unfitted. For a rank-one student $M = muu^\mathsf{T}$, the common amplitude term cancels from the log ratio of the squared components. Subtracting their growth rates gives equation (13), and evaluating it at $\psi \simeq \lambda_1$ and $\psi = 0$ gives the two slopes of Section 5. At the mixed endpoint, the ratio stops changing when $\kappa\psi^\star = 2(\lambda_2 - \lambda_1)$. Together with amplitude stationarity, this gives equation (14), and Appendix F.1 shows that this endpoint is the global minimum.

Proposition F.2 (Exact solution for the one-axis kurtosis example)**.** *Let $\Sigma = I$, let v_1 be a teacher eigenvector, and suppose $K_A = \kappa(v_1^\mathsf{T} A v_1) v_1 v_1^\mathsf{T}$ for every symmetric A, as in the residual formula equation (38). Define*

$$Y(t, \Psi) = \exp\big(tM_\star + \tfrac{\kappa}{2}\Psi v_1 v_1^\mathsf{T}\big) U_0. \tag{39}$$

The pair (Ψ, N) is determined from

$$\dot{\Psi} = \lambda_1 - v_1^\mathsf{T} Y N^{-1} Y^\mathsf{T} v_1, \qquad \dot{N} = 2Y^\mathsf{T} Y, \qquad \Psi(0) = 0, \quad N(0) = S_0^{-1}. \tag{40}$$

Then $M = YN^{-1}Y^\mathsf{T}$ solves the exact population flow.

Proof. The residual growth matrix is

$$H_{M_\star - M} = M_\star - M + \frac{\kappa}{2}\psi v_1 v_1^\mathsf{T}.$$

Therefore

$$\dot{M} = G(t)M + MG(t) - 2M^2, \qquad G(t) = M_\star + \frac{\kappa}{2}\psi(t) v_1 v_1^\mathsf{T}.$$

The teacher and $v_1 v_1^\mathsf{T}$ commute, so the solution of $\dot{Y} = G(t)Y$ is the exponential in equation (39). The first equation in equation (40) gives $\dot{\Psi} = \psi$. Differentiating $YN^{-1}Y^\mathsf{T}$ with $\dot{N} = 2Y^\mathsf{T}Y$ completes the proof. □

F.1 Endpoint of the one-axis kurtosis conflict

Along the solution of proposition F.2, the early growth advantage of the first axis reverses and the student converges to a mixed direction. We show that this endpoint is the global rank-one minimum of the loss.

Proposition F.3 (Exact endpoint of the one-axis conflict)**.** *Let $d = 2$, $k = 1$, $\Sigma = I$, and $M_\star = \mathrm{diag}(\lambda_1, \lambda_2)$ with*

$$g\lambda_1 > \lambda_2 > \lambda_1 > 0, \qquad g = 1 + \frac{\kappa}{2} > 1.$$

Suppose only the first coordinate has excess kurtosis κ. Write $M = muu^\mathsf{T}$ with $u = (\cos\alpha, \sin\alpha)$. The global rank-one minima have $m^\star = \lambda_2$ and

$$\cos^2\alpha^\star = \frac{g\lambda_1 - \lambda_2}{(g-1)\lambda_2}.$$

The two signs of $\sin\alpha\cos\alpha$ give symmetry-related off-diagonal matrices.

Proof. Set $x = \cos^2 \alpha$. Up to a constant, the exact loss is

$$\mathcal{L}(x, m) = g(mx - \lambda_1)^2 + (m(1-x) - \lambda_2)^2 + 2m^2 x(1-x).$$

At fixed x, minimization over m gives

$$m(x) = \frac{\lambda_2 + (g\lambda_1 - \lambda_2)x}{1 + (g-1)x^2}.$$

Substitution shows that minimizing the loss is equivalent to maximizing

$$\frac{[\lambda_2 + (g\lambda_1 - \lambda_2)x]^2}{1 + (g-1)x^2}.$$

The interior stationary point is

$$x^\star = \frac{g\lambda_1 - \lambda_2}{(g-1)\lambda_2}.$$

The conflict inequalities put $x^\star$ strictly between zero and one. The derivative changes from positive to negative there, so it is the global maximizer. Substitution into $m(x)$ gives $m^\star = \lambda_2$. □

For $(\lambda_1, \lambda_2, g) = (0.8, 1.4, 2.5)$, the formula gives $(u_1^\star)^2 = 2/7$ and $(u_2^\star)^2 = 5/7$. The gradient flow in Figure 2e converges to these values.

G Exact solution for spherical fourth moments

The solution below is related to the exact solution of Martin et al. (2024) for an uncentered quadratic network with Gaussian inputs. In their model, the trace term comes from the uncentered loss, even for Gaussian data. Here it comes from the spherical fourth cumulant of the centered model.

For identity covariance and spherical fourth moments, $K_{R_t} = \bar{\kappa}(\mathrm{tr}(R_t)I + 2R_t)$ for $R_t = M_\star - M(t)$, so

$$H_t = (1 + \bar{\kappa})R_t + \frac{\bar{\kappa}}{2}\,\mathrm{tr}(R_t)I. \tag{41}$$

The student matrix and the scalar multiple of I vanish in $(I - P)H_tU$, leaving the direction equation in the main text.

Theorem G.1 (Direction path and matrix solution)**.** *Let $\Sigma = I$ and $K_A = \bar{\kappa}(\mathrm{tr}(A)I + 2A)$ with $\bar{\kappa} > -2/(d+2)$. Define $c = \bar{\kappa}/[2(1+\bar{\kappa})]$ and use rescaled time $\tau = (1+\bar{\kappa})t$. Dots in this section denote derivatives with respect to τ. Then*

$$\dot{M} = M_\star M + M M_\star - 2M^2 + 2c\theta M, \qquad \theta = \mathrm{tr}(M_\star - M).$$

For $M_0 = U_0 S_0 U_0^\mathsf{T}$, set

$$\phi(\tau) = \int_0^\tau \theta(v)\, dv, \qquad Y(\tau) = e^{c\phi(\tau)} e^{\tau M_\star} U_0,$$

and

$$N(\tau) = S_0^{-1} + 2\int_0^\tau e^{2c\phi(v)} U_0^\mathsf{T} e^{2vM_\star} U_0\, dv.$$

Let $T_\star = \mathrm{tr}\, M_\star$ and $G_0(\tau) = U_0^\mathsf{T} e^{2\tau M_\star} U_0$. The accumulated residual ϕ and the matrix N solve the closed initial-value problem

$$\dot{\phi} = T_\star - e^{2c\phi}\,\mathrm{tr}\big(N^{-1}G_0(\tau)\big), \qquad \dot{N} = 2e^{2c\phi}G_0(\tau), \qquad \phi(0) = 0, \quad N(0) = S_0^{-1}. \tag{42}$$

Then $M(\tau) = Y(\tau)N(\tau)^{-1}Y(\tau)^\mathsf{T}$, and its projector is exactly the Gaussian projector in rescaled time.

Proof. Equation 41 gives the stated matrix flow after rescaling. The constructed matrix satisfies $\mathrm{tr}\, M = e^{2c\phi}\,\mathrm{tr}(N^{-1}G_0)$, so the first equation in equation (42) gives $\dot{\phi} = \theta$. Therefore

$$\dot{Y} = (M_\star + c\theta I)Y, \qquad \dot{N} = 2Y^\mathsf{T}Y.$$

Differentiation of $YN^{-1}Y^\mathsf{T}$ gives the flow. The scalar factor $e^{c\phi}$ does not change the range of Y, so the projector is the Gaussian formula. □

Proposition G.2 (Aligned amplitudes and fixed point)**.** *Once the student aligns with the first k teacher eigenvectors and S is diagonal, the amplitudes obey equation (16). If all k retained amplitudes are positive and $R_k = \sum_{i>k} \lambda_i$, the fixed point is*

$$s_j^\infty = \lambda_j + c\frac{R_k}{1+kc}. \tag{43}$$

Proof. Restrict the matrix flow in theorem G.1 to a student aligned with the teacher and with diagonal S. This gives

$$\dot{s}_j = 2s_j(\lambda_j + c\theta - s_j).$$

At a positive fixed point, $s_j = \lambda_j + c\theta$. Since $\theta = \sum_{i>k} \lambda_i + \sum_{j\le k}(\lambda_j - s_j)$,

$$\theta = R_k - kc\theta = \frac{R_k}{1+kc}.$$

Substitution gives the stated amplitudes. □

Holding all modes beyond the first m at zero replaces k by m and R_k by R_m in the same calculation, which gives the stage equilibria

$$\theta^{(m)} = \frac{R_m}{1+mc}, \qquad s_j^{(m)} = \lambda_j + c\,\theta^{(m)}, \quad j \le m. \tag{44}$$

With small initialization and separated learning stages, the full trajectory passes near them.

H Data weights for independent input coordinates

We assume independent input coordinates and a teacher diagonal in the same basis, so that

$$\Sigma = \mathrm{diag}(d_1, \dots, d_d), \qquad M_\star = \mathrm{diag}(\lambda_1, \dots, \lambda_d), \tag{45}$$

$$K_{R_t} = \mathrm{diag}\big(d_a^2 \kappa_a (R_t)_{aa}\big)_{a=1}^d. \tag{46}$$

Write $w_a = 1 + \kappa_a/2 > 0$ and $f_a^2 = d_a^2 w_a$, so that the loss is equation (17). The weight along axis a initially grows at rate $f_a^2\lambda_a$, and its squared weight at twice that rate. Along a represented coordinate, the amplitude obeys $\dot{m}_a = 2f_a^2 m_a(\lambda_a - m_a)$ and settles at the fitted amplitude λ_a.

For a represented axis i and an unrepresented axis β, the fixed point is stable against rotating weight from i toward β exactly when

$$d_\beta w_\beta \lambda_\beta - d_i \lambda_i < 0. \tag{47}$$

The loss ranking compares only diagonal students. Proposition H.2 and the benchmark below test them under the full weight dynamics.

Proposition H.1 (Fitted amplitude, loss ranking and initial growth)**.** *Assume equations (45) and (46). Define $w_a = 1 + \kappa_a/2 > 0$ and $f_a = d_a\sqrt{w_a}$. For a student diagonal in the coordinate basis:*

1. *a represented axis has fitted amplitude λ_a;*
2. *among diagonal rank-k students, the one with the lowest loss represents the k axes with the largest $f_a\lambda_a$;*
3. *at small initialization, the weight along axis a grows at rate $f_a^2\lambda_a$, and its squared weight at rate $2f_a^2\lambda_a$.*

Proof. For a diagonal residual $A = \mathrm{diag}(a_1, \dots, a_d)$,

$$\mathcal{L}(A) = \sum_a d_a^2 w_a a_a^2.$$

If axis a is represented, minimizing its term sets its residual to zero and gives amplitude λ_a. If it is omitted, its loss contribution is

$$d_a^2 w_a \lambda_a^2 = (f_a\lambda_a)^2.$$

The best diagonal student therefore represents the axes whose omission costs the most. The initial growth matrix is diagonal with entry

$$(H_{M_\star})_{aa} = d_a^2 w_a \lambda_a = f_a^2 \lambda_a.$$

This is the growth rate of the weight, and the squared weight grows at twice this rate. □

Proposition H.2 (Stability against rotations between coordinate axes)**.** *At a diagonal rank-k fixed point, rotate a represented axis i slightly toward an unrepresented axis β. The perturbation grows at rate*

$$d_\beta(d_\beta w_\beta \lambda_\beta - d_i \lambda_i).$$

The fixed point is stable against every such rotation if and only if equation (47) holds for every represented i and unrepresented β.

Proof. The cross term changes M at first order but has zero diagonal, so the coordinate fourth cumulant does not act on it. The unrepresented axis gets the initial growth $d_\beta^2 w_\beta \lambda_\beta$, and the student feedback from the represented axis subtracts $d_\beta d_i \lambda_i$. Their difference gives the rate. □

H.1 Held-out ranking benchmark

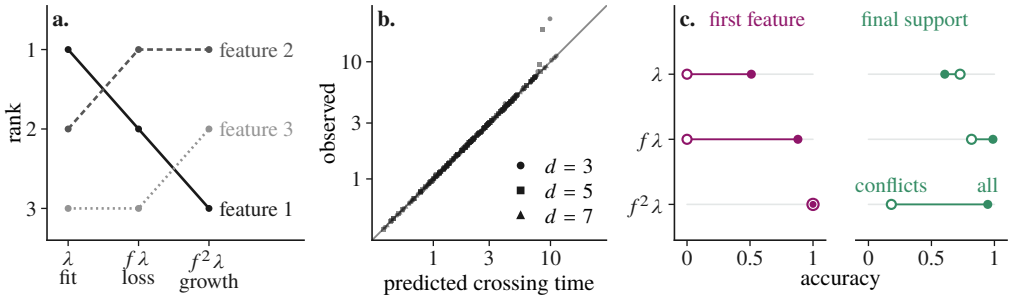


Figure A3: **Initial growth and eventual fit rank the same features differently.** (a) Three features ranked by fitted amplitude λ, loss reduction $f\lambda$, and initial weight growth $f^2\lambda$. (b) Held-out first-crossing times match the linearized prediction; marker shape gives the dimension. (c) $f^2\lambda$ predicts the first feature and $f\lambda$ the final features; filled markers are all problems, open markers the conflicts, in which the two rankings disagree.

Panel a uses $\lambda = (1.6, 1.2, 0.75)$ and $f = (0.55, 0.85, 1.05)$. Panel b shows 339 crossing pairs on logarithmic axes. With log-linear interpolation between reporting times, the observed crossings match the linearized prediction with median relative error 1.3×10^{-6}, and 94% lie within one percent. Panel c uses 192 problems, of which 170 converge. In 23 of them, the loss ranking $f\lambda$ and the growth ranking $f^2\lambda$ pick different axes, and 11 of these conflicts converge.

The rank-one test draws $d \in \{3, 5, 7\}$ uniformly, then draws

$$\lambda_a \sim \mathrm{LogUniform}(0.45, 2.2), \qquad d_a \sim \mathrm{LogUniform}(0.55, 1.9),$$
$$w_a \sim \mathrm{LogUniform}(0.60, 2.20).$$

The seed is 20260818. Before integration, we keep a draw only if the winners of the loss ranking and of the growth ranking each lead by at least two percent, and the loss winner satisfies equation (47) with a five percent margin. Of 352 draws, 144 fail the stability condition and 16 fail a margin condition, which leaves 192 test problems. Most exclusions come from the stability condition, because with covariance and kurtosis combined the diagonal loss winner is often unstable to rotations.

Every coordinate starts with squared weight $10^{-9}/d$. We integrate the weight flow with DOP853 to time 80, with relative tolerance 10^{-9} and absolute tolerance 10^{-12}. The first feature is the first coordinate whose squared weight crosses 10^{-5}. We report the final features only when the norm of the vector field is below 10^{-7}. Table A1 gives the result counts.

The median Kendall correlation between $f^2\lambda$ and the measured order of early growth is 1.00. In converged runs, the represented amplitudes match λ_a with median relative error 4.7×10^{-10}.

Outcome	Fitted amplitude λ	Loss ranking $f\lambda$	Initial growth $f^2\lambda$
First feature, all 192	51.0%	88.0%	100.0%
Final features, 170 converged	60.6%	98.8%	94.7%
First feature, 23 conflicts	0.0%	0.0%	100.0%
Final features, 11 converged conflicts	72.7%	81.8%	18.2%

Table A1: Held-out ranking results, as the fraction of problems in which each ranking predicts the outcome. Conflicts are problems in which the loss ranking and the growth ranking pick different axes. Unconverged runs count for the first feature and are left out of the final features.

I Experimental details and reproducibility

All runs use fixed seeds. We integrate every scalar closure from its own initial conditions, without using any values from the trajectory it predicts.

Figures 1 and A1c: isotropic dynamics. The rank-two student uses teacher eigenvalues $(2.4, 1.9, 1.2, 0.8, 0.5, 0.25, 0.1)$, one generic initial frame, and initialization scales $\varepsilon = 10^{-3}, \dots, 10^{-9}$. Direction accuracy is measured by $\sin^2\theta_{\max}$, the squared sine of the largest principal angle to the leading teacher subspace. It falls below 0.05 at $t = 2.89$ for every scale. The amplitude threshold in Figure 1c is half the weakest active teacher amplitude; its time has fitted slope 0.261 against the prediction $1/(2\lambda_2) = 0.263$. Figure A1c instead uses $\varepsilon = 10^{-7}$ and plots $1 - \sin^2\theta_{\max}$ alongside the largest eigenvalue of M divided by λ_1. Its shaded interval begins at $t = 2.892$ and ends when this amplitude fraction first reaches the fixed threshold $\alpha = 0.05$, at $t = 3.668$. This first-amplitude threshold precedes the weakest-amplitude threshold at $t = 5.04$. The closed matrix solution matches the flow to maximum relative error 3.4×10^{-8}.

Figure 2a,b: covariance switch. The conflict uses $\Sigma = \mathrm{diag}(1, 2.5, 0.7)$ and $M_\star = \mathrm{diag}(1, 0.28, 0.4)$. Feature B is coordinate 2, which has the largest early growth rate g_i, and feature A is coordinate 1, which has the largest $\widetilde{\lambda}_i$. The switch occurs at $t = 18.6$, and the three predicted log-odds slopes are -1.5, $+0.6$, and $+1.5$. The paths plot $\sqrt{q_i} = \sqrt{d_i}|w_i|$, and the weight fractions are $w_i^2 / \sum_j w_j^2$. The gray path in Figure 2a uses one dataset of size 1024. The scalar closure matches the matrix flow to 2.5×10^{-7} in q_i and 2×10^{-6} in the weight fractions.

Figure 3: spherical moments. Panel a uses the rank-two problem and three input laws described in the spherical sampled experiment below. Panels c and d instead use a rank-four aligned student, teacher spectrum $(3, 2.2, 1.5, 0.9, 0.6, 0.4)$, initial amplitudes 10^{-24}, and couplings $c = 0.30$ and $c = -0.12$. In panel d, s_4 starts to grow at rescaled time 14.24. The independently solved closure matches the flow to 2.3×10^{-7} and 3.1×10^{-5} in amplitude, and to 4.6×10^{-6} in this onset time. Black dashes mark the stage equilibria in equation (44).

Figure 2c–e: kurtosis and finite data. The one-axis kurtosis example in panels d and e uses $(\lambda_1, \lambda_2) = (0.8, 1.4)$ and $\kappa = 3$. The predicted slopes are -1.2 and $+1.2$, with zero first-axis residual at $t = 7.8$. The independent closure matches the flow to 1.7×10^{-8}. The numerical solver uses the equivalent normalization $Y(0) = U_0 S_0^{1/2}$ and $N(0) = I$, whereas the main text uses $Y(0) = U_0$ and $N(0) = S_0^{-1}$. Panel c uses 24 datasets per size and 12 minibatch runs at each batch size 64, 256, and 1024 on $n = 8192$.

I.1 Finite-sample feature switch

We fix the covariance example before drawing data:

$$\Sigma = \mathrm{diag}(1, 2.5, 0.7), \qquad M_\star = \mathrm{diag}(1, 0.28, 0.4).$$

For a dataset $\{x_i\}_{i=1}^n$, define $B_i = x_i x_i^\mathsf{T} - \Sigma$ and the sample growth matrix associated with a residual A by

$$\widehat{H}_A^{(n)} = \frac{1}{2n} \sum_{i=1}^{n} \langle B_i, A \rangle B_i. \tag{48}$$

Its expectation is $\Sigma A \Sigma$ under the Gaussian sampling law. We integrate

$$\dot{W} = \widehat{H}^{(n)}_{M_\star - WW^\top} W$$

to time 60. We center with the known covariance. Each initialization gives every coordinate the same squared weight $q_i = 10^{-8}$ and then applies a fixed log-normal perturbation and random signs.

The early prediction is the leading eigenvector of $\widehat{H}^{(n)}_{M_\star}$. The final prediction is the coordinate axis with the lowest empirical loss after its amplitude is optimized. A run counts as a complete switch when the capture of feature 2 exceeds 0.55 and the capture of feature 1 ends above 0.80. Dataset sizes are $128, 256, 512, 1024, 2048, 4096, 8192$, with 24 independent datasets and initializations at every size. Table A2 gives the results.

n	Complete switch	Early prediction	Final prediction	Median switch time
128	$17/24$	$22/24$	$23/24$	6.50
256	$18/24$	$22/24$	$22/24$	9.65
512	$23/24$	$24/24$	$24/24$	9.40
1024	$24/24$	$22/24$	$24/24$	10.25
2048	$24/24$	$24/24$	$24/24$	12.55
4096	$24/24$	$24/24$	$24/24$	13.45
8192	$24/24$	$24/24$	$24/24$	12.95

Table A2: Finite-dataset switch results. Switch time is censored for unsuccessful runs and summarized only among successful runs. Prediction columns compare empirical moment predictions with observed early and final dominant coordinates.

The minibatch test uses datasets of size 8192, batch sizes 64, 256, and 1024, 12 independent datasets and initializations per batch, Euler step 0.02, and horizon 60. All 36 runs satisfy the complete-switch criterion. This test checks that the switch is visible under minibatch training, and we do not claim a theorem for stochastic gradient descent.

I.2 Spherical sampled experiment

The spherical experiment uses $d = 6$, $k = 2$, teacher spectrum $(3, 2, 1.2, 0.7, 0, 0)$, and initial matrix scale 10^{-2}. The laws are Gaussian with $\bar{\kappa} = 0$, multivariate t_{12} with $\bar{\kappa} = 1/4$, and a Gaussian scale mixture with squared scale 4 with probability $1/7$ and squared scale $1/2$ otherwise, giving $\bar{\kappa} = 3/2$. Each law uses five independent runs and batch size 4096. The learning rate is divided by $1 + \bar{\kappa}$, and the horizontal axis is the rescaled population time. The final means of the leading amplitudes and their predicted values are

$$(2.9988, 3.1569, 3.3589) \quad \text{and} \quad (3.0000, 3.1583, 3.3563).$$