Decomposing SGD Dynamics in Neural Networks: Teacher-Induced Spikes and Variance Inflation

Samet Demir[1]*, **Zafer Doğan**[1,2]†
[1]MLIP Research Group, KUIS AI Center, Koç University [2]Department of EEE, Koç University
{sdemir20,zdogan}@ku.edu.tr

Abstract

We study two-layer neural networks trained by stochastic gradient descent (SGD) on a multi-index teacher model, focusing on how first-layer parameters orthogonal to the teacher subspace affect generalization. We decompose the first-layer weights into a signal component aligned with the teacher and an orthogonal complement, and analyze their joint SGD dynamics together with the second layer in the proportional limit. We show that although SGD updates to the complement are initially isotropic, the training dynamics induce teacher-dependent spectral spikes in the complement Gram matrix. This emergent anisotropy provides a concrete mechanism by which representation variance is amplified, and generalization degrades. To isolate this effect, we introduce Decomposed Dynamics, an analytically tractable surrogate whose asymptotic test-error dynamics coincide with those of SGD. In simulations with a ReLU student, freezing the directions of the orthogonal complement improves generalization relative to training them. Finally, we show that the signal component rapidly aligns with the second-layer weights, revealing strong cross-layer coupling even in this shallow setting. Together, our results provide a dynamical explanation for how feature learning outside the teacher subspace harms generalization.

1 Introduction

Understanding what stochastic gradient descent (SGD) *actually does* in neural networks—beyond reducing training loss—remains a central open problem in learning theory. Even for two-layer networks, which are simple enough to admit sharp asymptotic analysis, our theoretical understanding is fragmented across regimes and modeling assumptions. In the "lazy" or kernel regime, features remain nearly fixed during training, leading to kernel regression with random features or the neural tangent kernel (NTK) [1, 2]. At the opposite extreme, mean-field and particle-based approaches describe feature learning at infinite width via PDE or interacting particle dynamics [3, 4, 5, 6]. While both viewpoints have delivered major insights, they leave unresolved a basic structural question that arises repeatedly in teacher–student analyses of representation learning:

When the target depends on a low-dimensional teacher subspace, what role do the many first-layer parameters outside that subspace play in generalization?

Most existing teacher–student analyses focus on the alignment of learned features with teacher directions [7, 8, 9]. However, in high-dimensional overparameterized settings, the majority of first-layer degrees of freedom lie in the teacher-orthogonal complement. A common simplifying assumption—often justified by concentration or Gaussian equivalence arguments—is that this complement can be summarized by a few low-dimensional quantities or treated as isotropic noise. Yet SGD is

*Current affiliation: Oden Institute for Computational Engineering and Sciences, The University of Texas at Austin.
†Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

not passive: through coupling with the data and the second layer, it can reshape covariance structure, induce anisotropy, and affect test-time behavior in ways invisible if one tracks only teacher-aligned components.

In this work, we make the teacher-orthogonal complement explicit and study its dynamics directly. We decompose the first-layer weights into (i) a *signal matrix* supported on the teacher subspace and (ii) an orthogonal *complement matrix*. We analyze their joint SGD training with the second layer in the high-dimensional proportional limit. A central finding is that the Gram matrix of the complement does *not* remain bulk-like: SGD induces *spectral spikes*, revealing emergent anisotropic structure in the complement covariance. While reminiscent of finite-rank deformations and spiked random matrix models [10, 11, 12], here the spikes are *generated endogenously by training dynamics* rather than imposed by the data.

This characterization enables a tractable surrogate description of SGD. Combining our analysis of the complement Gram evolution with the unchanged SGD updates of the second layer and the signal matrix, we introduce **Decomposed Dynamics**, a simplified algorithm whose asymptotic test-error dynamics coincide with those of SGD. This surrogate isolates the effect of complement anisotropy. The decomposition also lets us address whether training the complement is beneficial for generalization. Counterintuitively, our experiments show that *freezing the complement directions* can outperform training them, consistent with avoiding variance inflation induced by complement anisotropy. We also identify a rapid alignment phenomenon in which the signal matrix becomes strongly coupled to the second-layer weights, underscoring that feature learning and last-layer adaptation are tightly intertwined even in shallow networks.

Beyond its theoretical implications, this mechanism suggests an explanation for why constraining or partially freezing feature representations can improve test performance in overparameterized models. Finally, our perspective connects naturally to stochastic-process approximations of SGD. Classical diffusion and stationary-distribution viewpoints explain aspects of implicit regularization [13, 14], while broader generalization phenomena—such as interpolation and double descent—highlight the importance of variance control [15, 16]. Our results sharpen this message for two-layer feature learning: the teacher-orthogonal complement provides a concrete mechanism by which SGD inflates variance through emergent anisotropy, and explicitly controlling—or freezing—this component can improve generalization.

Our contributions are summarized as follows:

1. Using the standard decomposition of the first-layer weights into a *signal matrix* and a teacher-orthogonal *complement matrix*, we analyze their *joint* SGD dynamics with the second layer in the high-dimensional proportional limit at the level of the complement Gram matrix.
2. We show that SGD injects coherent teacher-induced structure into the complement, which produces *spectral spikes* in its Gram matrix (Figure 1), revealing emergent anisotropic structure in teacher-orthogonal covariance and identifying a concrete mechanism for variance inflation.
3. Leveraging this characterization, we propose **Decomposed Dynamics**, a surrogate whose asymptotic generalization-error dynamics match those of SGD, and show in simulations that *fixing the complement directions* can outperform training them, while also uncovering fast alignment between the signal component and second-layer weights.

Related work — Prior scaling-limit analyses of SGD include the teacher-orthogonal complement, but typically through reduced state variables: the overlap matrices of a fixed- or growing-width student in the order-parameter ODEs of online learning [17, 7, 18], summary statistics such as overlaps, norms and the risk in high-dimensional limit theorems [19, 20, 21], two-time correlation and response functions in DMFT [22, 23, 24], and the distribution of neurons in infinite-width mean-field limits [3, 4, 5, 6]. These descriptions characterize alignment, risk and marginal dynamics, but do not resolve the teacher-induced structure of the covariance that SGD injects into the Gram matrix of the complement at proportional width, which is the state we track. Spikes in trained first-layer weights are known from one or a few large gradient steps, where they typically lie in teacher-relevant directions and help learning [25, 26, 9, 8]. Here, online SGD injects coherent structure into the complement itself, which carries no target signal and receives isotropic inputs, and Section 3.3 relates this structure to the test error. Appendix A gives a detailed discussion of the related work.

Notation — We write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$, where $\mathcal{F}_t$ denotes the σ-field generated by all randomness up to time t, including the initialization and samples $\{x_s\}_{s<t}$. In particular, (F_t, a_t) is $\mathcal{F}_t$-measurable,

so $\mathbb{E}_t$ averages over the new sample (x_t, y_t) only. For a matrix A, we use $\|A\|_2$, $\|A\|_F$, and $\mathrm{Tr}(A)$ to denote its operator norm, Frobenius norm, and trace, respectively. For an integer m, let $[m] := \{1, \dots, m\}$. We use standard big-O and little-o notation; write $A \lesssim B$ if $A \le CB$ for some constant C independent of the diverging dimensions; and write $\asymp$ to denote same order in the diverging dimensions.

2 Problem setting

We study the dynamics of stochastic gradient descent (SGD) in two-layer neural networks trained on data generated by a multi-index teacher. We describe the setting in detail below.

Teacher model (multi-index) — Let $x \sim \mathcal{N}(0, I_n)$ and consider a *multi-index* teacher of the form

$$y \;=\; g\big(\Xi^\top x\big), \;\; \textit{with} \;\; \Xi \in \mathbb{R}^{n \times k_\star}, \;\; \Xi^\top \Xi = I_{k_\star}, \tag{1}$$

where $k_\star = O(1)$ is fixed and $g : \mathbb{R}^{k_\star} \to \mathbb{R}$ is bounded and Lipschitz. Let $z_\star := \Xi^\top x \sim \mathcal{N}(0, I_{k_\star})$.

Student (two-layer neural network) — We consider the two-layer neural network model

$$\hat{y}(x;\, F, a) \;=\; \frac{1}{k} \sum_{i=1}^{k} a_i \, \sigma\big(f_i^\top x\big), \tag{2}$$

where $F = [f_1, \dots, f_k]^\top \in \mathbb{R}^{k \times n}$ and $a \in \mathbb{R}^k$ are the first and second layer weights. These parameters are trained on the squared loss $\ell(\hat{y}, y) := \frac{1}{2}(\hat{y}(x;\, F, a) - y)^2 + \frac{\lambda}{2k}\|a\|_2^2$ with regularization constant λ, and we evaluate performance using the generalization error (true risk) $\mathcal{L}_t := \mathbb{E}_{x,y}[\frac{1}{2}(\hat{y}(x;\, F_t, a_t) - y)^2]$.

Training We study the training dynamics of stochastic gradient descent (SGD) that uses a single sample (x_t, y_t) to update the parameters (F_t, a_t) at each step t (with step sizes $\eta_f, \eta_a > 0$) as follows:

$$f_{i,t+1} = f_{i,t} - \eta_f \nabla_{f_i} \ell_t \;=\; f_{i,t} - \frac{\eta_f}{k} \, e_t \, a_{i,t} \, \sigma'(f_{i,t}^\top x_t) \, x_t, \tag{3}$$

$$a_{t+1} = a_t - \eta_a \nabla_a \ell_t \;=\; a_t - \frac{\eta_a}{k} \Big(e_t \, \sigma(F_t x_t) + \lambda a_t \Big), \tag{4}$$

for all $i \in [k]$, where $\ell_t := \ell(\hat{y}(x_t; F_t, a_t), y_t)$, $e_t = \hat{y}(x_t; F_t, a_t) - y_t$ and $\sigma(F_t x_t) := (\sigma(f_{1,t}^\top x_t), \dots, \sigma(f_{k,t}^\top x_t))^\top$. Note that we will drop the time step t from our notation whenever it is not necessary. Each sample is fresh, i.e., independent of $\mathcal{F}_t$, so our analysis covers online one-pass SGD with batch size one. Extending it to sample reuse is an open direction.

Subspace decomposition — To explicitly identify the degrees of freedom that govern feature learning and generalization in the overparameterized regime, we decompose the first-layer weights into components aligned with the teacher subspace and its orthogonal complement. Prior work [27, 28, 25, 26, 29, 30] indicates that feature learning in the first layer is closely related to the teacher directions Ξ in (1). Motivated by this, we study the SGD dynamics of F_t through a decomposition into parts living in $\mathrm{span}(\Xi)$ and its orthogonal complement. Let $P_\parallel := \Xi\Xi^\top$ and $P_\perp := I - P_\parallel$ denote the orthogonal projectors onto the teacher subspace and its orthogonal complement, respectively. Each row of F_t admits the decomposition

$$f_{i,t} = \Xi m_{i,t} + f_{i,t}^\perp, \qquad m_{i,t} := \Xi^\top f_{i,t} \in \mathbb{R}^{k_\star}, \quad f_{i,t}^\perp := P_\perp f_{i,t} \in \mathbb{R}^n. \tag{5}$$

Define $r_{i,t} := \|f_{i,t}^\perp\|_2$ and $u_{i,t} := f_{i,t}^\perp / \|f_{i,t}^\perp\|_2 \in \mathrm{span}(\Xi)^\perp$ (with the convention $u_{i,t} \equiv 0$ when $r_{i,t} = 0$, an event of probability zero throughout under our assumptions). Let $M_t \in \mathbb{R}^{k \times k_\star}$ and $U_t \in \mathbb{R}^{k \times n}$ be the matrices with rows $m_{i,t}^\top$ and $u_{i,t}^\top$, respectively, and define $r_t := (r_{1,t}, \dots, r_{k,t})^\top$. In the remainder of this work, we focus on the joint dynamics of $(a_t, M_t, F_t^\perp)$, where the *complement matrix* is

$$F_t^\perp \;:=\; F_t P_\perp \;=\; \big[f_{1,t}^\perp, \cdots, f_{k,t}^\perp\big]^\top \;=\; \mathrm{diag}(r_t)\, U_t \;\in \mathbb{R}^{k \times n}. \tag{6}$$

2.1 Assumptions

Here, we list the standing assumptions of our theoretical development.

A1. Data samples (x_t, y_t) distributed as $x_t \overset{iid}{\sim} \mathcal{N}(0, I_n)$ and $y_t = g(\Xi^\top x_t)$.

A2. Teacher function $g : \mathbb{R}^{k_\star} \to \mathbb{R}$ is bounded, $k_\star = O(1)$, and g is L_g-Lipschitz with respect to $\|\cdot\|_2$ on $\mathbb{R}^{k_\star}$.

A3. The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ lies in $W^{3,\infty}(\mathbb{R})$; that is, σ, σ', σ'', and σ''' are bounded.

A4. The input dimension n and width of the neural network k jointly diverge with finite ratio, over training horizons proportional to n: $n, k \to \infty$ with $k/n \to \alpha \in (0, \infty)$ and $t \leq \tau n$ for fixed $\tau \in (0, \infty)$, with step sizes $\eta_a, \eta_f = O(1)$ and regularization $\lambda \geq 0$ fixed constants.

A5. The neural network initialization satisfies $f_{i,0} \overset{iid}{\sim} \mathcal{N}(0, I_n/n)$ and $a_{i,0}$ i.i.d. sub-Gaussian with parameter $K_{\mathrm{sg}} < \infty$ and $\mathbb{E}[a_{i,0}^2] > 0$, independent of F_0.

Discussion of assumptions — Boundedness of σ'' is needed for the Gaussian-interpolation and Stein's identity steps in the proofs of Lemmas 3.1–3.2 and Theorem 3.4 (which also uses bounded σ'''), the Gaussian–Sobolev bound for the conditional drift in Appendix E, and the linearization in Proposition 3.6. Extensions to non-smooth activations (e.g., ReLU) are not covered by the present analysis; see Appendix J.

A derived event — The following is a *consequence* of **A1**–**A5** together with the SGD updates (3)–(4), established as Lemma I.1 in Appendix I; we name it here as a labeled high-probability event so that downstream theoretical results can work on it.

A6. *Moment stability of the SGD trajectory.* For each fixed $\tau > 0$ there exists $B_\tau < \infty$ such that, with probability $1 - o(1)$ as $n \to \infty$, $\sup_{t \leq \tau n} \frac{1}{k} \sum_{i=1}^k (a_{i,t}^2 + \|m_{i,t}\|_2^2 + r_{i,t}^2) \leq B_\tau$.

3 Theoretical results

In this section, we present our main theoretical results. We begin by characterizing the covariance (diffusion) of the row-wise SGD updates of the *complement matrix* $F_t^\perp$. Although the update direction $g_t = P_\perp x_t \sim \mathcal{N}(0, P_\perp)$ and the initial $F_0^\perp$ are isotropic in the complement, the $k \times k$ covariance of the per-neuron drivers is not. We show that training thereby induces *teacher-dependent structure*, which leads to the emergence of spectral spikes (Figure 1). In particular, Lemmas 3.1–3.2 show that the covariance injected at every step contains a coherent, teacher-conditioned part S_t.

Building on this covariance characterization together with the mean (drift) of the complement updates, we introduce an analytically tractable surrogate algorithm, termed *Decomposed Dynamics*, whose test-error dynamics are asymptotically equivalent to those of SGD under mild conditions. We then quantify the resulting variance inflation caused by teacher-induced anisotropy in the complement. Finally, in an idealized single-index setting, we show that the signal vector m_t rapidly aligns with the second-layer weights a_t, highlighting strong cross-layer coupling in the joint training of the two layers.

For clarity of exposition, we state the main objects and results in the single-index case ($k_\star = 1$), with the extension to the multi-index setting deferred to Appendix J.1; all results except Proposition 3.6 extend to any fixed $k_\star$. In the single-index reduction, we write $\Xi = \xi \in \mathbb{R}^n$ and decompose the input as $x_t = z_{0,t}\xi + g_t$, where $z_{0,t} \sim \mathcal{N}(0, 1)$ and $g_t \sim \mathcal{N}(0, P_\perp)$ are independent. The teacher output then takes the form $y_t = g(z_{0,t})$.

3.1 Geometry of updates to the complement matrix

Define the per-neuron driver $v_{i,t} := e_t\, a_{i,t}\, \sigma'(f_{i,t}^\top x_t)$, and let $v_t = (v_{i,t})_{i=1}^k \in \mathbb{R}^k$. Writing $g_t := P_\perp x_t$, the complement update takes the rank-one form

$$F_{t+1}^\perp \;=\; F_t^\perp - \frac{\eta_f}{k}\, v_t g_t^\top, \tag{7}$$

showing that the geometry of the complement updates is governed by the accumulation of $v_t v_t^\top$. We now characterize this accumulation and show how it decomposes into a centered bulk and teacher-induced structure.

Lemma 3.1 (Complement geometry: normalized bulk isotropy and teacher-induced structure)**.** *Assume **A1–A5** and work on the event **A6**. Fix $\tau > 0$ and let $t \le \tau n$. Decompose $v_s = \mu_s + \tilde{v}_s$ with $\mu_s := \mathbb{E}_s[v_s \mid z_{0,s}]$ (an average over g_s only) and $\mathbb{E}_s[\tilde{v}_s \mid z_{0,s}] = 0$, and define the teacher-induced structure $S_s := \mathbb{E}_s[\mu_s \mu_s^\top] \in \mathbb{R}^{k\times k}$. Let*

$$\widetilde{C}_t := \frac{1}{n} I_k + \Big(\frac{\eta_f}{k}\Big)^2 \sum_{s<t} \tilde{v}_s \tilde{v}_s^\top, \qquad C_t := \frac{1}{n} I_k + \Big(\frac{\eta_f}{k}\Big)^2 \sum_{s<t} v_s v_s^\top. \tag{8}$$

Then, there exists $\varepsilon_n \to 0$ such that, with probability $1 - o(1)$,

$$\sup_{s\le t} \frac{1}{k^2} \|\widetilde{C}_s - c_s I_k\|_F^2 \le \varepsilon_n, \qquad c_s := \frac{1}{k} \operatorname{Tr}(\widetilde{C}_s), \tag{9}$$

$$C_t = \widetilde{C}_t + \Big(\frac{\eta_f}{k}\Big)^2 \sum_{s<t} S_s + R_t, \qquad \sup_{s\le \tau n} \|R_s\|_F \le \varepsilon_n. \tag{10}$$

Proof sketch. Lemma B.1 implies negligible normalized off-diagonal drift in $\widetilde{C}_t$ over $t \le \tau n$. Writing $\widetilde{C}_t$ as drift plus martingale, stability (**A6**) and bounded σ, σ' yield uniform control and (9). Expanding $v_s = \mu_s + \tilde{v}_s$ gives (10), with R_t controlled similarly (Appendix C). ■

Term by term, in $v_s v_s^\top = \tilde{v}_s \tilde{v}_s^\top + \mu_s \mu_s^\top + \mu_s \tilde{v}_s^\top + \tilde{v}_s \mu_s^\top$ the first term builds $\widetilde{C}_t$, the conditional mean S_s of the second builds the structured term, and the cross terms together with $\mu_s \mu_s^\top - S_s$ form R_t. With fresh samples these are martingale differences, so R_t is a martingale of vanishing Frobenius norm, whereas $S_s \succeq 0$ accumulates coherently. "Teacher-induced" refers to the conditioning on $z_{0,s}$ in μ_s; the neuron-space directions of S_s are set by the SGD state. The accumulated structure $(\eta_f/k)^2 \sum_{s<t} S_s$ admits a uniform-in-time finite-mode approximation (Appendix E).

We next analyze how this accumulated structure propagates to the Gram matrix of the complement $F_t^\perp$ and gives rise to spectral spikes.

Lemma 3.2 (One-step update of the complement Gram)**.** *Assume **A1–A5** and work on the event **A6**. Fix $\tau > 0$ and let $t \le \tau n$, with $n_\perp := \operatorname{Tr}(P_\perp) = n - 1$. Define the normalized complement Gram*

$$G_t := \frac{1}{n_\perp} F_t^\perp (F_t^\perp)^\top \in \mathbb{R}^{k\times k}, \tag{11}$$

the complement preactivations $h_t^\perp := F_t^\perp g_t \in \mathbb{R}^k$ with $g_t := P_\perp x_t \sim \mathcal{N}(0, P_\perp)$, the driver covariance $K_t := \mathbb{E}_t[v_t v_t^\top] = S_t + \mathbb{E}_t[\tilde{v}_t \tilde{v}_t^\top]$, and the Stein operator $J_t := \mathbb{E}_t[\partial v_t / \partial h_t^\perp] \in \mathbb{R}^{k\times k}$, i.e., $(J_t)_{ij} = \frac{1}{k}\mathbb{E}_t[a_{i,t} a_{j,t} \sigma'(h_{i,t})\sigma'(h_{j,t})] + \delta_{ij}\, a_{i,t} \mathbb{E}_t[e_t \sigma''(h_{i,t})]$, with $h_{i,t} := f_{i,t}^\top x_t$. Then, with probability $1 - o(1)$ uniformly over $t \le \tau n$,

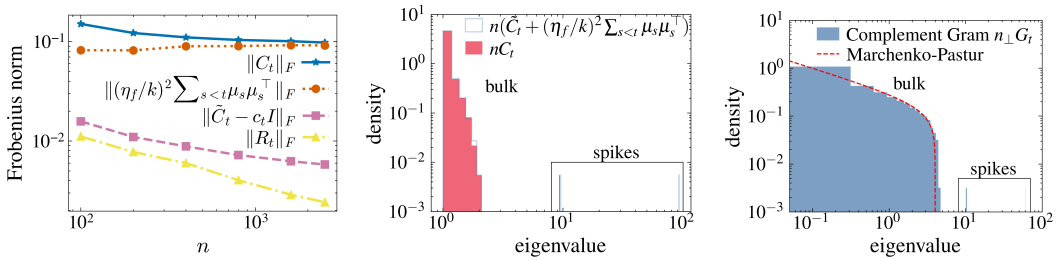
$$\mathbb{E}_t[G_{t+1}] = G_t - \frac{\eta_f}{k}\big(J_t G_t + G_t J_t\big) + \Big(\frac{\eta_f}{k}\Big)^2 K_t + E_t, \tag{12}$$

where

$$\sup_{t\le \tau n} \|E_t\|_F = O\big(n^{-3/2}\big). \tag{13}$$

Proof sketch. Expanding $G_{t+1} = \frac{1}{n_\perp} F_{t+1}^\perp (F_{t+1}^\perp)^\top$ with $F_{t+1}^\perp = F_t^\perp - (\eta_f/k) v_t g_t^\top$ yields G_t, linear terms in $\mathbb{E}_t[v_t h_t^{\perp\top}]$ and its transpose, and a quadratic term $(\eta_f/k)^2 \mathbb{E}_t[(\|g_t\|_2^2/n_\perp) v_t v_t^\top]$. Given $(\mathcal{F}_t, z_{0,t})$, $h_t^\perp \sim \mathcal{N}(0, n_\perp G_t)$, so Gaussian integration by parts gives exactly $\mathbb{E}_t[v_t h_t^{\perp\top}] = n_\perp J_t G_t$. Gaussian norm concentration gives the quadratic term as $(\eta_f/k)^2 K_t + O(n^{-3/2})$; see Appendix D. ■

Remark 3.3 (Bulk and teacher-induced spikes in the complement Gram). Lemmas 3.1 and 3.2 separate the mechanisms shaping the spectrum of G_t. First, Lemma 3.1 shows that, up to the centered accumulator $\widetilde{C}_t$ and a vanishing remainder, C_t is the accumulated teacher-induced component $(\eta_f/k)^2 \sum_{s<t} S_s$, which admits a uniform-in-time finite-mode approximation (Appendix E). Second, Lemma 3.2 yields the recursion $\mathbb{E}_t[G_{t+1}] = G_t - \frac{\eta_f}{k}(J_t G_t + G_t J_t) + (\eta_f/k)^2 K_t$, up to $O(n^{-3/2})$ error. The J_t term damps the current Gram through the first (Gauss–Newton) summand of J_t, while $(\eta_f/k)^2 K_t$ injects a positive semidefinite covariance at each step. The coherent part of K_t, the approximately low-rank teacher-induced structure S_t (Appendix E), creates outliers above a



(a) Frobenius norm of C_t (b) Spectral density of nC_t (c) Spectral density of $n_\perp G_t$

Figure 1: Visualization of Lemma 3.1 and Lemma 3.2. Panels (a)–(b) illustrate how the accumulation of teacher-induced structure, captured by the driver accumulator C_t in (8), leads to bulk-plus-spike behavior in its spectral density, in line with Lemma 3.1. Panel (c) shows the resulting bulk-plus-spike structure in the spectrum of $n_\perp G_t$, in line with Remark 3.3. Norms in (a) are unnormalized, and panels (a)–(b) use the realized $\mu_s \mu_s^\top$ in place of S_s. Parameters are $\alpha = 1.0$, $\tau = 2.0$, $\lambda = 0.01$, $k_\star = 1$, $g(\cdot) = \tanh(\cdot)$, $\sigma = \mathrm{ReLU}$, and $\eta_a = \eta_f = 0.8$ throughout, with input dimension $n = 1000$ for panels (b)–(c).

Marchenko–Pastur bulk in the spectrum of G_t; the lemmas establish this mechanism, and Figure 1c shows the resulting spectrum.

Figure 1 illustrates this bulk-plus-spikes phenomenon. As the input dimension n increases, $\|\widetilde{C}_t - c_t I_k\|_F$ and $\|R_t\|_F$ decrease, while the accumulated teacher-induced term $(\eta_f/k)^2 \sum_{s<t} S_s$ remains non-negligible and dominates the norm of C_t. The resulting spectrum of $n_\perp G_t = F_t^\perp (F_t^\perp)^\top$ exhibits a Marchenko–Pastur bulk together with additional spikes, in agreement with Remark 3.3.

How large are the spikes? — A natural question is whether the outliers in Figure 1c are large enough to matter or merely a finite-size effect. In the setting of Figure 1c, over five seeds (mean $\pm$ standard deviation), the top eigenvalue of $n_\perp G_t$ is $\lambda_1 = 63.3 \pm 1.6$, far above the Marchenko–Pastur edge 4 of the initial spectrum. Its per-neuron share $\lambda_1/k \approx 0.063$ is about 5.5% of the mean complement variance $\frac{1}{k}\sum_i r_{i,t}^2$ but $40 \pm 3\%$ of the mean signal variance $\frac{1}{k}\sum_i m_{i,t}^2$: the spike is small compared with the complement as a whole, yet comparable to the teacher-aligned signal. Halving n roughly halves λ_1 (ratio 0.52 ± 0.02), so λ_1/k stays nearly constant as n grows.

3.2 Equivalence of SGD and Decomposed Dynamics in test-error dynamics

Next, we focus on the characterization of the test-error (generalization) dynamics of the SGD by studying the updates of the complement $F_t^\perp$. Specifically, in our setting, if two algorithms have the same drift (the conditional mean of the update) and diffusion (its neuron-space second moment), then, given the stability estimates of Lemma F.1, they perform equivalently on the test metrics of interest (mainly generalization error), in the spirit of propagation-of-chaos for proportional-regime SGD [31, 32]. In the following theorem, we use this property to replace the original update of the complement matrix $F_t^\perp$ with a simpler surrogate update rule while preserving the test-error dynamics in the asymptotic limit.

Theorem 3.4 (Equivalence of SGD and Decomposed Dynamics in learning curves)**.** *Assume **A1–A5** and $|a_{i,0}| \le A$ almost surely for a constant A. Consider two training procedures initialized from the same (F_0, a_0):*

1. ***(SGD)**: stochastic gradient descent updates* (3)–(4).
2. ***(Decomposed Dynamics)**: a surrogate that (i) updates $(a_t^{\mathrm{DD}}, M_t^{\mathrm{DD}})$ by the SGD rule (Appendix F.2), and (ii) evolves its complement matrix $(F_t^\perp)^{\mathrm{DD}}$ by the following update, which has the same conditional mean as the SGD complement update:*

$$(F_{t+1}^\perp)^{\mathrm{DD}} \;=\; (I - \widetilde{\Lambda}_t)\,(F_t^\perp)^{\mathrm{DD}} \;+\; \zeta_t, \tag{14}$$

where $\mathrm{vec}(\zeta_t^\top) \sim \mathcal{N}(0, \Gamma_t \otimes P_\perp)$ is generated from fresh standard Gaussian innovations, independent across time and of the fresh sample driving the (a_t, M_t) updates.

Furthermore, $\widetilde{\Lambda}_t$ and Γ_t are defined as

$$\Gamma_t \;:=\; \Big(\frac{\eta_f}{k}\Big)^2 \mathbb{E}_t\big[v_t v_t^\top\big] \in \mathbb{R}^{k\times k}, \qquad \widetilde{\Lambda}_t \;:=\; \frac{\eta_f}{k}\, J_t, \tag{15}$$

with J_t the Stein operator of Lemma 3.2; both are evaluated at the current DD state.

Then, for any fixed $\tau > 0$, $\sup_{0\le t\le \tau n} \big|\mathcal{L}_t^{\mathrm{SGD}} - \mathcal{L}_t^{\mathrm{DD}}\big| \xrightarrow{\mathbb{P}} 0$.

Proof. Decomposed Dynamics matches (i) the conditional drift and diffusion of (a_t, M_t), (ii) the complement second moment via $\mathrm{vec}(\zeta_t^\top) \sim \mathcal{N}(0, \Gamma_t \otimes P_\perp)$ with (15), and (iii) the conditional mean of the SGD complement increment, which by Stein's identity is $-\widetilde{\Lambda}_t F_t^\perp$. Hence the conditional mean increments of the reduced state (a_t, M_t, Σ_t) with $\Sigma_t = F_t^\perp (F_t^\perp)^\top$, which determines the test loss, agree up to $O(n^{-3/2})$ per step in the norm $\|\cdot\|_\star$ of Appendix F.4. By Lemma F.1, a discrete Grönwall argument transports this to the test loss. The full proof is provided in Appendix F. ■

Figure 2 illustrates our numerical results comparing the SGD with the equivalent surrogate (Decomposed Dynamics) for different learning rates $\eta = \eta_f = \eta_a$, various $\alpha = k/n$ ratios, and changing activation functions $\sigma(\cdot)$. The results affirm that the Decomposed Dynamics closely match the learning curves of the SGD in terms of the generalization (test) error under various conditions, even in moderate dimensions. This surrogate is useful for characterizing the behavior of the complement matrix $F_t^\perp$, as it separates the injected covariance Γ_t, which contains the teacher-induced part $(\eta_f/k)^2 S_t$, from the drift $-\widetilde{\Lambda}_t F_t^\perp$.

3.3 Variance inflation by the teacher-induced structure

Next, we study the impact of the teacher-induced complement structure (spikes) on the generalization error in the following proposition. To begin, we observe that $h_t^\perp := F_t^\perp g_t \sim \mathcal{N}(0, F_t^\perp (F_t^\perp)^\top)$ for a given F_t. Remark 3.3 and Figure 1 indicate that the complement Gram admits an approximate bulk-plus-structure decomposition: $n_\perp G_t = F_t^\perp (F_t^\perp)^\top \approx \hat{c}_t I_k + B_t^{\mathrm{acc}} (B_t^{\mathrm{acc}})^\top$ where $\hat{c}_t I_k$ captures the isotropic (bulk) part while $B_t^{\mathrm{acc}} (B_t^{\mathrm{acc}})^\top$ captures the accumulated teacher-induced structure (spikes), with B_t^{acc} the accumulated loading matrix defined in (54) of Appendix E, and $\hat{c}_t$ the bulk-eigenvalue level of $n_\perp G_t$ (with $\hat{c}_0 \approx 1$ from the i.i.d. Gaussian initialization). Replacing the covariance of $h_t^\perp$ by this approximation, we define an effective predictor as follows:

$$\hat{y}_t^{\mathrm{eff}} \;:=\; \frac{1}{k}\sum_{i=1}^{k} a_{i,t}\, \sigma\Big(m_{i,t}^\top z_\star + \big(\sqrt{\hat{c}_t}\varepsilon + B_t^{\mathrm{acc}}\, s\big)_i\Big), \tag{16}$$

where $z_\star \sim \mathcal{N}(0, I_{k_\star})$, $\varepsilon \sim \mathcal{N}(0, I_k)$, $s \sim \mathcal{N}(0, I)$ are independent random vectors, independent of $\mathcal{F}_t$ (the dimension of s is the column count of B_t^{acc}).

Proposition 3.5 (Variance Inflation by Teacher-Induced Complement Structure)**.** *Assume **A1–A5** and work on the event **A6**; fix $\tau > 0$, $t \le \tau n$. Let $\mathcal{F}_t$ be the σ-field generated by the randomness up to time t so that $(a_t, M_t, \hat{c}_t, B_t^{\mathrm{acc}})$ are measurable (deterministic). Define the corresponding effective test loss*

$$\mathcal{L}_t^{\mathrm{eff}} \;:=\; \frac{1}{2}\,\mathbb{E}\Big[\big(\hat{y}_t^{\mathrm{eff}} - g(z_\star)\big)^2 \,\Big|\, \mathcal{F}_t\Big]. \tag{17}$$

Then $\mathcal{L}_t^{\mathrm{eff}}$ admits the decomposition

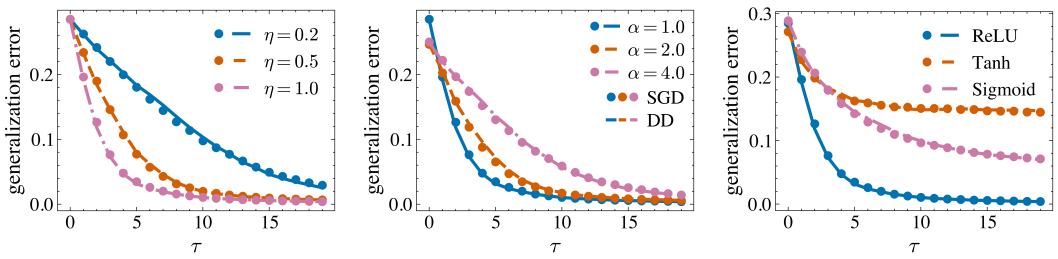
$$2\mathcal{L}_t^{\mathrm{eff}} \;=\; \mathcal{L}_{\mathrm{bias},t} \;+\; \mathcal{L}_{\mathrm{res},t} \;+\; \mathcal{L}_{\mathrm{str},t}, \tag{18}$$

where

$$\mathcal{L}_{\mathrm{bias},t} := \mathbb{E}_{z_\star}\left[\Big(\mathbb{E}[\hat{y}_t^{\mathrm{eff}} \mid z_\star, \mathcal{F}_t] - g(z_\star)\Big)^2\right], \quad \mathcal{L}_{\mathrm{res},t} := \mathbb{E}_{z_\star}\mathbb{E}_s\big[\mathrm{Var}_\varepsilon\big(\hat{y}_t^{\mathrm{eff}} \mid z_\star, s, \mathcal{F}_t\big)\big], \tag{19}$$

$$\mathcal{L}_{\mathrm{str},t} := \mathbb{E}_{z_\star}\Big[\mathrm{Var}_s\Big(\mathbb{E}_\varepsilon[\hat{y}_t^{\mathrm{eff}} \mid z_\star, s, \mathcal{F}_t]\Big)\Big] \;\ge\; 0. \tag{20}$$

Moreover, $\mathcal{L}_{\mathrm{str},t} = 0$ whenever $B_t^{\mathrm{acc}} = 0$ (no teacher-induced structure in the complement channel). Let $\omega_{i,t}(z_\star) := a_{i,t}\,\mathbb{E}_\varepsilon[\sigma'(m_{i,t}^\top z_\star + \sqrt{\hat{c}_t}\,\varepsilon_i)]$ be the slope-weighted readout. Then $\mathcal{L}_{\mathrm{str},t} > 0$ whenever $(B_t^{\mathrm{acc}})^\top \omega_t(z_\star) \neq 0$ for some $z_\star \in \mathbb{R}^{k_\star}$.



(a) learning rate $\eta_a = \eta_f = \eta$ ($\alpha = 1.0$ and $\sigma = \mathrm{ReLU}$)

(b) width $\alpha = k/n$ ($\eta_a = \eta_f = 1.0$, $\sigma = \mathrm{ReLU}$)

(c) activation function σ ($\eta_a = \eta_f = 1.0$ and $\alpha = 1.0$)

Figure 2: Illustration of Theorem 3.4: asymptotic equivalence of test-error curves for the SGD and the Decomposed Dynamics under various conditions. Namely, (a) shows the effect of different learning rate $\eta_f = \eta_a = \eta$, (b) demonstrates the impact of different width $\alpha := k/n$, and finally, (c) indicates the influence of the activation function $\sigma(\cdot)$. Here, $n = 500, \tau = 20.0, k_\star = 1, g(\cdot) = \mathrm{ReLU}(\cdot)$, $\lambda = 0$. Markers show SGD and lines show Decomposed Dynamics (DD).

Proof. This is an iterated law-of-total-variance identity: first condition on $z_\star$, then condition on $(z_\star, s)$ and integrate over ε. Full proof is given in Appendix G. ■

Proposition 3.5 isolates the variance of the predictor along the structure (spike) directions, $\mathcal{L}_{\mathrm{str},t} \geq 0$, which we call "teacher-induced variance inflation". In Section 4 (Figure 3a), we illustrate that the Decomposed Dynamics with fixed U_t (fixed complement directions dropping the teacher-induced spikes from the learning dynamics similar to setting $B_t^{\mathrm{acc}} = 0$) achieves better (lower) test-error curve than the SGD, consistent with the prediction that teacher-induced structure can increase the test error by variance inflation.

3.4 Coupling of the first and second layer weights

Finally, we investigate the coupling of the first layer F_t and the second layer a_t in an idealized (oracle) version of the dynamics in the following proposition. Focusing on the single-index case $k_\star = 1$, where F_t is decomposed as $F_t = m_t \xi^\top + F_t^\perp$ for some $m_t \in \mathbb{R}^k$, we show that m_t and a_t get aligned quickly during the early steps of the training. Let $\kappa_g := \mathbb{E}[g(z)z]$ and $\kappa_\sigma := \mathbb{E}[\sigma'(z)]$, $z \sim \mathcal{N}(0,1)$, be the first Hermite coefficients of g and σ, and let $\cos_t := a_t^\top m_t / (\|a_t\|_2 \|m_t\|_2)$.

Proposition 3.6 (Fast alignment of m_t and a_t)**.** *Let $k_\star = 1$ and assume **A1–A5** with $\eta_a, \eta_f > 0$ and $|a_{i,0}| \leq A$ almost surely for a constant A. Suppose that g and σ are odd with $\kappa_g \kappa_\sigma \neq 0$, and let (a_t, m_t) follow the oracle no-spike dynamics of Appendix H, in which the complement channel is replaced by synthetic independent standard Gaussians. Then for every $\varepsilon > 0$ there is $\tau_0(\varepsilon) > 0$ such that, with probability $1 - o(1)$ over initialization,*

$$1 - \cos_t^2 \;\leq\; \varepsilon \qquad \text{for all } k^{2/3} \leq t \leq \tau_0(\varepsilon)\, k. \tag{21}$$

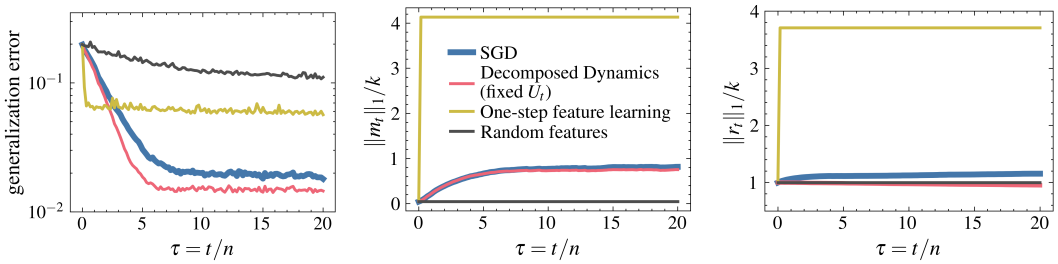
In particular, m_t aligns with a_t within $o(k)$ steps.

Proof. See Appendix H. ■

Here, Proposition 3.6 states that the teacher-aligned part of the first layer lines up with the second layer within a vanishing fraction of training. While the proposition is proved for the oracle no-spike dynamics, Figure 4b shows similarly rapid alignment under SGD.

4 Additional simulation results

In this section, to illustrate the characteristics of feature learning, we compare the dynamics of four versions of two-layer neural networks: (i) the lazy training regime, where the first layer is fixed at initialization (random features [1]); (ii) one-step feature learning, where the first layer is trained with one giant step of gradient descent and then frozen, as used in the literature to explain feature learning [25, 26, 9, 8, 29]; (iii) joint SGD training in the mean-field regime, following (3)–(4); and (iv)



(a) generalization (b) mean $|m|$ (subspace strength) (c) mean r (complement norm)

Figure 3: Comparison of learning algorithms in the decomposition variables. Panels show (a) test error, (b) mean teacher-subspace strength $\|m_t\|_1/k$, and (c) mean complement norm $\|r_t\|_1/k$ for full SGD, Decomposed Dynamics with fixed complement directions U_t, one-step feature learning, and random features. Parameters are $n = 300$, $\alpha = 1.0$ ($k = 300$), $\tau = 20.0$, $k_\star = 1$, $g(\cdot) = \tanh(\cdot)$, $\sigma(\cdot) = \mathrm{ReLU}(\cdot)$, $\eta_a = \eta_f = 1.0$ for the SGD/DD baselines, ridge $\lambda = 10^{-4}$, and single-sample training; the one-step baseline uses one initial first-layer update estimated with 1000 samples, and all methods use the same initialization; random features use $\eta_a = 60$ and one-step learning $\eta_a = 1.5$ with a first-layer step $\eta_f/k = \sqrt{n}$.

Decomposed Dynamics (DD) with fixed U_t, where the complement directions stay at initialization while the signal matrix M_t and the complement norms r_t are the trained parameters of the first layer, updated by the components of the SGD step along Ξ and along $u_{i,0}$, respectively (this variant uses the SGD step itself rather than the surrogate update (14); below, DD refers to this variant). While all algorithms use SGD for the second layer, the distinction lies in the training of the first layer. For random features and one-step feature learning, which serve as baselines for feature learning, the first layer is fixed while the second layer is trained. For SGD and DD, the two layers are trained simultaneously, capturing the coupled nature of feature learning.[3]

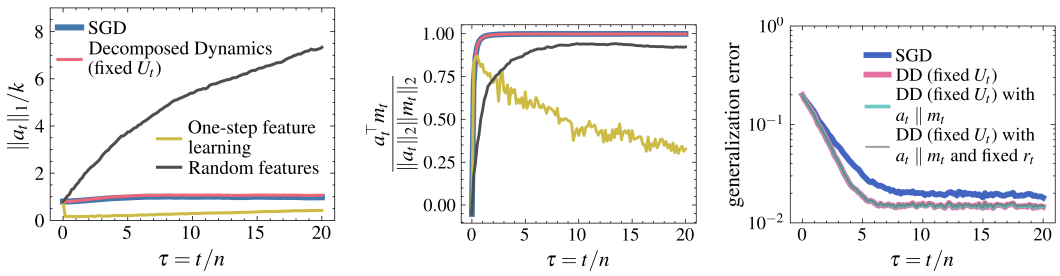
Generalization dynamics — We start by comparing these algorithms in terms of generalization error over time in Figure 3a. SGD ultimately outperforms one-step feature learning, which in turn improves on random features. One-step feature learning leads to a steep decrease at the beginning, while SGD enjoys a slower but steadier decrease over time. In this ReLU setting, DD with fixed complement directions improves further on SGD. Since SGD and DD grow the signal component at a similar rate (Figure 3b), this improvement does not come from faster learning of the teacher direction. Freezing removes the teacher-induced spike channel, but it also stops the noise-driven growth of r_t seen under SGD (Figure 3c), so the comparison does not isolate the spike channel. The result is consistent with the prediction that training the complement directions can inflate variance even when the teacher-aligned component learns useful signal.

First-layer dynamics — Recalling $F_t = M_t\Xi^\top + \mathrm{diag}(r_t)U_t$, we track the signal matrix M_t and the complement norms r_t in Figures 3b–c. Figure 3b shows that the norm of M_t rises during feature learning, indicating increasing alignment between the first-layer weights and the teacher subspace, since M_t captures the aligned part of the first layer. SGD and DD show similar growth of M_t, whereas the norms of both M_t and r_t remain fixed for random features and jump sharply with the giant step of one-step feature learning. In Figure 3c, $\|r_t\|_1/k$ grows under SGD and stays nearly constant under DD, which motivates the fixed-r_t reduction tested in Figure 4c and the unit-variance complement channel of Proposition 3.6.

Second-layer dynamics — In Figure 4a, the norm $\|a_t\|_1/k$ of the second-layer weights changes drastically for random features and one-step feature learning to compensate for the fixed first layer, whereas SGD and DD display similar, steady behavior with a slight increase over time.

Coupling of the layers — Figure 4b shows the coupled dynamics of the two layers through the alignment between the signal matrix M_t of the first layer and the second-layer weights a_t. SGD

[3]Code for all figures is available at
`https://github.com/KU-MLIP/Decomposing-SGD-Dynamics-in-Neural-Networks`.



(a) mean $|a|$ (b) alignment of a and m (c) setting $m_t \parallel a_t$

Figure 4: Second-layer/signal coupling and reduced Decomposed Dynamics in the setting of Figure 3. Panels show (a) mean second-layer magnitude $\|a_t\|_1/k$, (b) cosine alignment $a_t^\top m_t/(\|a_t\|_2\|m_t\|_2)$ for the same four algorithms, and (c) test-error curves for DD reductions that keep U_t fixed and additionally impose $m_t \parallel a_t$, with and without fixed r_t, against SGD and DD with fixed U_t.

and DD have nearly identical alignment evolution over time: the alignment increases rapidly for both, consistent with Proposition 3.6. One-step feature learning aligns after the large update and then gradually loses alignment, with fluctuations, while the second layer adapts to the frozen first layer. Random features also gain alignment over time, even though their first layer is fixed: the second layer upweights the fixed neurons that are already better aligned with the target subspace.

Unnecessary degrees of freedom — For DD, the near-constancy of r_t and the fast alignment of M_t with a_t motivate us to fix r_t at initialization and to set $M_t = a_t\nu_t^\top$ with a new trainable parameter $\nu_t \in \mathbb{R}^{k_\star}$ (updated by projecting the DD step for M_t onto a_t). Figure 4c shows that fixing r_t at initialization and setting the signal matrix parallel to the second layer leaves the test-error dynamics essentially unchanged, while the number of trainable parameters drops from $k(n+1)$ to $k + k_\star$. Thus, especially when $k_\star \ll n$, many raw first-layer degrees of freedom appear unnecessary for this teacher–student task. These variants use the teacher subspace, so they are diagnostic interventions; a teacher-agnostic analogue would train a low-rank update on frozen random features.

Summary — Together, the simulations show that SGD builds spikes in the complement Gram matrix (Figure 1), freezing the complement directions improves the test error in the ReLU setting (Figure 3a), the signal component aligns quickly with the second layer (Figure 4b), and a reduced model with $k + k_\star$ trainable parameters matches the frozen-direction variant (Figure 4c).

5 Conclusion

We decomposed two-layer feature learning into a teacher-aligned signal component and a high-dimensional teacher-orthogonal complement. Even when complement updates are initially isotropic, SGD injects *teacher-dependent structure* that creates *spectral spikes* in the complement Gram matrix, giving a mechanism for emergent anisotropy and variance inflation. Decomposed Dynamics isolates this geometry through its complement update and matches SGD learning curves asymptotically, and the decomposition enables counterfactual experiments showing that constraining or freezing the complement can improve generalization. The same viewpoint reveals rapid cross-layer alignment between the signal component and second-layer weights, suggesting far fewer effective degrees of freedom than the raw parameterization. These results sharpen the role of nominally "irrelevant" directions and motivate interventions such as complement regularization, freezing, or low-rank training. In misspecified settings or under task shift, however, the nominal complement may carry signal, and structure learned there could help.

Acknowledgments and Disclosure of Funding

This work was supported in part by TÜBİTAK under the ARDEB 1001 program (Project No. 124E063). S.D. was also supported by a KUIS AI Center AI Fellowship and a TÜBİTAK BİDEB 2211 PhD Scholarship. The authors declare no competing interests.

References

[1] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

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

[3] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[4] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

[5] Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022.

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

[7] 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, 2019.

[8] Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, 25(349):1–65, 2024.

[9] Hugo Cui, Luca Pesce, Yatin Dandi, Florent Krzakala, Yue Lu, Lenka Zdeborova, and Bruno Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. In *International Conference on Machine Learning*, 2024.

[10] Iain M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2):295–327, 2001.

[11] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, 2005.

[12] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011.

[13] Stephan Mandt, Matthew D. Hoffman, and David M. Blei. Stochastic gradient descent as approximate Bayesian inference. *Journal of Machine Learning Research*, 18(134):1–35, 2017.

[14] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, 2017.

[15] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949–986, 2022.

[16] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[17] David Saad and Sara A. Solla. On-line learning in soft committee machines. *Physical Review E*, 52(4):4225–4243, 1995.

[18] Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 35, pages 23244–23255, 2022.

[19] 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.

[20] Luca Arnaboldi, Ludovic Stephan, Florent Krzakala, and Bruno Loureiro. From high-dimensional & mean-field dynamics to dimensionless ODEs: A unifying approach to SGD in two-layers networks. In *Conference on Learning Theory*, 2023.

[21] 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.

[22] Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in Gaussian mixture classification. In *Advances in Neural Information Processing Systems*, 2020.

[23] Cédric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborová. Rigorous dynamical mean-field theory for stochastic gradient descent methods. *SIAM Journal on Mathematics of Data Science*, 6(2):400–427, 2024.

[24] Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. In *Advances in Neural Information Processing Systems*, 2022.

[25] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, 2022.

[26] Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. In *International Conference on Machine Learning*, 2024.

[27] Alireza Mousavi-Hosseini, Sejun Park, Manuela Girotti, Ioannis Mitliagkas, and Murat A Erdogdu. Neural networks efficiently learn low-dimensional representations with SGD. In *International Conference on Learning Representations*, 2023.

[28] Alireza Mousavi-Hosseini, Denny Wu, and Murat A. Erdogdu. Learning multi-index models with neural networks via mean-field Langevin dynamics. In *International Conference on Learning Representations*, 2025.

[29] Samet Demir and Zafer Dogan. Asymptotic analysis of two-layer neural networks after one gradient step under Gaussian mixtures data with structure. In *International Conference on Learning Representations*, 2025.

[30] Rishi Sonthalia, Michael Murray, and Guido Montufar. Low rank gradients and where to find them. In *Advances in Neural Information Processing Systems*, 2025.

[31] Alain-Sol Sznitman. Topics in propagation of chaos. In Paul-Louis Hennequin, editor, *École d'Été de Probabilités de Saint-Flour XIX — 1989*, volume 1464 of *Lecture Notes in Mathematics*, pages 165–251. Springer, 1991.

[32] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.

[33] Alexandru Damian, Jason D. Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, 2022.

[34] Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. In *Advances in Neural Information Processing Systems*, 2022.

[35] Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. The merged-staircase property: A necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, 2022.

[36] Emmanuel Abbe, Enric Boix-Adserà, and Theodor Misiakiewicz. SGD learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *Conference on Learning Theory*, 2023.

[37] Phan-Minh Nguyen and Huy Tuan Pham. A rigorous framework for the mean field limit of multilayer neural networks. *Mathematical Statistics and Learning*, 6:201–357, 2023.

[38] Colin Wei, Jason D. Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets v.s. their induced kernel. In *Advances in Neural Information Processing Systems*, 2019.

[39] Lénaïc Chizat. Sparse optimization on measures with over-parameterized gradient descent. *Mathematical Programming*, 194:487–532, 2022.

[40] Leticia F. Cugliandolo and Jorge Kurchan. On the out of equilibrium relaxation of the Sherrington–Kirkpatrick model. *Journal of Physics A: Mathematical and General*, 27(17):5749–5772, 1994.

[41] Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, and Lenka Zdeborová. Who is afraid of big bad minima? Analysis of gradient-flow in a spiked matrix-tensor model. In *Advances in Neural Information Processing Systems*, 2019.

[42] 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*, 2022.

[43] Hong Hu and Yue M. Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 69(3):1932–1964, 2023.

[44] Andrea Montanari and Basil Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*. PMLR, 2022.

[45] Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning. In *International Conference on Machine Learning*, 2023.

[46] Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. In *Advances in Neural Information Processing Systems*, 2020.

[47] Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: A spiked random matrix perspective. In *Advances in Neural Information Processing Systems*, 2023.

[48] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

[49] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12:2914, 2021.

[50] Michel Talagrand. *Mean field models for spin glasses: Volume I: Basic examples*, volume 54. Springer Science & Business Media, 2010.

[51] Ivan Nourdin and Giovanni Peccati. *Normal Approximations with Malliavin Calculus: From Stein's Method to Universality*, volume 192 of *Cambridge Tracts in Mathematics*. Cambridge University Press, 2012.

A Related Work

Our work touches several closely related lines in the theory of feature learning and high-dimensional SGD: teacher–student models, mean-field and DMFT analyses of training dynamics, Gaussian universality, spiked random matrix theory, and implicit regularization in overparameterized models. Below we situate our contributions against each, focusing on distinctions rather than a comprehensive survey.

A.1 Teacher–Student Models and Feature Learning

Teacher–student frameworks have played a central role in theoretical studies of representation learning, as they allow generalization and optimization dynamics to be analyzed in a controlled setting. A large body of work studies two-layer networks trained by SGD or related algorithms on targets generated by a low-dimensional teacher, with particular emphasis on how learned features align with teacher directions [7, 8, 9]. Closely related, [33] prove that, for degree-p polynomial targets with a rank-r expected Hessian, two-layer networks learn rank-r multi-index representations with gradient descent in $O(d^2 r + dr^p)$ samples for input dimension d, [34] establish single-index recoverability with shallow networks driven by the information exponent, and [28] show that mean-field Langevin dynamics can learn multi-index models with complexity governed by an effective dimension. The staircase line of work [35, 36] characterizes which sparse and multi-index targets SGD recovers efficiently, identifying *leap complexity* as the governing scale. Together, these works characterize when and how shallow networks learn teacher-relevant structure, including alignment dynamics, phase transitions, time scales, and effective dimension.

We complement these analyses by resolving SGD dynamics along *both* teacher-aligned and teacher-orthogonal directions. Existing work typically concentrates on the aligned components and carries the orthogonal complement through low-dimensional state variables (fixed-width overlap matrices, norms, scalar kernels) or treats it as isotropic noise. [18] extend the order-parameter ODEs of online learning to widths and learning rates that scale as powers of the input dimension; in their normalization our setting (width and learning rate both proportional to n) lies on the line of their phase diagram on which the SGD variance term in the update of the student overlap matrix (to leading order, the covariance injected into the complement) is of the same order as the learning term. Beyond fixed width, they use this term only through its scaling and its effect on the asymptotic risk, not through its structure across neurons. We show that the covariance SGD injects into the complement at every step contains a coherent, teacher-induced part (Lemmas 3.1–3.2), which produces anisotropic, spiked complement covariance (Remark 3.3 and Figure 1), with concrete consequences for generalization (Section 3.3). The fast-alignment phenomenon between m_t and a_t (Proposition 3.6) further sharpens the timescale picture in this line.

A.2 SGD Dynamics Beyond the NTK Regime

A major theme in the theory of neural network training is the distinction between "lazy" training regimes, where features remain essentially fixed, and regimes in which features evolve substantially during optimization. The neural tangent kernel (NTK) framework formalizes the former perspective, relating training dynamics to kernel regression with random features [1, 2]. While this viewpoint has yielded powerful tools for understanding optimization and generalization, it inherently suppresses feature learning effects.

At the opposite extreme, mean-field and particle-based approaches describe feature learning at infinite width using PDE or interacting particle systems [3, 4, 5, 6], with rigorous multilayer extensions in [37]. [38] establish a fundamental separation between mean-field (feature-learning) and NTK (lazy) regimes for two-layer networks, and [39] analyzes sparse measure-valued gradient descent in the over-parameterized limit. These analyses capture nontrivial feature evolution but often abstract away fine-grained covariance structure by averaging over particles or weights. Our equivalence proof (Theorem 3.4) rests on a stability estimate for the reduced state (a_t, M_t, Σ_t) (Lemma F.1), in the spirit of propagation of chaos, while retaining anisotropic complement covariances absent in standard mean-field descriptions.

Dynamical mean-field theory (DMFT), originating in spin-glass physics [40], provides a complementary self-consistent description of SGD in high dimensions. In machine-learning settings, [22]

apply DMFT to SGD on Gaussian-mixture classification, with rigorous foundations in [23]; [19] establish high-dimensional limit theorems with effective dynamics and critical scalings, and [41] apply DMFT to gradient flow on a spiked matrix-tensor landscape. [24] derive a self-consistent dynamical field theory of kernel evolution in wide networks. [20] bridge high-dimensional and mean-field descriptions of two-layer SGD via dimensionless ODEs, showing that infinite-width dynamics remain close to a low-dimensional subspace spanned by target principal directions. Our equivalence proof for Decomposed Dynamics uses stability estimates (Lemma F.1) in the spirit of [23, 19], but specialized to the $(M_t, F_t^\perp)$ split of the first layer.

Our analysis sits between the lazy and mean-field regimes: features evolve nontrivially under SGD, but the decomposition $F_t = M_t \Xi^\top + F_t^\perp$ separates a low-dimensional teacher subspace from a high-dimensional complement whose structure we track. This is enough to derive an analytically tractable surrogate (Decomposed Dynamics, Theorem 3.4) whose test-error dynamics match those of SGD in the proportional limit for bounded $a_{i,0}$. The takeaway is that even when mean-field descriptions are accurate at the level of marginal distributions, higher-order structure in the teacher-orthogonal directions can shape generalization.

A.3 Gaussian Universality and Equivalence

A growing body of work establishes *Gaussian universality* (or *Gaussian equivalence*): in high-dimensional asymptotics, certain non-Gaussian feature distributions can be replaced, without changing learning curves at leading order, by Gaussian surrogates with matched first and second moments. [42] formalize this principle for shallow networks acting on data from generative models, and [43] prove universality laws for random-feature models via a Lindeberg/Stein argument. [44] extend universality to general empirical risk minimization in the proportional regime, and [45] establish deterministic equivalents and error universality for deep random features.

Our results provide a complementary, training-time form of equivalence. The Decomposed Dynamics surrogate (Theorem 3.4) replaces the non-Gaussian rank-one complement increment $-(\eta_f/k) v_t g_t^\top$ by a Gaussian increment with the same conditional mean and, to leading order, the same neuron-space second moment, in line with the universality principle; the bulk-plus-spike replacement $n_\perp G_t \approx \hat{c}_t I_k + B_t^{\mathrm{acc}} (B_t^{\mathrm{acc}})^\top$ in the effective predictor (Eq. (16)) further approximates the covariance of the (exactly Gaussian) complement preactivations. Whereas the works above establish equivalence at fixed (typically untrained) features, or apply it to one-pass SGD of a fixed-width student [42], our equivalence is for the test-loss-relevant reduced state (a_t, M_t, Σ_t) at proportional width along the SGD trajectory, which the Gaussian surrogate reproduces through its conditional mean increments.

A.4 Random Matrix Theory and Spiked Models

The emergence of spectral spikes in covariance or Gram matrices is a classical topic in random matrix theory, with foundational results in spiked covariance and finite-rank deformation models [10, 11, 12]. In these settings, low-rank structure is imposed exogenously through the data-generating process, and spectral outliers encode signal directions. Closer to neural network training, [46] establish the spectra of the Conjugate Kernel and Neural Tangent Kernel in the linear-width regime via iterated Marchenko–Pastur fixed-point equations, providing the bulk reference for training-time spectra. [47] study learning in the presence of low-dimensional input structure via a spiked random matrix perspective, where spikes originate in the data covariance. Closer to our setting, one or a few large gradient steps create spikes in the first-layer weights that align with teacher-relevant (polynomial) directions and improve learning [25, 26, 9, 8].

In contrast, the spikes we identify are *dynamically generated by SGD* (Remark 3.3): the data distribution imposes no anisotropy in the teacher-orthogonal subspace, yet a few spectral outliers emerge from the interaction between SGD updates, the second-layer weights, and the teacher-aligned signal, above a bulk resembling Marchenko–Pastur (Figure 1). This yields a training-time mechanism for spike formation, within reach of random matrix theory, that differs from classical spiked models and from one-step spikes: ours form in the teacher-orthogonal complement under online SGD and carry no target signal.

A.5 Implicit Regularization, Variance Control, and Generalization

A complementary line of work studies SGD through stochastic-process and diffusion approximations, relating training dynamics to stationary distributions and implicit regularization effects [13, 14]. Separately, work on interpolation, ridgeless regression, and double descent highlights the central role of variance control in overparameterized models [15, 16]. Closer to the proportional regime studied here, [48] derive precise asymptotics for the generalization error of random features and characterize the double-descent curve and the role of effective ridge regularization, while [49] decompose generalization through the eigenstructure of the data–feature kernel, providing a spectrum-aware bias–variance picture.

Our results sharpen these perspectives for two-layer feature learning. Proposition 3.5 identifies the teacher-orthogonal complement as a concrete, structural source of predictor variance, arising from dynamically induced anisotropy rather than data covariance. Empirically (Figure 3), freezing the complement directions improves generalization, exposing a much smaller effective number of degrees of freedom than the raw $k \times n$ parameterization and clarifying how such effects coexist with active feature learning in the teacher-aligned subspace.

B Neuron decoupling

This appendix provides a decoupling bound for the *centered* SGD drivers conditional on the teacher signal. We present the single-index case ($k_\star = 1$) to streamline notation.

B.1 Filtration and notation

Fix an SGD iterate t. Let $\mathcal{F}_t$ be the σ-field generated by all randomness up to time t (including initialization and samples $\{x_s\}_{s<t}$), so that (F_t, a_t) is $\mathcal{F}_t$-measurable. At time t, a fresh sample $x_t \sim \mathcal{N}(0, I_n)$ is drawn independently of $\mathcal{F}_t$.

Decompose $x_t = z_{0,t}\xi + g_t$ with $z_{0,t} = \xi^\top x_t \sim \mathcal{N}(0,1)$ and $g_t = P_\perp x_t \sim \mathcal{N}(0, P_\perp)$ independent of $z_{0,t}$. For each neuron i define the $\mathcal{F}_t$-measurable decomposition

$$f_{i,t} = m_{i,t}\xi + r_{i,t}u_{i,t}, \qquad m_{i,t} = \xi^\top f_{i,t}, \qquad r_{i,t} = \|P_\perp f_{i,t}\|_2, \qquad u_{i,t} = \frac{P_\perp f_{i,t}}{\|P_\perp f_{i,t}\|_2} \in \xi^\perp. \tag{22}$$

Then the preactivation is

$$h_{i,t} = f_{i,t}^\top x_t = m_{i,t}z_{0,t} + r_{i,t}z_{i,t}, \qquad z_{i,t} := u_{i,t}^\top g_t. \tag{23}$$

Define the student prediction and error

$$\hat{y}_t = \frac{1}{k}\sum_{\ell=1}^{k} a_{\ell,t}\sigma(h_{\ell,t}), \qquad y_t = g(z_{0,t}), \qquad e_t = \hat{y}_t - y_t,$$

and the per-neuron SGD driver

$$v_{i,t} := e_t\, a_{i,t}\, \sigma'(h_{i,t}). \tag{24}$$

Define the conditional mean and centered driver

$$\mu_{i,t}(z_{0,t}) := \mathbb{E}[v_{i,t} \mid \mathcal{F}_t, z_{0,t}], \qquad \tilde{v}_{i,t} := v_{i,t} - \mu_{i,t}(z_{0,t}). \tag{25}$$

B.2 A quantitative two-neuron decoupling bound (conditional, centered)

We control $\mathrm{Cov}(\tilde{v}_{i,t}, \tilde{v}_{j,t})$ via a Gaussian interpolation argument over the complement Gaussian vector $z_t = (z_{1,t}, \ldots, z_{k,t})$ conditional on $(\mathcal{F}_t, z_{0,t})$.

Lemma B.1 (Conditional decoupling for centered drivers, weighted form)**.** *Assume **A1–A5** and work on the event **A6**. Fix $\tau > 0$, $t \le \tau n$ and $i \neq j$, and condition on $(\mathcal{F}_t, z_{0,t})$. Let $\rho_{pq,t} := u_{p,t}^\top u_{q,t}$, $\Delta_t := \max_{p\neq q} |\rho_{pq,t}| \le 1$, and the empirical coupled-second-moment $H_t := \frac{1}{k}\sum_{p=1}^{k} a_{p,t}^2 r_{p,t}^2$.*

Define the per-pair weights

$$W_{ij,t}^{(\rho)} := |a_{i,t}a_{j,t}|\, r_{i,t}\, r_{j,t},$$
$$W_{ij,t}^{(\Delta)} := |a_{i,t}a_{j,t}|\, \big(1 + r_{i,t} + r_{j,t}\big), \tag{26}$$
$$W_{ij,t}^{(0)} := a_{i,t}^2\, r_{i,t}^2\, |a_{j,t}| + a_{j,t}^2\, r_{j,t}^2\, |a_{i,t}| + |a_{i,t}a_{j,t}|\, H_t.$$

Then

$$\big|\operatorname{Cov}(\tilde{v}_{i,t}, \tilde{v}_{j,t} \mid \mathcal{F}_t, z_{0,t})\big| \;\le\; C\Big[W_{ij,t}^{(\rho)}\, |\rho_{ij,t}| \;+\; W_{ij,t}^{(\Delta)}\, \Delta_t \;+\; \tfrac{1}{k}\, W_{ij,t}^{(0)}\Big], \tag{27}$$

where $C < \infty$ is a constant depending only on $\|\sigma\|_\infty$, $\|\sigma'\|_\infty$, $\|\sigma''\|_\infty$, $\|g\|_\infty$, and B_τ from Lemma I.1. The fourth-moment-type quantity H_t is left explicit: it is bounded by $B'_\tau(n) = O(\log n)$ via Lemma I.2, but this is consumed only by downstream sums, not absorbed here. Consequently, since the weights and $\rho_{ij,t}, \Delta_t$ are $\mathcal{F}_t$-measurable,

$$\big|\mathbb{E}[\tilde{v}_{i,t}\tilde{v}_{j,t} \mid \mathcal{F}_t]\big| \;\le\; C\Big[W_{ij,t}^{(\rho)}\, |\rho_{ij,t}| \;+\; W_{ij,t}^{(\Delta)}\, \Delta_t \;+\; \tfrac{1}{k}\, W_{ij,t}^{(0)}\Big].$$

Proof. Fix $(\mathcal{F}_t, z_{0,t})$. The complement projection vector $z_t = (z_{1,t}, \dots, z_{k,t})$ is centered Gaussian with covariance $\Sigma_t^u = U_t U_t^\top$, i.e., $(\Sigma_t^u)_{pq} = \rho_{pq,t}$ and $(\Sigma_t^u)_{pp} = 1$. Write $\tilde{v}_{i,t} = F_i(z_t) - \mathbb{E}[F_i(z_t) \mid \mathcal{F}_t, z_{0,t}]$ where

$$F_i(z_t) = e_t(z_t)\, a_{i,t}\, \sigma'(m_{i,t}z_{0,t} + r_{i,t}z_{i,t})$$

and $e_t(z_t) = \frac{1}{k}\sum_\ell a_{\ell,t}\sigma(m_{\ell,t}z_{0,t} + r_{\ell,t}z_{\ell,t}) - g(z_{0,t})$. Since subtracting constants does not affect covariance, it suffices to bound $\operatorname{Cov}(F_i(z_t), F_j(z_t))$.

Under **A3**, F_i is Lipschitz in z_t with square-integrable weak gradient. Gaussian interpolation [50, Lemma 1.3.1] for Gaussian vectors implies

$$\operatorname{Cov}(F_i(z_t), F_j(z_t)) = \int_0^1 \mathbb{E}\Big[\nabla F_i(Z)^\top \Sigma_t^u\, \nabla F_j(Z^{(\theta)}) \,\Big|\, \mathcal{F}_t, z_{0,t}\Big]\, d\theta, \tag{28}$$

where $(Z, Z^{(\theta)})$ are jointly Gaussian with marginal $\mathcal{N}(0, \Sigma_t^u)$ and cross-covariance $\mathbb{E}[Z\, Z^{(\theta)\top}] = \theta\, \Sigma_t^u$. Taking absolute values and applying Cauchy–Schwarz yields

$$|\operatorname{Cov}(F_i, F_j)| \;\le\; \sum_{p,q=1}^{k} |\rho_{pq,t}| \Big(\mathbb{E}[(\partial_p F_i(Z))^2 \mid \mathcal{F}_t, z_{0,t}]\Big)^{1/2} \Big(\mathbb{E}[(\partial_q F_j(Z))^2 \mid \mathcal{F}_t, z_{0,t}]\Big)^{1/2}. \tag{29}$$

We bound the weak derivatives using the chain rule. For $p \neq i$,

$$\partial_p F_i(z_t) = (\partial_p e_t(z_t)) \cdot a_{i,t}\sigma'(h_{i,t}) = \frac{1}{k}\, a_{p,t}r_{p,t}\sigma'(h_{p,t}) \cdot a_{i,t}\sigma'(h_{i,t}),$$

hence by bounded σ',

$$\mathbb{E}[(\partial_p F_i)^2 \mid \mathcal{F}_t, z_{0,t}] \;\le\; \frac{C}{k^2}\, a_{i,t}^2 a_{p,t}^2 r_{p,t}^2 \qquad (p \neq i). \tag{30}$$

For $p = i$,

$$\partial_i F_i(z_t) = (\partial_i e_t)\, a_{i,t}\sigma'(h_{i,t}) + e_t \cdot a_{i,t}r_{i,t}\, \sigma''(h_{i,t}), \qquad \partial_i e_t = \frac{1}{k} a_{i,t}r_{i,t}\sigma'(h_{i,t}),$$

so bounded σ', σ'' implies

$$\mathbb{E}[(\partial_i F_i)^2 \mid \mathcal{F}_t, z_{0,t}] \;\le\; C\Big(\frac{a_{i,t}^4 r_{i,t}^2}{k^2} + a_{i,t}^2 r_{i,t}^2\, \mathbb{E}[e_t^2 \mid \mathcal{F}_t, z_{0,t}]\Big). \tag{31}$$

On the high-probability event in **A6** we have $\frac{1}{k}\sum_\ell a_{\ell,t}^2 \le B_\tau$ and $\frac{1}{k}\sum_\ell r_{\ell,t}^2 \le B_\tau$, and since σ, g are bounded,

$$|e_t(z_t)| \;\le\; \|\sigma\|_\infty \Big(\tfrac{1}{k}\textstyle\sum_\ell a_{\ell,t}^2\Big)^{1/2} + \|g\|_\infty \;\le\; \|\sigma\|_\infty \sqrt{B_\tau} + \|g\|_\infty,$$

uniformly in z_t. In particular $\max_\ell a_{\ell,t}^2 \le \sum_\ell a_{\ell,t}^2 \le kB_\tau$, so $a_{i,t}^4/k^2 \le B_\tau\, a_{i,t}^2$, and (31) simplifies to

$$\mathbb{E}[(\partial_i F_i)^2 \mid \mathcal{F}_t, z_{0,t}] \;\le\; C_1\, a_{i,t}^2\, r_{i,t}^2, \qquad \mathbb{E}[(\partial_p F_i)^2 \mid \mathcal{F}_t, z_{0,t}] \;\le\; \frac{C_1}{k^2}\, a_{i,t}^2\, a_{p,t}^2\, r_{p,t}^2 \quad (p \neq i), \tag{32}$$

with C_1 depending only on $\|\sigma\|_\infty, \|\sigma'\|_\infty, \|\sigma''\|_\infty, \|g\|_\infty, B_\tau$, and the analogous bounds for $\partial_q F_j$. Insert (32) into (29) and partition $(p,q) \in \{1,\dots,k\}^2$ into four blocks.

(T1) $(p,q) \in \{i,j\} \times \{i,j\}$. The four cells contribute, using $|\rho_{ii}| = |\rho_{jj}| = 1$ and $|\rho_{ij}| = |\rho_{ji}|$:

- $(p,q) = (i,j)$: $|\rho_{ij,t}| \cdot C_1\,|a_{i,t}a_{j,t}|r_{i,t}r_{j,t} \;=\; C_1\,W^{(\rho)}_{ij,t}\,|\rho_{ij,t}|$.
- $(p,q) = (j,i)$: same with an extra factor $|a_{i,t}a_{j,t}|/k^2$ (both derivatives off-diagonal); subleading and absorbed into $C_1\,W^{(\rho)}_{ij,t}\,|\rho_{ij,t}|$ via $|a_{i,t}a_{j,t}|/k^2 \le B_\tau/k$.
- $(p,q) = (i,i)$: $1 \cdot \frac{C_1}{k}\,a_{i,t}^2 r_{i,t}^2 |a_{j,t}|$, part of $\frac{1}{k}\,W^{(0)}_{ij,t}$.
- $(p,q) = (j,j)$: symmetric, part of $\frac{1}{k}\,W^{(0)}_{ij,t}$.

(T2) $p \in \{i,j\}$, $q \notin \{i,j\}$, *and (T3) symmetric.* For $p = i$, using $|\rho_{iq,t}| \le \Delta_t$ for $q \ne i$ and Cauchy–Schwarz $\sum_q |a_{q,t}|r_{q,t} \le \sqrt{\sum a^2}\sqrt{\sum r^2} \le kB_\tau$,

$$\sum_{q \notin \{i,j\}} |\rho_{iq,t}| \cdot \tfrac{C_1}{k}|a_{i,t}a_{j,t}|r_{i,t}|a_{q,t}|r_{q,t} \;\le\; C_1 B_\tau\,\Delta_t\,|a_{i,t}a_{j,t}|r_{i,t} \;\le\; C_1 B_\tau\,\Delta_t\,W^{(\Delta)}_{ij,t}.$$

The case $p = j$ produces an r_j-analogue carrying an additional factor $|a_{j,t}|/k \le \sqrt{B_\tau/k}$ (both derivatives off-diagonal), absorbed into $C_1 B_\tau\,\Delta_t\,W^{(\Delta)}_{ij,t}$. (T3) is symmetric.

(T4) $p, q \notin \{i,j\}$. Split into diagonal $p = q$ and off-diagonal $p \ne q$:

- Diagonal: $\sum_{p \notin \{i,j\}} 1 \cdot \frac{C_1}{k^2}|a_{i,t}a_{j,t}|a_{p,t}^2 r_{p,t}^2 \le \frac{C_1}{k^2}|a_{i,t}a_{j,t}| \sum_p a_{p,t}^2 r_{p,t}^2 = \frac{C_1}{k}\,|a_{i,t}a_{j,t}|\,H_t$, which is exactly the $|a_i a_j| H_t$ piece of $\frac{1}{k}W^{(0)}_{ij,t}$ (no fourth-moment input needed at this stage).
- Off-diagonal: $\sum_{p \ne q \notin \{i,j\}} |\rho_{pq,t}| \cdot \frac{C_1}{k^2}|a_{i,t}a_{j,t}||a_{p,t}a_{q,t}|r_{p,t}r_{q,t} \le \Delta_t\,\frac{C_1}{k^2}|a_{i,t}a_{j,t}|\big(\sum_p |a_{p,t}|r_{p,t}\big)^2 \le C_1 B_\tau^2\,\Delta_t\,|a_{i,t}a_{j,t}|$, which is $\le C_1 B_\tau^2\,\Delta_t\,W^{(\Delta)}_{ij,t}$.

Summing (T1)–(T4) and absorbing B_τ, B_τ^2 into a single constant C yields (27). ∎

C Proof of Lemma 3.1 (complement geometry)

C.1 Decomposition of the driver covariance accumulator

Expanding $v_s v_s^\top$ gives

$$v_s v_s^\top = \tilde v_s \tilde v_s^\top + \mu_s \mu_s^\top + \mu_s \tilde v_s^\top + \tilde v_s \mu_s^\top,$$

where $\mu_s = \mu_s(z_{0,s})$ is the conditional mean vector and $\tilde v_s = v_s - \mu_s$. Summing and inserting into (8) yields (10) with

$$\widetilde C_t = \frac{1}{n} I_k + \Big(\frac{\eta_f}{k}\Big)^2 \sum_{s=0}^{t-1} \tilde v_s \tilde v_s^\top, \qquad S_s = \mathbb{E}[\mu_s \mu_s^\top \mid \mathcal{F}_s],$$

and

$$R_t = \Big(\frac{\eta_f}{k}\Big)^2 \sum_{s=0}^{t-1} \Big(\mu_s \tilde v_s^\top + \tilde v_s \mu_s^\top\Big) + \Big(\frac{\eta_f}{k}\Big)^2 \sum_{s=0}^{t-1} \Big(\mu_s \mu_s^\top - \mathbb{E}[\mu_s \mu_s^\top \mid \mathcal{F}_s]\Big).$$

Frobenius-level martingale control for R_t. Let $\mathcal{G}_s := \sigma(\mathcal{F}_s, z_{0,s})$ denote the σ-field obtained by enlarging $\mathcal{F}_s$ with $z_{0,s}$. By definition, μ_s is $\mathcal{G}_s$-measurable, and $\mathbb{E}[\tilde v_s \mid \mathcal{G}_s] = 0$, hence $\mu_s \tilde v_s^\top + \tilde v_s \mu_s^\top$ is a martingale difference for the filtration $(\mathcal{G}_{s+1})_{s \ge 0}$. Separately, $\mu_s \mu_s^\top - \mathbb{E}[\mu_s \mu_s^\top \mid \mathcal{F}_s]$ is centered under the original $\mathcal{F}_s$ (with $S_s := \mathbb{E}[\mu_s \mu_s^\top \mid \mathcal{F}_s]$), and is a martingale difference for $(\mathcal{F}_{s+1})_{s \ge 0}$ since $\mu_s \mu_s^\top \in \mathcal{F}_{s+1}$.

The martingale bounds in this appendix are applied to the increments stopped at the $(\mathcal{F}_s)$-stopping time $T := \inf\{s : \|a_s\|_2^2 > kB_\tau\} \wedge \lfloor \tau n \rfloor$, for which they hold whenever $s < T$. The stopped processes $R_{t \wedge T}$ and $N_{t \wedge T}$ (below) are martingales with the stated second moments, and $T = \lfloor \tau n \rfloor$ on **A6**, so they coincide with R_t and N_t there. For $s < T$ (so $\|a_s\|_2^2 \le kB_\tau$) and bounded σ, σ', we have

$$\|\mu_s\|_2^2 \le \mathbb{E}[\|v_s\|_2^2 \mid \mathcal{G}_s] \lesssim \|a_s\|_2^2 = O(k), \qquad \mathbb{E}[\|\tilde v_s\|_2^2 \mid \mathcal{G}_s] \lesssim \|a_s\|_2^2 = O(k),$$

and therefore

$$\mathbb{E}\big[\|\mu_s\tilde{v}_s^\top + \tilde{v}_s\mu_s^\top\|_F^2 \mid \mathcal{G}_s\big] \;\le\; C\,\|\mu_s\|_2^2\,\mathbb{E}[\|\tilde{v}_s\|_2^2 \mid \mathcal{G}_s] \;=\; O(k^2),$$

and similarly $\mathbb{E}[\|\mu_s\mu_s^\top - \mathbb{E}[\mu_s\mu_s^\top \mid \mathcal{F}_s]\|_F^2 \mid \mathcal{F}_s] = O(k^2)$. Hence, by orthogonality of martingale differences in $L^2(\|\cdot\|_F)$,

$$\mathbb{E}\big[\|R_t\|_F^2\big] \;\lesssim\; \Big(\frac{\eta_f}{k}\Big)^4 \sum_{s=0}^{t-1} O(k^2) \;=\; O\Big(\frac{t}{k^2}\Big) \;=\; O\Big(\frac{1}{n}\Big),$$

since $t \le \tau n$ and $k \asymp n$. By Doob's maximal inequality, $\mathbb{E}[\sup_{t\le\tau n}\|R_t\|_F^2] \le 4\,\mathbb{E}[\|R_{\lfloor\tau n\rfloor}\|_F^2] = O(1/n)$, so $\sup_{t\le\tau n}\|R_t\|_F \le \varepsilon_n$ with high probability for some deterministic $\varepsilon_n \to 0$. This gives the remainder bound in (10).

C.2 Residual isotropy

By Lemma B.1, on the event **A6**, and using the trivial bounds $|\rho_{ij,t}|, \Delta_t \le 1$,

$$\big|\mathbb{E}[\tilde{v}_{i,t}\tilde{v}_{j,t} \mid \mathcal{F}_t]\big| \;\le\; C\,\Omega_{ij,t}, \qquad \Omega_{ij,t} \;:=\; W^{(\rho)}_{ij,t} + W^{(\Delta)}_{ij,t} + \tfrac{1}{k}\,W^{(0)}_{ij,t}, \qquad i \ne j, \tag{33}$$

with the weights from (26) and C the constant of Lemma B.1. By the explicit form of the weights, A6, and Lemma I.2 (which gives $H_t \le B'_\tau(n) = O(\log n)$ and $\sum_p a_p^2 r_p^2 \le k B'_\tau(n)$), the per-pair envelope satisfies

$$\sum_{i\ne j} \Omega_{ij,t}^2 \;\le\; k^2\,\mathrm{polylog}(n), \tag{34}$$

the dominant contribution being $\sum_{i,j}(W^{(\rho)}_{ij,t})^2 \;=\; \big(\sum_p a_{p,t}^2 r_{p,t}^2\big)^2 \;\le\; (kB'_\tau(n))^2$; the $W^{(\Delta)}$ and $\frac{1}{k}W^{(0)}$ contributions are bounded by $O(k^2\log n)$ and $O(k\log^2 n)$, respectively.

Write

$$\widetilde{C}_t = \frac{1}{n}I_k + \Big(\frac{\eta_f}{k}\Big)^2 \sum_{s<t} \tilde{v}_s\tilde{v}_s^\top, \qquad c_t = \frac{1}{k}\,\mathrm{Tr}(\widetilde{C}_t),$$

and decompose into drift plus martingale, $\widetilde{C}_t = D_t + N_t$, with $D_t := \frac{1}{n}I_k + \Big(\frac{\eta_f}{k}\Big)^2 \sum_{s<t}\mathbb{E}[\tilde{v}_s\tilde{v}_s^\top \mid \mathcal{F}_s]$ and N_t the centered remainder. Then $\|\widetilde{C}_t - c_t I_k\|_F \le \|D_t - \frac{1}{k}\,\mathrm{Tr}(D_t)I_k\|_F + 2\|N_t\|_F$.

Drift, off-diagonal control. For $i \ne j$, by (33) and Cauchy–Schwarz on the time sum,

$$|(D_t)_{ij}|^2 \;\le\; \frac{C^2\eta_f^4\,t}{k^4}\sum_{s<t}\Omega_{ij,s}^2.$$

Summing over $i \ne j$, applying (34) at each $s < t$, and using $t \le \tau n \lesssim k$,

$$\sum_{i\ne j}|(D_t)_{ij}|^2 \;\le\; \frac{C^2\eta_f^4\,t^2}{k^4}\cdot k^2\,\mathrm{polylog}(n) \;=\; O(\mathrm{polylog}(n)), \qquad \text{hence} \qquad \frac{1}{k^2}\sum_{i\ne j}(D_t)_{ij}^2 \;=\; o(1).$$

Combined with the diagonal term,

$$\frac{1}{k^2}\Big\|D_t - \tfrac{1}{k}\,\mathrm{Tr}(D_t)I_k\Big\|_F^2 \le \frac{1}{k^2}\sum_{i\ne j}(D_t)_{ij}^2 + \frac{1}{k^2}\sum_i\Big((D_t)_{ii} - \tfrac{1}{k}\,\mathrm{Tr}(D_t)\Big)^2 = o(1),$$

where the diagonal term is $o(1)$ as follows. Since $D_t = \frac{1}{n}I_k + (\eta_f/k)^2\sum_{s<t}\mathbb{E}[\tilde{v}_s\tilde{v}_s^\top \mid \mathcal{F}_s]$ is a sum of PSD matrices, $D_t \succeq 0$, and taking trace gives

$$\mathrm{Tr}(D_t) \;=\; \frac{k}{n} + \Big(\frac{\eta_f}{k}\Big)^2\sum_{s<t}\mathbb{E}\big[\|\tilde{v}_s\|_2^2 \,\big|\, \mathcal{F}_s\big].$$

On **A6**, $\mathbb{E}[\|\tilde{v}_s\|_2^2 \mid \mathcal{F}_s] \le C\sum_p a_{p,s}^2 = O(k)$ uniformly, so $\mathrm{Tr}(D_t) \le k/n + (\eta_f/k)^2\cdot t\cdot O(k) = O(1)$ since $t \le \tau n \asymp k$. Then $\sum_i((D_t)_{ii} - \frac{1}{k}\,\mathrm{Tr}(D_t))^2 \le \sum_i (D_t)_{ii}^2 \le \|D_t\|_F^2 \le \mathrm{Tr}(D_t)^2 = O(1)$ by PSD-ness, giving $\frac{1}{k^2}\sum_i((D_t)_{ii} - \frac{1}{k}\,\mathrm{Tr}(D_t))^2 = o(1)$.

Martingale. Since $\|\tilde{v}_s\tilde{v}_s^\top\|_F = \|\tilde{v}_s\|_2^2$ and $|\,\mathrm{Tr}(N_t)| \le \sqrt{k}\|N_t\|_F$, it suffices to control $\|N_t\|_F$. Using orthogonality of martingale differences in $L^2(\|\cdot\|_F)$,

$$\mathbb{E}[\|N_t\|_F^2] \;\le\; \Big(\frac{\eta_f}{k}\Big)^4\sum_{s=0}^{t-1}\mathbb{E}\big[\|\tilde{v}_s\tilde{v}_s^\top\|_F^2\big] = \Big(\frac{\eta_f}{k}\Big)^4\sum_{s=0}^{t-1}\mathbb{E}[\|\tilde{v}_s\|_2^4].$$

For $s < T$ and bounded σ, σ', we have $\|\tilde{v}_s\|_2 \le \|v_s\|_2 + \|\mu_s\|_2 \lesssim \|a_s\|_2$ and $\|a_s\|_2^2 \le kB_\tau$, hence $\|\tilde{v}_s\|_2^4 = O(k^2)$ and

$$\mathbb{E}[\|N_t\|_F^2] \;=\; O\Big(\frac{t}{k^2}\Big) = O\Big(\frac{1}{n}\Big).$$

By Doob's maximal inequality, $\mathbb{E}[\sup_{t\le\tau n}\|N_t\|_F^2] = O(1/n)$, hence $\sup_{t\le\tau n}\|N_t\|_F = o(1)$ with high probability, and therefore

$$\sup_{t\le\tau n}\frac{1}{k^2}\|N_t - \tfrac{1}{k}\operatorname{Tr}(N_t)I_k\|_F^2 \;\le\; \sup_{t\le\tau n}\frac{4}{k^2}\|N_t\|_F^2 \;=\; o(1).$$

Combining the drift and martingale bounds yields (9) on the high-probability events specified by Lemmas I.1–I.2, which jointly hold with probability $1 - o(1)$. The drift contribution is $o(1)$ in normalized Frobenius norm via the weighted form of Lemma B.1 together with the bound $H_t \le B'_\tau(n)$ from Lemma I.2, and the martingale contribution is $o(1)$ by Doob's inequality applied to the second-moment bound on $\|\tilde{v}_s\|_2^2$. ■

D Proof of Lemma 3.2 (normalized complement Gram recursion)

We prove Lemma 3.2. For clarity we write the argument in the single-index case ($k_\star = 1$); the multi-index case follows by conditioning on $z_{\star,t} = \Xi^\top x_t$ throughout.

D.1 Exact one-step identity

Recall $F_t^\perp := F_t P_\perp \in \mathbb{R}^{k\times n}$ and the complement update

$$F_{t+1}^\perp = F_t^\perp - \frac{\eta_f}{k} v_t g_t^\top, \qquad g_t := P_\perp x_t \sim \mathcal{N}(0, P_\perp), \tag{35}$$

where x_t is fresh and independent of $\mathcal{F}_t$, hence g_t is independent of $\mathcal{F}_t$. Let $n_\perp := \operatorname{Tr}(P_\perp)$ and define

$$G_t := \frac{1}{n_\perp} F_t^\perp (F_t^\perp)^\top \in \mathbb{R}^{k\times k}, \qquad h_t^\perp := F_t^\perp g_t \in \mathbb{R}^k.$$

Expanding gives the exact identity

$$G_{t+1} = G_t - \frac{\eta_f}{k\, n_\perp}\Big(v_t h_t^{\perp\top} + h_t^\perp v_t^\top\Big) + \Big(\frac{\eta_f}{k}\Big)^2 \frac{\|g_t\|_2^2}{n_\perp}\, v_t v_t^\top. \tag{36}$$

Taking $\mathbb{E}_t[\cdot]$ yields

$$\mathbb{E}_t[G_{t+1}] = G_t - \frac{\eta_f}{k\, n_\perp}\Big(\mathbb{E}_t[v_t h_t^{\perp\top}] + \mathbb{E}_t[h_t^\perp v_t^\top]\Big) + \Big(\frac{\eta_f}{k}\Big)^2 \mathbb{E}_t\left[\frac{\|g_t\|_2^2}{n_\perp}\, v_t v_t^\top\right]. \tag{37}$$

D.2 Quadratic term: remove $\|g_t\|_2^2/n_\perp$

Let $\Delta_t := \|g_t\|_2^2/n_\perp - 1$. Since $g_t \sim \mathcal{N}(0, P_\perp)$,

$$\mathbb{E}_t[\Delta_t] = 0, \qquad \mathbb{E}_t[\Delta_t^2] = \operatorname{Var}(\|g_t\|_2^2/n_\perp) = \frac{2}{n_\perp}.$$

Define

$$E_t^{(g)} := \Big(\frac{\eta_f}{k}\Big)^2 \mathbb{E}_t\big[\Delta_t\, v_t v_t^\top\big].$$

By conditional Cauchy–Schwarz,

$$\|\mathbb{E}_t[\Delta_t v_t v_t^\top]\|_F \le \sqrt{\mathbb{E}_t[\Delta_t^2]}\cdot\sqrt{\mathbb{E}_t[\|v_t v_t^\top\|_F^2]} = \sqrt{\frac{2}{n_\perp}}\cdot\sqrt{\mathbb{E}_t[\|v_t\|_2^4]}.$$

Under **A2–A3**, $|\sigma|$ and $|\sigma'|$ are bounded and g is bounded, hence $|e_t| \lesssim 1 + \|a_t\|_2/\sqrt{k}$. On the **A6** event, $\sum_i a_{i,t}^2 = O(k)$ uniformly over $t \le \tau n$, so $|e_t| = O(1)$, $\|v_t\|_2^2 \lesssim \sum_{i=1}^k a_{i,t}^2 = O(k)$ and $\mathbb{E}_t[\|v_t\|_2^4] = O(k^2)$. Thus

$$\|\mathbb{E}_t[\Delta_t v_t v_t^\top]\|_F = O\Big(\frac{k}{\sqrt{n_\perp}}\Big), \qquad \Rightarrow \qquad \|E_t^{(g)}\|_F = O\Big(\frac{1}{k\sqrt{n_\perp}}\Big) = O\Big(\frac{1}{n^{3/2}}\Big),$$

since $k \asymp n_\perp \asymp n$. Therefore,

$$\Big(\frac{\eta_f}{k}\Big)^2 \mathbb{E}_t\Big[\frac{\|g_t\|_2^2}{n_\perp} v_t v_t^\top\Big] = \Big(\frac{\eta_f}{k}\Big)^2 K_t + E_t^{(g)}, \qquad K_t := \mathbb{E}_t[v_t v_t^\top] = S_t + \mathbb{E}_t[\tilde{v}_t \tilde{v}_t^\top]. \tag{38}$$

The cross terms in K_t vanish because μ_t is a function of $z_{0,t}$ and $\mathbb{E}_t[\tilde{v}_t \mid z_{0,t}] = 0$.

D.3 Linear term: exact Stein identity

Condition on $(\mathcal{F}_t, z_{0,t})$. Since the rows of $F_t^\perp$ lie in $\mathrm{range}(P_\perp)$, $h_t^\perp = F_t^\perp g_t \sim \mathcal{N}(0, \Sigma_t)$ with $\Sigma_t := F_t^\perp (F_t^\perp)^\top = n_\perp G_t$, and Gaussian integration by parts in $g_t \sim \mathcal{N}(0, P_\perp)$ [51], using $\nabla_g v_{i,t} = F_t^{\perp\top} \partial v_{i,t} / \partial h_t^\perp$ (a.e., by **A3**; valid also when Σ_t is singular, e.g., $k > n_\perp$), gives, for all i, j,

$$\mathbb{E}\big[v_{i,t} h_{j,t}^\perp \,\big|\, \mathcal{F}_t, z_{0,t}\big] = \sum_{\ell=1}^{k} (\Sigma_t)_{j\ell}\, \mathbb{E}\Big[\frac{\partial v_{i,t}}{\partial h_{\ell,t}^\perp} \,\Big|\, \mathcal{F}_t, z_{0,t}\Big],$$

$$\frac{\partial v_{i,t}}{\partial h_{\ell,t}^\perp} = \frac{1}{k}\, a_{i,t} a_{\ell,t} \sigma'(h_{i,t}) \sigma'(h_{\ell,t}) + \delta_{i\ell}\, e_t a_{i,t} \sigma''(h_{i,t}).$$

Since Σ_t is $\mathcal{F}_t$-measurable, averaging over $z_{0,t}$ gives $\mathbb{E}_t[v_t h_t^{\perp\top}] = J_t \Sigma_t = n_\perp J_t G_t$ exactly, and, by symmetry of J_t and Σ_t, $\mathbb{E}_t[h_t^\perp v_t^\top] = n_\perp G_t J_t$. Hence

$$\frac{\eta_f}{k\, n_\perp}\Big(\mathbb{E}_t[v_t h_t^{\perp\top}] + \mathbb{E}_t[h_t^\perp v_t^\top]\Big) = \frac{\eta_f}{k}\big(J_t G_t + G_t J_t\big). \tag{39}$$

D.4 Assemble terms and conclude

Combining (37), (38) and (39) gives (12) with $E_t := E_t^{(g)}$, so that $\|E_t\|_F = O(n^{-3/2})$ uniformly over $t \le \tau n$ (with probability $1 - o(1)$), which is (13). ■

E Teacher-induced covariance is effectively finite rank

This appendix proves that the *teacher-induced* neuron-space covariance

$$S_t \;=\; \mathbb{E}_{z_\star}\big[\mu_t(z_\star)\mu_t(z_\star)^\top \,\big|\, \mathcal{F}_t\big], \qquad \mu_{i,t}(z_\star) := \mathbb{E}[v_{i,t} \mid \mathcal{F}_t, z_\star],$$

is well-approximated (in normalized trace norm, hence in normalized Frobenius norm) by a matrix of rank at most $N_D = \binom{k_\star + D}{D}$, which grows only polylogarithmically in n for a truncation degree such as $D = \lceil (\log n)^2 \rceil$, for which $\log n/(D+1) \to 0$, and that the *accumulated* structure is controlled uniformly over time.

Throughout, $z_\star \sim \mathcal{N}(0, I_{k_\star})$ with $k_\star = O(1)$ fixed.

E.1 Hermite basis and notation

Let $\gamma_{k_\star}$ denote the standard Gaussian measure on $\mathbb{R}^{k_\star}$. Let $\{H_\alpha\}_{\alpha \in \mathbb{N}^{k_\star}}$ be the multivariate orthonormal Hermite basis for $L^2(\gamma_{k_\star})$ [51], indexed by multi-indices $\alpha = (\alpha_1, \dots, \alpha_{k_\star})$ with total degree $|\alpha| = \sum_j \alpha_j$, and satisfying $\mathbb{E}[H_\alpha(z_\star) H_\beta(z_\star)] = \mathbf{1}\{\alpha = \beta\}$.

For each i, expand (in $L^2(\gamma_{k_\star})$)

$$\mu_{i,t}(z_\star) = \sum_{\alpha \in \mathbb{N}^{k_\star}} \beta_{i,t}^{(\alpha)}\, H_\alpha(z_\star), \qquad \beta_{i,t}^{(\alpha)} := \mathbb{E}_{z_\star}[\mu_{i,t}(z_\star) H_\alpha(z_\star) \mid \mathcal{F}_t]. \tag{40}$$

Define coefficient vectors $b_t^{(\alpha)} \in \mathbb{R}^k$ by stacking coefficients across neurons:

$$b_t^{(\alpha)} := \big(\beta_{1,t}^{(\alpha)}, \dots, \beta_{k,t}^{(\alpha)}\big)^\top, \qquad \alpha \in \mathbb{N}^{k_\star}. \tag{41}$$

E.2 Exact PSD decomposition of S_t

Lemma E.1 (Hermite coefficient decomposition)**.** *Conditional on $\mathcal{F}_t$,*

$$S_t \;=\; \sum_{\alpha \in \mathbb{N}^{k_\star}} b_t^{(\alpha)} \big(b_t^{(\alpha)}\big)^\top, \tag{42}$$

where the series converges in Frobenius norm.

Proof. Insert (40) into $\mu_t(z_\star)\mu_t(z_\star)^\top$ and take $\mathbb{E}_{z_\star}[\cdot \mid \mathcal{F}_t]$. Orthonormality removes cross-terms, yielding (42). Convergence follows from Parseval: $\mathbb{E}[\mu_{i,t}(z_\star)^2 \mid \mathcal{F}_t] = \sum_\alpha (\beta_{i,t}^{(\alpha)})^2 < \infty$ under bounded moments. ■

E.3 Spectral tail control from Lipschitz regularity

Lemma E.2 (Hermite tail bound via Gaussian Sobolev energy)**.** *Let $f \in W^{1,2}(\gamma_{k_\star})$ with Hermite coefficients $\{\hat{f}(\alpha)\}$ in the basis $\{H_\alpha\}$. Then for any integer $D \geq 0$,*

$$\sum_{|\alpha|>D} \hat{f}(\alpha)^2 \;\leq\; \frac{1}{D+1}\,\mathbb{E}\big[\|\nabla f(z_\star)\|_2^2\big], \qquad z_\star \sim \mathcal{N}(0, I_{k_\star}), \tag{43}$$

where ∇f is the weak gradient (defined a.e.).

Proof. Let $f = \sum_{m\geq 0} f^{(m)}$ be the orthogonal decomposition into homogeneous Wiener chaoses (total degree m). Then $\|f^{(m)}\|_2^2 = \sum_{|\alpha|=m} \hat{f}(\alpha)^2$. A classical identity for the OU generator [51] gives

$$\mathbb{E}[\|\nabla f\|_2^2] \;=\; \sum_{m\geq 1} m\,\|f^{(m)}\|_2^2.$$

Therefore,

$$\sum_{m>D} \|f^{(m)}\|_2^2 \;\leq\; \frac{1}{D+1}\sum_{m>D} m\,\|f^{(m)}\|_2^2 \;\leq\; \frac{1}{D+1}\,\mathbb{E}[\|\nabla f\|_2^2],$$

which is (43) after re-expanding by multi-indices. ■

E.4 Uniform Sobolev bound for $\mu_{i,t}(z_\star)$

Lemma E.3 (Averaged Gaussian Sobolev bound for conditional mean drivers)**.** *Assume **A1–A5**. Fix $\tau > 0$ and $t \leq \tau n$. There exists a sequence $G(n)$, with $G^2(n) = O(\log n)$, depending only on L_g, $\|g\|_\infty$, the bounds $\|\sigma\|_\infty, \|\sigma'\|_\infty, \|\sigma''\|_\infty$, the constant B_τ from Lemma I.1, and the polylog rate $B'_\tau(n) = O(\log n)$ from Lemma I.2, such that with probability $1 - o(1)$,*

$$\frac{1}{k}\sum_{i=1}^{k} \mathbb{E}_{z_\star}\big[\|\nabla_{z_\star}\mu_{i,t}(z_\star)\|_2^2 \,\big|\, \mathcal{F}_t\big] \;\leq\; G^2(n), \quad \textit{i.e.,} \quad \sum_{i=1}^{k} \mathbb{E}_{z_\star}\big[\|\nabla_{z_\star}\mu_{i,t}\|_2^2 \,\big|\, \mathcal{F}_t\big] \leq k\,G^2(n). \tag{44}$$

Proof. Fix t and condition on $\mathcal{F}_t$. Recall

$$v_{i,t} = e_t\,a_{i,t}\,\sigma'(h_{i,t}), \qquad e_t = \frac{1}{k}\sum_{\ell=1}^{k} a_{\ell,t}\sigma(h_{\ell,t}) - g(z_\star), \qquad h_{\ell,t} = m_{\ell,t}^\top z_\star + r_{\ell,t} z_{\ell,t},$$

with $z_{\ell,t}$ complement Gaussians independent of $z_\star$ given $\mathcal{F}_t$. By **A2** and Rademacher's theorem, g has a weak gradient with $\|\nabla g\|_2 \leq L_g$ a.e. Since $\sigma \in W^{2,\infty}$, the map $z_\star \mapsto e_t$ is a.e. differentiable with

$$\nabla_{z_\star} e_t \;=\; \frac{1}{k}\sum_{\ell=1}^{k} a_{\ell,t}\sigma'(h_{\ell,t})\,m_{\ell,t} \;-\; \nabla g(z_\star),$$

and $\nabla_{z_\star}\sigma'(h_{i,t}) = \sigma''(h_{i,t})\,m_{i,t}$ a.e.

Commuting derivative and conditional expectation (justified by dominated convergence using bounded $\sigma, \sigma', \sigma''$),

$$\nabla_{z_\star}\mu_{i,t}(z_\star) \;=\; \mathbb{E}\Big[\,a_{i,t}\,\sigma'(h_{i,t})\,\nabla_{z_\star} e_t + e_t\,a_{i,t}\,\sigma''(h_{i,t})\,m_{i,t}\;\Big|\;\mathcal{F}_t, z_\star\Big].$$

Applying Jensen and Cauchy–Schwarz to bound $\|\nabla\mu_{i,t}\|_2^2$,

$$\|\nabla_{z_\star}\mu_{i,t}(z_\star)\|_2^2 \;\leq\; 2\|\sigma'\|_\infty^2\,a_{i,t}^2\,\mathbb{E}[\|\nabla e_t\|_2^2 \mid \mathcal{F}_t, z_\star] + 2\|\sigma''\|_\infty^2\,a_{i,t}^2\,\|m_{i,t}\|_2^2\,\mathbb{E}[e_t^2 \mid \mathcal{F}_t, z_\star]. \tag{45}$$

On the high-probability event of Lemma I.1, $|e_t| \leq C$ uniformly, and by Cauchy–Schwarz on the formula for ∇e_t, $\|\nabla e_t\|_2^2 \leq 2\|\sigma'\|_\infty^2 \frac{1}{k^2}(\sum_\ell a_\ell^2)(\sum_\ell \|m_\ell\|^2) + 2L_g^2 \leq 2\|\sigma'\|_\infty^2 B_\tau^2 + 2L_g^2$.

Averaging (45) over $z_\star$ and summing over i,

$$\sum_{i=1}^{k} \mathbb{E}_{z_\star}\big[\|\nabla \mu_{i,t}\|_2^2 \mid \mathcal{F}_t\big] \;\le\; C\sum_{i=1}^{k} a_{i,t}^2 + C\sum_{i=1}^{k} a_{i,t}^2 \|m_{i,t}\|_2^2 \;\le\; CkB_\tau + CkB'_\tau(n),$$

where the first bound uses Lemma I.1 and the second uses (89) from Lemma I.2. Setting $G^2(n) := C(B_\tau + B'_\tau(n)) = O(\log n)$ yields (44). ■

E.5 Finite-rank approximation of S_t

Theorem E.4 (Teacher-induced covariance is approximately finite-rank (normalized trace))**.** *Assume **A1–A5** and work on the event **A6**; fix $\tau > 0$, $t \le \tau n$. For any integer $D \ge 0$, define the truncated PSD matrix*

$$S_t^{(\le D)} \;:=\; \sum_{|\alpha| \le D} b_t^{(\alpha)} \big(b_t^{(\alpha)}\big)^\top, \tag{46}$$

where $b_t^{(\alpha)}$ are from (41)*. Then $S_t^{(\le D)}$ has rank at most*

$$N_D \;:=\; \#\{\alpha \in \mathbb{N}^{k_\star} : |\alpha| \le D\} \;=\; \binom{k_\star + D}{D}, \tag{47}$$

and with probability $1 - o(1)$,

$$\frac{1}{k}\operatorname{tr}\big(S_t - S_t^{(\le D)}\big) \;\le\; \frac{G^2(n)}{D+1}, \qquad G^2(n) = O(\log n), \tag{48}$$

and hence the same bound for $\frac{1}{k}\|S_t - S_t^{(\le D)}\|_F$, where $G^2(n)$ is the Sobolev constant from Lemma E.3 (which depends on B_τ and on the coupled bound $B'_\tau(n) = O(\log n)$ from Lemma I.2).

Proof. By Lemma E.1, $S_t - S_t^{(\le D)} = \sum_{|\alpha|>D} b^{(\alpha)} b^{(\alpha)\top}$ is PSD, hence $\|S_t - S_t^{(\le D)}\|_F \le \operatorname{tr}(S_t - S_t^{(\le D)})$ (by PSD-ness of $S_t - S_t^{(\le D)}$, which is a sum of rank-1 PSD terms). Moreover,

$$\operatorname{tr}(S_t - S_t^{(\le D)}) = \sum_{i=1}^{k} \sum_{|\alpha|>D} (\beta_{i,t}^{(\alpha)})^2.$$

Apply Lemma E.2 to each $\mu_{i,t}$ conditional on $\mathcal{F}_t$ and sum over i:

$$\operatorname{tr}(S_t - S_t^{(\le D)}) \le \frac{1}{D+1}\sum_{i=1}^{k} \mathbb{E}_{z_\star}\big[\|\nabla \mu_{i,t}(z_\star)\|_2^2 \,\big|\, \mathcal{F}_t\big] \le \frac{kG^2(n)}{D+1}$$

with high probability by Lemma E.3. Divide by k to obtain (48). ■

E.6 A time-uniform truncation and control of the accumulated covariance

Fix a macroscopic horizon $T_{\max} = \lfloor \tau n \rfloor$. For each time s, recall $S_s = \mathbb{E}_{z_\star}[\mu_s(z_\star)\mu_s(z_\star)^\top \mid \mathcal{F}_s]$ and let $S_s^{(\le D)}$ be (46).

Lemma E.5 (Uniform-in-time Hermite tail bound)**.** *Assume **A1–A5** and work on the event **A6**; fix $\tau > 0$. With probability $1 - o(1)$, for all $s \le T_{\max}$ and all integers $D \ge 0$,*

$$\frac{1}{k}\operatorname{tr}\big(S_s - S_s^{(\le D)}\big) \;\le\; \frac{G^2(n)}{D+1}, \qquad G^2(n) = O(\log n), \tag{49}$$

where $G^2(n)$ is the (slowly-growing) Sobolev constant from Lemma E.3. Consequently,

$$\frac{1}{k}\big\|S_s - S_s^{(\le D)}\big\|_F \;\le\; \frac{G^2(n)}{D+1}. \tag{50}$$

Proof. Apply Lemma E.2 to each $\mu_{i,s}$ conditional on $\mathcal{F}_s$ and sum over i. Although Lemma E.3 is stated at a fixed time, its proof's bound on $\sum_i \mathbb{E}_{z_\star}[\|\nabla \mu_{i,s}\|_2^2 \mid \mathcal{F}_s]$ depends only on the quantities bounded in **A6** and on $\sum_i a_{i,s}^2 \|m_{i,s}\|_2^2$, all of which are controlled uniformly in $s \le T_{\max}$ on a single $1 - o(1)$ event by Lemmas I.1–I.2 (which themselves give $\sup_{s \le \tau n}$ statements). Hence the same Sobolev constant $G^2(n)$ applies to every $s \le T_{\max}$ on this joint event. Finally, use PSD-ness to pass from trace to Frobenius norm. ■

Theorem E.6 (Accumulated teacher-induced covariance is uniformly approximately finite-mode)**.** *Assume **A1–A5** and work on the event **A6**; fix $\tau > 0$. Define the accumulated teacher-induced covariance*

$$\Sigma_t^{\mathrm{teach}} := \sum_{s=0}^{t-1} S_s, \qquad \Sigma_t^{\mathrm{teach},(\le D)} := \sum_{s=0}^{t-1} S_s^{(\le D)}.$$

(We use the superscript teach *to disambiguate from the complement covariance $\Sigma_t = n_\perp G_t$ used in Appendix F.) Then with probability $1 - o(1)$, uniformly for all $1 \le t \le T_{\max}$ and all $D \ge 0$,*

$$\frac{1}{kt}\,\mathrm{tr}\big(\Sigma_t^{\mathrm{teach}} - \Sigma_t^{\mathrm{teach},(\le D)}\big) \;\le\; \frac{G^2(n)}{D+1}, \qquad G^2(n) = O(\log n), \tag{51}$$

and hence the same bound for $\frac{1}{kt}\|\Sigma_t^{\mathrm{teach}} - \Sigma_t^{\mathrm{teach},(\le D)}\|_F$. Moreover, the truncation uses a time-invariant teacher feature map of dimension $R = N_D = \binom{k_\star + D}{D}$.

Proof. By triangle inequality and Lemma E.5,

$$\big\|\Sigma_t^{\mathrm{teach}} - \Sigma_t^{\mathrm{teach},(\le D)}\big\|_F \;\le\; \sum_{s=0}^{t-1} \big\|S_s - S_s^{(\le D)}\big\|_F \;\le\; \sum_{s=0}^{t-1} \mathrm{tr}(S_s - S_s^{(\le D)}) \;\le\; t\,k\,\frac{G^2(n)}{D+1},$$

and since the second sum equals $\mathrm{tr}(\Sigma_t^{\mathrm{teach}} - \Sigma_t^{\mathrm{teach},(\le D)})$, division by kt yields (51). The "time-invariant feature map" claim follows from the fact that the truncation uses the fixed subspace $\mathcal{V}_D = \mathrm{span}\{H_\alpha : |\alpha| \le D\}$, which depends only on the (static) teacher variable distribution $z_\star \sim \mathcal{N}(0, I_{k_\star})$. ■

E.7 Time-invariant structure source

Define the fixed feature map $\varphi_D : \mathbb{R}^{k_\star} \to \mathbb{R}^R$ by listing all Hermite polynomials up to degree D:

$$\varphi_D(z_\star) := \big(H_\alpha(z_\star)\big)_{|\alpha| \le D}, \qquad R = N_D.$$

For each time s, define the loading matrix $B_s \in \mathbb{R}^{k \times R}$ whose columns are the coefficient vectors $b_s^{(\alpha)}$ for $|\alpha| \le D$:

$$\mu_s^{(\le D)}(z_\star) \;=\; \sum_{|\alpha| \le D} b_s^{(\alpha)}\, H_\alpha(z_\star) \;=\; B_s\, \varphi_D(z_\star), \tag{52}$$

so that

$$S_s^{(\le D)} \;=\; \mathbb{E}_{z_\star}\big[\mu_s^{(\le D)}(z_\star)\mu_s^{(\le D)}(z_\star)^\top \mid \mathcal{F}_s\big] \;=\; B_s B_s^\top, \tag{53}$$

because $\mathbb{E}[\varphi_D(z_\star)\varphi_D(z_\star)^\top] = I_R$ by orthonormality.

Accumulated loading matrix. Iterating the mean Gram recursion of Lemma 3.2, treating J_q and K_s as deterministic, gives, up to the propagated initial Gram and the second-order terms $(\eta_f/k)^2 J_q G_q J_q$, $G_t \approx (\eta_f/k)^2 \sum_{s<t} \Pi_{s,t}\, K_s\, \Pi_{s,t}^\top$ with damping product $\Pi_{s,t} := (I - \frac{\eta_f}{k} J_{t-1}) \cdots (I - \frac{\eta_f}{k} J_{s+1})$ (empty product $= I$; the contribution injected at step s is propagated through the damping at steps $s+1, \dots, t-1$). Splitting $K_s = S_s + \mathbb{E}_s[\tilde v_s \tilde v_s^\top]$ and truncating $S_s \approx S_s^{(\le D)} = B_s B_s^\top$, we define the *propagated accumulated loading matrix* at the covariance scale $\Sigma_t = n_\perp G_t$:

$$B_t^{\mathrm{acc}} \;:=\; \frac{\eta_f \sqrt{n_\perp}}{k}\big[\Pi_{0,t} B_0 \,\big|\, \Pi_{1,t} B_1 \,\big|\, \cdots \,\big|\, \Pi_{t-1,t} B_{t-1}\big] \;\in\; \mathbb{R}^{k \times tR}, \tag{54}$$

so that $B_t^{\mathrm{acc}} (B_t^{\mathrm{acc}})^\top = n_\perp (\eta_f/k)^2 \sum_{s<t} \Pi_{s,t}\, S_s^{(\le D)}\, \Pi_{s,t}^\top$ captures the propagated teacher-induced contribution to Σ_t (at the same scale used in the bulk-plus-spike approximation of the main text and the variance-inflation channel (16)). The column count tR may be of order n, so B_t^{acc} is generally not low-rank in neuron space, but its columns are indexed by the fixed Hermite feature modes $\{H_\alpha : |\alpha| \le D\}$. The construction also assigns the propagated initial Gram and the propagated centered part to the bulk $\hat c_t I_k$, so it gives a qualitative picture of Σ_t.

Remark E.7 (What "accumulated finite mode" means here). Theorem E.6 shows that, uniformly over macroscopic training time, the accumulated teacher-induced covariance per unit time is well-approximated (in normalized trace norm) by the sum of the same fixed-mode truncations:

$$\frac{1}{t}\sum_{s=0}^{t-1} S_s \approx \frac{1}{t}\sum_{s=0}^{t-1} B_s B_s^\top, \qquad \text{with a fixed structure source } \varphi_D(z_\star) \in \mathbb{R}^R.$$

The structure source φ_D is determined only by the static teacher variable $z_\star$ and does not rotate with time. However, while each per-step factor $S_s^{(\le D)} = B_s B_s^\top$ has rank $\le R$, the accumulated covariance $\sum_{s<t} B_s B_s^\top$ in $\mathbb{R}^{k\times k}$ may have rank up to $\min(tR, k)$, since the neuron-side column spaces of B_s vary with s. The "finite-mode" guarantee is therefore on the teacher side ($\varphi_D \in \mathbb{R}^R$ fixed), not on the neuron-side rank of the accumulated covariance. The accumulated loading matrix B_t^{acc} defined in (54) is used in the variance-inflation analysis of Proposition 3.5 (Appendix G).

F Proof of Theorem 3.4 (asymptotic equivalence)

The proof has four steps:

1. Identify the test loss as a functional of the state (a_t, M_t, Σ_t) with $\Sigma_t = n_\perp G_t = F_t^\perp (F_t^\perp)^\top$ (Sec. F.1).
2. Match drift and diffusion of the parameters (a_t, M_t) between SGD and DD (Sec. F.2).
3. Compute the conditional mean of the complement increment by Stein's identity, show that the directional decoherence ρ^{dec} is produced by ζ_t, and derive the spike covariance Γ_t (Sec. F.3).
4. Prove stability estimates for the reduced state (Lemma F.1) and transport the matching of conditional mean increments to the test loss by a discrete Grönwall argument on the reduced state (Sec. F.4).

F.1 Step 1: Test loss as a functional of the covariance state

The test loss $\mathcal{L}_t = \frac{1}{2}\mathbb{E}_{x,y}[(\hat{y}(x; F_t, a_t) - y)^2]$ depends on F_t only through the joint distribution of the preactivations $h_{i,t} = m_{i,t}^\top z_\star + h_{i,t}^\perp$. Since $z_\star = \Xi^\top x \sim \mathcal{N}(0, I_{k_\star})$ and $h_t^\perp \mid \mathcal{F}_t \sim \mathcal{N}(0, \Sigma_t)$ *exactly*, with covariance state

$$\Sigma_t \;:=\; n_\perp G_t \;=\; F_t^\perp (F_t^\perp)^\top \;\in\; \mathbb{R}^{k\times k}, \tag{55}$$

the test loss is a deterministic functional

$$\mathcal{L}_t \;=\; \Phi(a_t, M_t, \Sigma_t) \;:=\; \tfrac{1}{2}\,\mathbb{E}_{z_\star, h^\perp}\Big[\Big(\tfrac{1}{k}\textstyle\sum_{i=1}^k a_{i,t}\,\sigma(m_{i,t}^\top z_\star + h_i^\perp) - g(z_\star)\Big)^2\Big], \tag{56}$$

where $h^\perp \sim \mathcal{N}(0, \Sigma_t)$ is jointly Gaussian. Step 4 uses that Φ is Lipschitz in the norm $\|\cdot\|_\star$ on the region $\mathcal{R}_n$ defined there. Moreover, given $\mathcal{F}_t$ the SGD step depends on the fresh sample only through $(z_{\star,t}, h_t^\perp, \|g_t\|_2^2)$, and $\|g_t\|_2^2 = h_t^{\perp\top}\Sigma_t^+ h_t^\perp + \chi_t$ with $\chi_t \sim \chi^2_{n_\perp - \mathrm{rank}(\Sigma_t)}$ independent of $(z_{\star,t}, h_t^\perp)$. Hence the conditional law of $X_{t+1} := (a_{t+1}, M_{t+1}, \Sigma_{t+1})$ depends on $F_t^\perp$ only through Σ_t, and (X_t) is a Markov chain; the same holds for Decomposed Dynamics, whose complement rows stay in $\mathrm{range}(P_\perp)$ by (14) and whose noise ζ_t is rotation invariant within $\mathrm{range}(P_\perp)$. (The structural decomposition $\Sigma_t \approx \hat{c}_t I + B_t^{\mathrm{acc}}(B_t^{\mathrm{acc}})^\top$ of Appendix E is not used in the equivalence proof; it serves as the interpretive framework in the variance-inflation analysis of Proposition 3.5.)

F.2 Step 2: drift+diffusion of (a_t, M_t) match between SGD and DD

By construction (Theorem 3.4 item 2(i)), Decomposed Dynamics applies the SGD updates (3)–(4) verbatim to $(a_t^{\mathrm{DD}}, M_t^{\mathrm{DD}})$:

$$\begin{aligned} a_{i,t+1}^{\mathrm{DD}} &= a_{i,t}^{\mathrm{DD}} - \tfrac{\eta_a}{k}\big(e_t^{\mathrm{DD}}\,\sigma(h_{i,t}^{\mathrm{DD}}) + \lambda a_{i,t}^{\mathrm{DD}}\big), \\ m_{i,t+1}^{\mathrm{DD}} &= m_{i,t}^{\mathrm{DD}} - \tfrac{\eta_f}{k}\,e_t^{\mathrm{DD}}\,a_{i,t}^{\mathrm{DD}}\,\sigma'(h_{i,t}^{\mathrm{DD}})\,z_{\star,t}, \end{aligned}$$

with $h_{i,t}^{\mathrm{DD}} = (m_{i,t}^{\mathrm{DD}})^\top z_\star + h_{i,t}^{\perp,\mathrm{DD}}$ and $h_t^{\perp,\mathrm{DD}} \mid \mathcal{F}_t \sim \mathcal{N}(0, \Sigma_t^{\mathrm{DD}})$ where $\Sigma_t^{\mathrm{DD}} := F_t^{\perp,\mathrm{DD}}(F_t^{\perp,\mathrm{DD}})^\top$. Both channels are exact Gaussians; equivalence reduces to comparing the evolutions of Σ_t and Σ_t^{DD} (Steps 3–4).

The conditional laws of the $(a_{i,t}, m_{i,t})$ increments under SGD and under DD agree whenever the complete reduced states (a_t, M_t, Σ_t) and $(a_t^{\mathrm{DD}}, M_t^{\mathrm{DD}}, \Sigma_t^{\mathrm{DD}})$ agree (Step 1); by Lemma F.1 (S1), on the region $\mathcal{R}_n$ of Step 4 the mismatch of the conditional means of these increments is controlled by the $\|\cdot\|_\star$-distance of the states. Step 3 below shows that DD's complement update has the same conditional mean increment as SGD's; together with the matching here, this is the input to Step 4.

F.3 Step 3: complement drift and spike covariance

Write the SGD complement matrix as $F_t^\perp = F_t P_\perp$ and recall the exact rank-one recursion

$$\Delta F_t^\perp := F_{t+1}^\perp - F_t^\perp \;=\; -\frac{\eta_f}{k}\, v_t g_t^\top, \qquad g_t = P_\perp x_t \sim \mathcal{N}(0, P_\perp),$$

where $v_{i,t} = e_t a_{i,t} \sigma'(h_{i,t})$ and $h_{i,t} = m_{i,t}^\top z_{\star,t} + (f_{i,t}^\perp)^\top g_t$.

(i) Conditional mean: the friction term and the full drift. Fix a neuron i and condition on $\mathcal{F}_t$. Let $f_{i,t}^\perp \in \mathbb{R}^n$ be the i-th row of $F_t^\perp$ and $r_{i,t}^2 = \|f_{i,t}^\perp\|_2^2$. Define the complement preactivation (scalar)

$$h_{i,t}^\perp := (f_{i,t}^\perp)^\top g_t.$$

By Gaussian regression in $\mathrm{range}(P_\perp)$, we may decompose

$$g_t \;=\; \frac{h_{i,t}^\perp}{r_{i,t}^2}\, f_{i,t}^\perp \;+\; g_{\perp,t}^{(i)}, \qquad g_{\perp,t}^{(i)} \;\perp\; f_{i,t}^\perp, \qquad \mathbb{E}\Big[g_{\perp,t}^{(i)} g_{\perp,t}^{(i)\top} \,\Big|\, \mathcal{F}_t\Big] = P_\perp - \frac{f_{i,t}^\perp (f_{i,t}^\perp)^\top}{r_{i,t}^2}. \tag{57}$$

Substituting into the row update $f_{i,t+1}^\perp = f_{i,t}^\perp - (\eta_f/k) v_{i,t} g_t$ gives the exact identity

$$f_{i,t+1}^\perp = \Big(1 - \frac{\eta_f}{k}\frac{v_{i,t} h_{i,t}^\perp}{r_{i,t}^2}\Big)\, f_{i,t}^\perp - \frac{\eta_f}{k}\, v_{i,t}\, g_{\perp,t}^{(i)}. \tag{58}$$

The second term $-(\eta_f/k) v_{i,t}\, g_{\perp,t}^{(i)}$ is orthogonal in $\mathrm{range}(P_\perp)$ to $f_{i,t}^\perp$ *by construction* (since $g_{\perp,t}^{(i)} \perp f_{i,t}^\perp$ from the regression decomposition (57)). Its conditional mean $\mathbb{E}[v_{i,t}\, g_{\perp,t}^{(i)} \mid \mathcal{F}_t]$ is generally nonzero — by Stein's identity it equals $\Sigma_t' \,\mathbb{E}[\nabla_{g_\perp^{(i)}} v_{i,t} \mid \mathcal{F}_t]$ where $\Sigma_t' = P_\perp - f_{i,t}^\perp (f_{i,t}^\perp)^\top / r_{i,t}^2$ — but it lives in $\mathrm{range}(\Sigma_t')$, the orthogonal complement of $f_{i,t}^\perp$ within $\mathrm{range}(P_\perp)$. The projection onto $f_{i,t}^\perp$ is therefore zero, but the orthogonal mean is first order in η_f/k and coherent across rows; it is part of the full drift computed below.

Taking $\mathbb{E}[\cdot \mid \mathcal{F}_t]$ in (58) and projecting onto the direction $f_{i,t}^\perp / r_{i,t}$ yields the scalar friction identity

$$\mathbb{E}[(f_{i,t+1}^\perp)^\top u_{i,t} \mid \mathcal{F}_t] = \big(1 - \rho_{i,t}^{\mathrm{fric}}\big)\, r_{i,t}, \qquad \rho_{i,t}^{\mathrm{fric}} = \frac{\eta_f}{k}\frac{\mathbb{E}[v_{i,t} h_{i,t}^\perp \mid \mathcal{F}_t]}{r_{i,t}^2},$$

which is the projection on $f_{i,t}^\perp$ of the full drift computed next.

Full conditional mean. Summing the $f_{i,t}^\perp$-aligned and orthogonal parts, Gaussian integration by parts over $g_t \sim \mathcal{N}(0, P_\perp)$ given $(\mathcal{F}_t, z_{\star,t})$ gives $\mathbb{E}[v_{i,t} g_t \mid \mathcal{F}_t, z_{\star,t}] = \sum_\ell \mathbb{E}[\partial v_{i,t}/\partial h_{\ell,t}^\perp \mid \mathcal{F}_t, z_{\star,t}]\, f_{\ell,t}^\perp$, hence, after averaging over $z_{\star,t}$,

$$\mathbb{E}_t[\Delta F_t^\perp] = -\frac{\eta_f}{k} J_t F_t^\perp = -\widetilde{\Lambda}_t F_t^\perp,$$

with J_t of Lemma 3.2. Its projection on $f_{i,t}^\perp$ is the friction term above; its off-diagonal entries couple the rows. Since $\mathbb{E}_t[\zeta_t] = 0$, (14) has exactly this conditional mean increment.

(ii) Directional decoherence. The orthogonal perturbation in (58) rotates the direction of $f_{i,t}^\perp$ within $\mathrm{range}(P_\perp)$. We compute the rate of this decoherence and show that Decomposed Dynamics reproduces it through ζ_t. Let $u_{i,t} = f_{i,t}^\perp / r_{i,t}$ and define the orthogonal increment

$$\delta_{i,t}^\perp := -\frac{\eta_f}{k}\, v_{i,t}\, g_{\perp,t}^{(i)}, \qquad u_{i,t}^\top \delta_{i,t}^\perp = 0.$$

A second-order Taylor expansion of the normalization map $x \mapsto x/\|x\|_2$ yields, for neurons with $r_{i,t}$ bounded below and $|a_{i,t}| \le A_n$ (event (85)),

$$\mathbb{E}\big[u_{i,t+1}^\top u_{i,t} \,\big|\, \mathcal{F}_t\big] = 1 - \frac{1}{2 r_{i,t}^2}\, \mathbb{E}\big[\|\delta_{i,t}^\perp\|_2^2 \,\big|\, \mathcal{F}_t\big] + o\Big(\frac{1}{n}\Big). \tag{59}$$

We use $\mathbb{E}\big[\|g_{\perp,t}^{(i)}\|_2^2 \mid \mathcal{F}_t\big] = \mathrm{Tr}(P_\perp - f_i^\perp (f_i^\perp)^\top / r_i^2) = n_\perp - 1$ and write $\mathbb{E}_t[v_{i,t}^2 \|g_\perp^{(i)}\|^2] = \mathbb{E}_t[v_{i,t}^2]\,(n_\perp - 1) + \mathrm{Cov}_t(v_{i,t}^2, \|g_\perp^{(i)}\|^2)$. The covariance is bounded by Cauchy–Schwarz: $|\mathrm{Cov}_t(v_i^2, \|g_\perp^{(i)}\|^2)| \le \sqrt{\mathrm{Var}_t(v_i^2)}\sqrt{\mathrm{Var}_t(\|g_\perp^{(i)}\|^2)}$, with $\mathrm{Var}_t(\|g_\perp^{(i)}\|^2) = 2(n_\perp - 1)$ (chi-square)

and $\mathrm{Var}_t(v_i^2) \le \mathbb{E}_t[v_i^4] \le C a_{i,t}^4$ pointwise on the high-probability event of Lemma I.1 (since $|v_i| \le C|a_i|$ from $|e_t|, \|\sigma'\|_\infty \le C$ on the same event). Hence

$$\mathbb{E}\big[\|\delta_{i,t}^\perp\|_2^2 \,\big|\, \mathcal{F}_t\big] \;=\; \Big(\frac{\eta_f}{k}\Big)^2 \Big[\, \mathbb{E}[v_{i,t}^2 \mid \mathcal{F}_t]\,(n_\perp - 1) + O\big(a_{i,t}^2\sqrt{n_\perp}\big)\,\Big],$$

where, for such neurons, the O-term contributes $O(n^{-3/2}\log n) = o(1/n)$ to (59). Inserting into (59) and using $n_\perp - 1 = n_\perp + o(n)$ yields the leading-order decay

$$\mathbb{E}[u_{i,t+1}^\top u_{i,t} \mid \mathcal{F}_t] = 1 - \rho_{i,t}^{\mathrm{dec}} + o\Big(\frac{1}{n}\Big), \qquad \rho_{i,t}^{\mathrm{dec}} = \frac{n_\perp}{2}\Big(\frac{\eta_f}{k}\Big)^2 \frac{\mathbb{E}[v_{i,t}^2 \mid \mathcal{F}_t]}{r_{i,t}^2}.$$

For Decomposed Dynamics, row i of ζ_t has $\mathbb{E}_t\|\zeta_{i,t}\|_2^2 = n_\perp(\eta_f/k)^2\mathbb{E}_t[v_{i,t}^2]$, and the drift $-\widetilde{\Lambda}_t F_t^\perp$ changes $u_{i,t}^\top u_{i,t+1}$ only at order $o(1/n)$, so (14) gives the same $\mathbb{E}_t[u_{i,t+1}^\top u_{i,t}] = 1 - \rho_{i,t}^{\mathrm{dec}} + o(1/n)$.

(iii) Spike covariance. The SGD complement increment satisfies

$$\Delta F_t^\perp \Delta(F_t^\perp)^\top = \Big(\frac{\eta_f}{k}\Big)^2 v_t v_t^\top \, \|g_t\|_2^2.$$

Writing $\|g_t\|_2^2 = n_\perp + (\|g_t\|_2^2 - n_\perp)$ with the second term having $\mathbb{E}_t[\,\cdot\,] = 0$ and $\mathrm{Var}_t(\,\cdot\,) = 2n_\perp$ (chi-square), Cauchy–Schwarz on the covariance gives, entry-wise,

$$\mathbb{E}_t\big[(v_i v_j)\,\|g_t\|_2^2\big] = n_\perp\,\mathbb{E}_t[v_i v_j] + O\big(\sqrt{n_\perp}\cdot\sqrt{\mathbb{E}_t[v_i^2 v_j^2]}\big) = n_\perp\,(K_t)_{ij} + O\big(|a_i a_j|\sqrt{n_\perp}\big),$$

hence in Frobenius norm,

$$\mathbb{E}[\Delta F_t^\perp \Delta(F_t^\perp)^\top \mid \mathcal{F}_t] = n_\perp\,\Gamma_t + E_t^{\mathrm{cov}}, \qquad \|E_t^{\mathrm{cov}}\|_F \le C(\eta_f/k)^2\sqrt{n_\perp}\,\|a_t\|_2^2 = O(n^{-1/2}),$$

with $E_t^{\mathrm{cov}} = n_\perp E_t$ for E_t of Lemma 3.2, and $\Gamma_t = (\eta_f/k)^2 K_t$ as defined in (15). In the surrogate, $\mathrm{vec}(\zeta_t^\top) \sim \mathcal{N}(0, \Gamma_t \otimes P_\perp)$ ensures $\mathbb{E}[\zeta_t\zeta_t^\top \mid \mathcal{F}_t] = \mathrm{Tr}(P_\perp)\Gamma_t = n_\perp\,\Gamma_t$, matching the leading-order covariance of the SGD increment.

(iv) Summary of Step 3. The update (14) has SGD's conditional mean increment $-\widetilde{\Lambda}_t F_t^\perp$ (i) and its leading neuron-space second moment $n_\perp\Gamma_t$ (iii); the directional decoherence (ii) is produced by ζ_t. Higher moments are not matched: SGD's linear Gram increment $-\frac{\eta_f}{k}(v_t h_t^{\perp\top} + h_t^\perp v_t^\top)$ correlates v_t with the same $h_t^\perp$, whereas, given $\mathcal{F}_t$, ζ_t is a centered Gaussian independent of the fresh sample, so the conditional variances of projections of $\Delta\Sigma_t$ differ.

F.4 Step 4: Grönwall argument for the test loss

For either procedure write $X_t = (a_t, M_t, \Sigma_t)$, a Markov chain with $\mathcal{L}_t = \Phi(X_t)$ (Step 1). For $x = (a, M, \Sigma)$ let $\|x\|_\star := k^{-1/2}\big(\|a\|_2 + \|M\|_F + \|\mathrm{diag}(\Sigma)\|_2\big) + k^{-1}\|\Sigma\|_2$, where $\|\mathrm{diag}(\Sigma)\|_2 := (\sum_i \Sigma_{ii}^2)^{1/2}$, let $\mathcal{R}_n := \{x : \Sigma \succeq 0,\ \max_i|a_i| \le \bar{A}_\tau,\ \max_i(\|m_i\|_2^2 + \Sigma_{ii}) \le \bar{Q}_\tau\}$ with $\bar{A}_\tau := e^{2C_a\tau/\alpha}(A+1)$ for $C_a := \eta_a\|\sigma\|_\infty \max(\|\sigma\|_\infty, \|g\|_\infty)$ and A the bound on $|a_{i,0}|$ in Theorem 3.4 (for n large, $\bar{A}_\tau$ bounds $\max_i|a_{i,t}|$; proof of Lemma F.1) and $\bar{Q}_\tau$ from Lemma F.1, and let $b(x)$ be the conditional mean increment of Decomposed Dynamics at state x.

Lemma F.1 (Stability of the reduced dynamics)**.** *Assume **A1–A5** and $|a_{i,0}| \le A$ almost surely, and let Decomposed Dynamics update (a_t, M_t) verbatim by the SGD rule, as in Step 2. There are $\bar{Q}_\tau, L, L_\Phi < \infty$, independent of n, such that, for all n large enough: **(S1)** $\|b(x) - b(y)\|_\star \le \frac{L}{n}\|x - y\|_\star$ and $|\Phi(x) - \Phi(y)| \le L_\Phi\|x - y\|_\star$ for all $x, y \in \mathcal{R}_n$; **(S2)** with probability $1 - o(1)$, $X_t^{\mathrm{SGD}} \in \mathcal{R}_n$ and $X_t^{\mathrm{DD}} \in \mathcal{R}_n$ for all $t \le \tau n$; **(S3)** for each procedure, the martingale part $\mathcal{M}_t := \sum_{s<t}\big(X_{s+1} - X_s - \mathbb{E}[X_{s+1} - X_s \mid X_s]\big)$ satisfies $\sup_{t\le\tau n}\|\mathcal{M}_t\|_\star = O_{\mathbb{P}}(n^{-1/2})$.*

Proof. Write $s_j := \|\sigma^{(j)}\|_\infty$, finite for $j \le 3$ by **A3**; below, C denotes constants that depend only on τ, α, A, η_a, η_f, λ, $k_\star$, $\|g\|_\infty$ and $s_0, \dots, s_3$, and n is large enough that $n/k \le 2/\alpha$ and $\eta_a\lambda \le k$.

Drift map. Fix $x \in \mathcal{R}_n$, let $z \sim \mathcal{N}(0, I_{k_\star})$ and $h^\perp \sim \mathcal{N}(0, \Sigma)$ be independent, and set $h := Mz + h^\perp$, $e := \frac{1}{k}a^\top\sigma(h) - g(z)$, $w := a \circ \sigma'(h)$, $v := e\,w$ and $D_v := \partial v/\partial h = \frac{1}{k}ww^\top + \mathrm{diag}(\gamma)$, $\gamma := e\,a \circ \sigma''(h)$. Then $J = \mathbb{E}D_v$, $K = \mathbb{E}[vv^\top]$ and, by Stein's identity, $J\Sigma = \mathbb{E}[vh^{\perp\top}]$. With $c_a := \eta_a n/k$ and $c_f := \eta_f n/k$,

$$nb(x) = \Big(-c_a\big(\mathbb{E}[e\,\sigma(h)] + \lambda a\big),\ -c_f\mathbb{E}[vz^\top],\ -c_f(J\Sigma + \Sigma J) + c_f\tfrac{\eta_f}{k}J\Sigma J + c_f^2\tfrac{n_\perp}{n}K\Big). \quad (60)$$

On $\mathcal{R}_n$, pointwise in $(z, h^\perp)$: $|e| \le C_e := s_0\bar{A}_\tau + \|g\|_\infty$, $|w_i| \le s_1\bar{A}_\tau$, $|v_i| \le C_v := C_e s_1 \bar{A}_\tau$ and $\|D_v\|_2 \le C_D := s_1^2\bar{A}_\tau^2 + C_e s_2 \bar{A}_\tau$, so $\|J\|_2 \le C_D$; moreover $\|\Sigma\|_2 \le \operatorname{Tr}\Sigma \le k\bar{Q}_\tau$.

(S1). $\mathcal{R}_n$ is convex. Mollifying σ does not increase any s_j, and b, Φ depend continuously on $(\sigma, \sigma', \sigma'')$, so we may assume $\sigma \in C^\infty$. Let $x, y \in \mathcal{R}_n$, $\delta = (\delta a, \delta M, \delta\Sigma) := y - x$ and $x_\theta := x + \theta\delta$. Gaussian interpolation [50, Lemma 1.3.1] gives $\frac{d}{d\theta}\mathbb{E}_{x_\theta}[F] = \mathbb{E}[\partial_{(a,M)}F[\delta a, \delta M]] + \frac{1}{2}\mathbb{E}\langle\nabla^2_{h^\perp}F, \delta\Sigma\rangle$ for smooth $F(a, M; z, h^\perp)$ of polynomial growth. Applying it to u_i, $v_i z$, $v_i h_j^\perp$ and $v_i v_j$, with $\nabla^2_{h^\perp}(v_i h_j^\perp) = h_j^\perp \nabla^2 v_i + \nabla v_i \mathbf{e}_j^\top + \mathbf{e}_j \nabla v_i^\top$ ($\mathbf{e}_j$ the j-th standard basis vector), gives for $u \in \{e\,\sigma(h), v\}$ and the random vector $\hat{\delta}u := \partial_{(a,M)}u[\delta a, \delta M] + \frac{1}{2}\big(\langle\nabla_h^2 u_i, \delta\Sigma\rangle\big)_{i\le k}$

$$\begin{aligned} &\tfrac{d}{d\theta}\mathbb{E}u = \mathbb{E}\hat{\delta}u, \qquad \tfrac{d}{d\theta}\mathbb{E}[vz^\top] = \mathbb{E}[\hat{\delta}v\, z^\top], \qquad \tfrac{d}{d\theta}(J\Sigma) = \mathbb{E}[\hat{\delta}v\, h^{\perp\top}] + J\delta\Sigma, \\ &\tfrac{d}{d\theta}K = \mathbb{E}[\hat{\delta}v\, v^\top + v\,\hat{\delta}v^\top] + \mathbb{E}[D_v\,\delta\Sigma\, D_v]. \end{aligned} \tag{61}$$

Explicitly, with $\delta e := \frac{1}{k}\big(\sigma(h)^\top\delta a + w^\top\delta M z\big)$,

$$\begin{aligned} \hat{\delta}v_i = {}& \delta e\, w_i + e\sigma'(h_i)\delta a_i + e\, a_i\sigma''(h_i)\,\delta m_i^\top z \\ &+ \tfrac{1}{2k}w_i\langle a\circ\sigma''(h), \operatorname{diag}\delta\Sigma\rangle + \tfrac{1}{k}a_i\sigma''(h_i)(\delta\Sigma\, w)_i + \tfrac{1}{2}e\, a_i\sigma'''(h_i)\,\delta\Sigma_{ii}, \end{aligned}$$

and $\hat{\delta}(e\,\sigma)_i$ has the same terms with $a_i\sigma^{(j)}(h_i)$ replaced by $\sigma^{(j-1)}(h_i)$ and without the δa_i term; σ''' enters only through the last term of $\hat{\delta}v_i$. Using $k\,\delta e^2 \le 2s_0^2\|\delta a\|_2^2 + 2s_1^2\bar{A}_\tau^2\|\delta M z\|_2^2$, $\|w\|_2^2 \le k s_1^2\bar{A}_\tau^2$, $\langle a\circ\sigma''(h), \operatorname{diag}\delta\Sigma\rangle^2 \le k s_2^2\bar{A}_\tau^2\|\operatorname{diag}\delta\Sigma\|_2^2$, $\|\delta\Sigma\, w\|_2 \le \|\delta\Sigma\|_2\|w\|_2$ and $\mathbb{E}\|\delta M z\|_2^2 = \|\delta M\|_F^2$, we obtain

$$\mathbb{E}\|\hat{\delta}u\|_2^2 \le C\big(\|\delta a\|_2^2 + \|\delta M\|_F^2 + \|\operatorname{diag}\delta\Sigma\|_2^2 + k^{-1}\|\delta\Sigma\|_2^2\big) \le Ck\|\delta\|_\star^2. \tag{62}$$

For random $\psi \in \mathbb{R}^k$ and $\varphi \in \mathbb{R}^p$, Cauchy–Schwarz gives $\|\mathbb{E}[\psi\varphi^\top]\|_2 \le (\mathbb{E}\|\psi\|_2^2)^{1/2}\|\mathbb{E}[\varphi\varphi^\top]\|_2^{1/2}$ and, if $p = k$, $\|\operatorname{diag}\mathbb{E}[\psi\varphi^\top]\|_2 \le (\mathbb{E}\|\psi\|_2^2)^{1/2}\max_i(\mathbb{E}\varphi_i^2)^{1/2}$. Here $\|\mathbb{E}[\varphi\varphi^\top]\|_2 = 1$ for $\varphi \in \{1, z\}$, while $\|\mathbb{E}[\varphi\varphi^\top]\|_2 \le k\max(\bar{Q}_\tau, C_v^2)$ and $\max_i\mathbb{E}\varphi_i^2 \le \max(\bar{Q}_\tau, C_v^2)$ for $\varphi \in \{h^\perp, v\}$. With $\|\cdot\|_F \le \sqrt{k_\star}\|\cdot\|_2$ on $\mathbb{R}^{k\times k_\star}$, (62) therefore bounds every expectation of the form $\mathbb{E}[\hat{\delta}u\,\varphi^\top]$ in (61) by $C\|\delta\|_\star$ in each part of $\|\cdot\|_\star$ ($k^{-1/2}$ times the Euclidean, Frobenius or diagonal norm, and k^{-1} times the operator norm). The diagonal bound with $\varphi = h^\perp$ is the only place in (S1) where the per-neuron bound $\Sigma_{ii} \le \bar{Q}_\tau$, rather than its average, is used. For the remaining terms, $\|J\delta\Sigma\|_2 \le C_D\|\delta\Sigma\|_2$ and $\|\mathbb{E}[D_v\delta\Sigma D_v]\|_2 \le C_D^2\|\delta\Sigma\|_2$, while $(J\delta\Sigma)_{ii} = \frac{1}{k}\mathbb{E}[w_i(\delta\Sigma w)_i] + \mathbb{E}[\gamma_i]\,\delta\Sigma_{ii}$ and $(D_v\delta\Sigma D_v)_{ii} = \frac{w_i^2}{k^2}w^\top\delta\Sigma w + \frac{2}{k}w_i\gamma_i(\delta\Sigma w)_i + \gamma_i^2\,\delta\Sigma_{ii}$ give diagonals of Euclidean norm at most $C(\|\operatorname{diag}\delta\Sigma\|_2 + k^{-1/2}\|\delta\Sigma\|_2)$, by $|w_i| \le s_1\bar{A}_\tau$ and $\|w\|_2 \le \sqrt{k}\,s_1\bar{A}_\tau$. Integrating over $\theta \in [0,1]$, the changes of $\mathbb{E}[e\,\sigma(h)]$, $\mathbb{E}[vz^\top]$, $P := J\Sigma$ and K between x and y are at most $C\|\delta\|_\star$ in the corresponding parts of $\|\cdot\|_\star$. Finally, with primes denoting evaluation at y, $J\Sigma J - J'\Sigma'J' = (P - P')J + J'(P - P')^\top + J'(\Sigma' - \Sigma)J$, and $\|\operatorname{diag}(\cdot)\|_2 \le \sqrt{k}\|\cdot\|_2$, so the term $c_f\frac{\eta_f}{k}J\Sigma J$ inherits the bound on P. Summing the blocks of (60) proves the bound on b. For $\Phi = \frac{1}{2}\mathbb{E}[e^2]$, $\frac{d}{d\theta}\Phi(x_\theta) = \mathbb{E}[e\,\delta e] + \frac{1}{2k^2}\mathbb{E}[w^\top\delta\Sigma\, w] + \frac{1}{2k}\mathbb{E}\big[e\langle a\circ\sigma''(h), \operatorname{diag}\delta\Sigma\rangle\big]$, which is at most $L_\Phi\|\delta\|_\star$ by the same bounds (without σ''').

(S2). For both procedures $|e_t| \le s_0\max_i|a_{i,t}| + \|g\|_\infty$, so their common a-update gives $\max_i|a_{i,t+1}| \le (1 + C_a/k)\max_i|a_{i,t}| + C_a/k$ (as in Appendix I, Step 2) and hence, since $|a_{i,0}| \le A$, deterministically $\max_i|a_{i,t}| \le e^{C_a t/k}(A+1) \le \bar{A}_\tau$ for $t \le \tau n$; thus $|v_{i,t}| \le C_v$. Let $q_{i,t} := \|m_{i,t}\|_2^2 + (\Sigma_t)_{ii}$, $T := \min\{t : \max_i q_{i,t} > \bar{Q}_\tau\} \wedge \lfloor\tau n\rfloor$, and split $q_{i,t+1} - q_{i,t} = \beta_{i,t} + \xi_{i,t}$ into its conditional mean and a martingale increment. For SGD, $q_{i,t+1} - q_{i,t} = -\frac{2\eta_f}{k}v_i h_i + \frac{\eta_f^2}{k^2}v_i^2\|x_t\|_2^2$; for Decomposed Dynamics it is $-\frac{2\eta_f}{k}v_i m_i^\top z + \frac{\eta_f^2}{k^2}v_i^2\|z\|_2^2 - \frac{2\eta_f}{k}(J\Sigma)_{ii} + \frac{\eta_f^2}{k^2}(J\Sigma J)_{ii} + 2\langle\tilde{f}_i, \zeta_i\rangle + \|\zeta_i\|_2^2$, with $\tilde{f}_i$ the i-th row of $\tilde{F} := (I - \widetilde{\Lambda}_t)F_t^\perp$ and $\zeta_i \sim \mathcal{N}(0, \Gamma_{ii}P_\perp)$, $\Gamma_{ii} \le (\eta_f/k)^2 C_v^2$. Since $|\mathbb{E}_t[v_i h_i]|, |(J\Sigma)_{ii}| = |\mathbb{E}_t[v_i h_i^\perp]|$ and $|\mathbb{E}_t[v_i m_i^\top z]|$ are at most $C_v\sqrt{q_{i,t}}$, and, for $t < T$, $(J\Sigma J)_{ii} \le (\sum_l |J_{il}|\Sigma_{ll}^{1/2})^2 \le C\bar{Q}_\tau$ (as $|J_{il}| \le s_1^2\bar{A}_\tau^2/k$ for $l \ne i$), we get for $t < T$ and n large: $\beta_{i,t} \le \frac{C_0}{k}(1 + \sqrt{q_{i,t}})$ with C_0 not depending on $\bar{Q}_\tau$, and $\mathbb{E}_t[\xi_{i,t}^4] \le C(1 + \bar{Q}_\tau)^2 k^{-4}$. By Burkholder's inequality, $\mathbb{E}\max_{t\le\tau n}|\sum_{s<t\wedge T}\xi_{i,s}|^4 \le C(\tau n)^2(1 + \bar{Q}_\tau)^2 k^{-4} = O(k^{-2})$, so with probability $1 - O(1/k)$ these k martingales all stay in $[-1, 1]$; moreover $\max_i q_{i,0} \le 2$ with probability $1 - ke^{-n/8}$, since $nq_{i,0} \sim \chi_n^2$. On these events, $q_{i,t} \le 3 + c_\tau(1 + \sqrt{\bar{Q}_\tau})$ for all i and $t \le T$,

with $c_\tau := 2C_0\tau/\alpha$, and this is $< \bar{Q}_\tau$ for $\bar{Q}_\tau := 4(1+c_\tau)^2$; hence $T = \lfloor \tau n \rfloor$. Finally $\Sigma_t \succeq 0$ for both procedures.

(S3). For $t < T$, the increments $\xi_t := X_{t+1} - X_t - \mathbb{E}_t[X_{t+1} - X_t]$ satisfy $\mathbb{E}_t\|\xi_t^a\|_2^2 + \mathbb{E}_t\|\xi_t^M\|_F^2 + \mathbb{E}_t\|\operatorname{diag}\xi_t^\Sigma\|_2^2 \le C/k$ and $\mathbb{E}_t\|\xi_t^\Sigma\|_F^2 \le C$: for SGD, from $\Delta\Sigma_t = -\frac{\eta_f}{k}(v_t h_t^{\perp\top} + h_t^\perp v_t^\top) + \frac{\eta_f^2}{k^2}\|g_t\|_2^2 v_t v_t^\top$ and $\mathbb{E}_t\|h_t^\perp\|_2^2 = \operatorname{Tr}\Sigma_t \le k\bar{Q}_\tau$; for Decomposed Dynamics, from $(\xi_t^\Sigma)_{ii} = 2\langle \tilde{f}_i, \zeta_i\rangle + \|\zeta_i\|_2^2 - \mathbb{E}_t\|\zeta_i\|_2^2$ with $\sum_i \Gamma_{ii}\|\tilde{f}_i\|_2^2 \le \max_i \Gamma_{ii} \operatorname{Tr}(\tilde{F}\tilde{F}^\top) = O(1/k)$ and $\operatorname{Var}_t\|\zeta_i\|_2^2 = 2n_\perp\Gamma_{ii}^2$, and from $\mathbb{E}_t\|\tilde{F}\zeta_t^\top\|_F^2 = \operatorname{Tr}\Gamma_t \operatorname{Tr}(\tilde{F}\tilde{F}^\top)$ and $\mathbb{E}_t\|\zeta_t\|_F^4 \le 3(n_\perp \operatorname{Tr}\Gamma_t)^2$. Doob's L^2 maximal inequality, applied to each block of the martingales stopped at T, gives $\mathbb{E}\sup_{t\le\tau n}\|\mathcal{M}_{t\wedge T}\|_\star^2 \le C\tau n/k^2 = O(1/n)$, and $\mathbb{P}(T < \lfloor\tau n\rfloor) = o(1)$ by (S2). ■

The constants grow with τ through $\bar{A}_\tau$ and $\bar{Q}_\tau$. The bounded initialization and the bound on σ''' in **A3** make L finite and independent of n in this proof: for Gaussian $a_{i,0}$, $\max_i a_i^2 \asymp \log k$, and the term $a_i a_j \mathbb{E}[e^2\sigma''(h_i)\sigma''(h_j)]$ of the Σ_{ij}-derivative of K generically makes the constant in (S1) at least of order $\log k$; if σ'' has a jump at 0, the m-drift of a neuron with $m_i = 0$ is generically only $\frac{1}{2}$-Hölder in Σ_{ii} near $\Sigma_{ii} = 0$.

Proof of Theorem 3.4. *(1) Drift mismatch.* By (S2), SGD lies in $\mathcal{R}_n$ for all $t \le \tau n$ with probability $1 - o(1)$. By Steps 2–3, its conditional mean increment at $x \in \mathcal{R}_n$ is $b(x) + \Delta b(x)$, where the (a, M) components of Δb vanish and its Σ component is $n_\perp E_t - (\eta_f/k)^2 J_t \Sigma J_t$, with E_t from Lemma 3.2. By Appendix F.3, Step (iii), $|(n_\perp E_t)_{ij}| \le C(\eta_f/k)^2\sqrt{n}\,|a_i a_j|$, so $k^{-1}\|n_\perp E_t\|_2 \le k^{-1}\|n_\perp E_t\|_F = O(n^{-3/2})$ and $k^{-1/2}\|\operatorname{diag}(n_\perp E_t)\|_2 \le Ck^{-5/2}\sqrt{n}\,\bar{A}_\tau\|a\|_2 = O(n^{-3/2})$. Since $\|J_t\|_2 \le C_D$ and $\|\Sigma\|_2 \le \operatorname{Tr}\Sigma \le k\bar{Q}_\tau$, the term $(\eta_f/k)^2 J_t\Sigma J_t$ contributes $O(n^{-3/2})$ to $\|\Delta b(x)\|_\star$ (using $\|\operatorname{diag}(\cdot)\|_2 \le \|\cdot\|_F \le \|J_t\|_2^2 \operatorname{Tr}\Sigma$). Hence $\sum_{s<\tau n}\|\Delta b(X_s^{\mathrm{SGD}})\|_\star = O(n^{-1/2})$. *(2) Grönwall.* Let $D_t := X_t^{\mathrm{SGD}} - X_t^{\mathrm{DD}}$, so $D_0 = 0$ and $D_t = \sum_{s<t}\big(b(X_s^{\mathrm{SGD}}) - b(X_s^{\mathrm{DD}})\big) + \sum_{s<t}\Delta b(X_s^{\mathrm{SGD}}) + \mathcal{M}_t^{\mathrm{SGD}} - \mathcal{M}_t^{\mathrm{DD}}$. On the event that both chains stay in $\mathcal{R}_n$, (S1) gives $\|D_t\|_\star \le \frac{L}{n}\sum_{s<t}\|D_s\|_\star + \vartheta_n$ with $\vartheta_n := \sum_{s<\tau n}\|\Delta b(X_s^{\mathrm{SGD}})\|_\star + 2\sup_{t\le\tau n}\max\big(\|\mathcal{M}_t^{\mathrm{SGD}}\|_\star, \|\mathcal{M}_t^{\mathrm{DD}}\|_\star\big)$, and the discrete Grönwall inequality gives $\sup_{t\le\tau n}\|D_t\|_\star \le \vartheta_n e^{L\tau} = O_{\mathbb{P}}(n^{-1/2})$ by (1), (S2) and (S3). *(3) Test loss.* By (S1), $\sup_{t\le\tau n}|\mathcal{L}_t^{\mathrm{SGD}} - \mathcal{L}_t^{\mathrm{DD}}| \le L_\Phi \sup_{t\le\tau n}\|D_t\|_\star = O_{\mathbb{P}}(n^{-1/2})$. ■

The argument uses only conditional mean increments; the unmatched fluctuations of (iv) enter only through (S3). The proof applies to any procedure with the same drift map b whose increments obey the per-step bounds used in Lemma F.1 (so that (S2)–(S3) hold).

G Proof of Proposition 3.5 (variance inflation by teacher-induced structure)

We prove Proposition 3.5 by iterating the law of total variance for the effective predictor (16). Throughout we condition on the sigma field $\mathcal{F}_t$ by using the notation $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_t]$, so that $(a_{i,t}, m_{i,t})$ and $(\hat{c}_t, B_t^{\mathrm{acc}})$ are measurable (deterministic).

Setup. Let $z_\star \sim \mathcal{N}(0, I_{k_\star})$, $\varepsilon \sim \mathcal{N}(0, I_k)$, $s \sim \mathcal{N}(0, I_{tR})$ be independent of each other and of $\mathcal{F}_t$, and define $\hat{y}_t^{\mathrm{eff}}$ as in (16). Define

$$y^{\mathrm{eff}} := g(z_\star), \qquad 2\mathcal{L}_t^{\mathrm{eff}} = \mathbb{E}\big[(\hat{y}_t^{\mathrm{eff}} - y^{\mathrm{eff}})^2 \,\big|\, \mathcal{F}_t\big].$$

Step 1: decompose by conditioning on $z_\star$. Conditioning first on $\sigma(z_\star, \mathcal{F}_t)$ and applying the law of total variance yields

$$\mathbb{E}\big[(\hat{y}_t^{\mathrm{eff}} - y^{\mathrm{eff}})^2 \,\big|\, \mathcal{F}_t\big] = \mathbb{E}_{z_\star}\left[\Big(\mathbb{E}[\hat{y}_t^{\mathrm{eff}} \mid z_\star, \mathcal{F}_t] - y^{\mathrm{eff}}\Big)^2 + \operatorname{Var}\big(\hat{y}_t^{\mathrm{eff}} \mid z_\star, \mathcal{F}_t\big)\right]. \tag{63}$$

The first term in (63) is exactly $\mathcal{L}_{\mathrm{bias},t}$ from (19). It remains to decompose the conditional variance term.

Step 2: decompose the conditional variance by conditioning on s. Given $(z_\star, \mathcal{F}_t)$, the only randomness in $\hat{y}_t^{\mathrm{eff}}$ is from the independent pair (ε, s). Apply the law of total variance with conditioning σ-field $\sigma(z_\star, s, \mathcal{F}_t)$:

$$\operatorname{Var}\big(\hat{y}_t^{\mathrm{eff}} \mid z_\star, \mathcal{F}_t\big) = \mathbb{E}_s\big[\operatorname{Var}_\varepsilon\big(\hat{y}_t^{\mathrm{eff}} \mid z_\star, s, \mathcal{F}_t\big)\big] + \operatorname{Var}_s\big(\mathbb{E}_\varepsilon\big[\hat{y}_t^{\mathrm{eff}} \mid z_\star, s, \mathcal{F}_t\big]\big). \tag{64}$$

Taking $\mathbb{E}_{z_\star}[\cdot]$ of both sides identifies the first term as $\mathcal{L}_{\mathrm{res},t}$ from (19) and the second as $\mathcal{L}_{\mathrm{str},t}$ from (20).

Step 3: combine and conclude nonnegativity / strictness. Substituting (64) into (63) gives the decomposition (18). Nonnegativity of $\mathcal{L}_{\mathrm{str},t}$ is immediate since it is an expectation of a variance. Moreover, $\mathcal{L}_{\mathrm{str},t} = 0$ whenever $B_t^{\mathrm{acc}} = 0$ since then $\hat{y}_t^{\mathrm{eff}}$ is independent of s and the inner variance vanishes almost surely. Conversely, if $\mathbb{E}_\varepsilon[\hat{y}_t^{\mathrm{eff}} \mid z_\star, s, \mathcal{F}_t]$ is not almost surely constant in s on a set of $z_\star$ of positive probability, then the variance in (20) is positive on that set and hence $\mathcal{L}_{\mathrm{str},t} > 0$. The condition of Proposition 3.5 ensures this. Indeed, differentiating $\mathbb{E}_\varepsilon[\hat{y}_t^{\mathrm{eff}} \mid z_\star, s, \mathcal{F}_t] = \frac{1}{k}\sum_i a_i\, \tilde{\sigma}_{\hat{c}_t}(m_i^\top z_\star + (B_t^{\mathrm{acc}} s)_i)$ with $\tilde{\sigma}_{\hat{c}_t}(u) := \mathbb{E}_{\varepsilon_1}[\sigma(u + \sqrt{\hat{c}_t}\varepsilon_1)]$, the s-derivative at $s = 0$ has j-th component $\frac{1}{k}\sum_i a_i\, \tilde{\sigma}'_{\hat{c}_t}(m_i^\top z_\star)\,(B_t^{\mathrm{acc}})_{ij} = \frac{1}{k}\big((B_t^{\mathrm{acc}})^\top \omega_t(z_\star)\big)_j$, since $\tilde{\sigma}'_{\hat{c}_t}(u) = \mathbb{E}_{\varepsilon_1}[\sigma'(u + \sqrt{\hat{c}_t}\varepsilon_1)]$. By that condition, this derivative is non-zero at some $z_\star$, hence, by continuity in $z_\star$, on an open set of $z_\star$, which has positive probability. On that set the conditional mean (of class C^1 in s) is not almost surely constant in s, since s has full support; hence $\mathcal{L}_{\mathrm{str},t} > 0$. ■

Remark (the cross-neuron term). For the network itself, with $h_i := m_{i,t}^\top z_\star + h_i^\perp$ and $h^\perp \sim \mathcal{N}(0, \Sigma_t)$ independent of $z_\star$, write $2\mathcal{L}_t = \mathrm{Bias}^2 + \mathrm{VarDiag} + \mathrm{Cross}$ with $\mathrm{Bias}^2 := \mathbb{E}_{z_\star}(\mathbb{E}[\hat{y} \mid z_\star] - g(z_\star))^2$ for $\hat{y} := \frac{1}{k}\sum_i a_{i,t}\sigma(h_i)$, $\mathrm{VarDiag} := \frac{1}{k^2}\mathbb{E}_{z_\star}\sum_i a_{i,t}^2 \operatorname{Var}(\sigma(h_i) \mid z_\star)$ and $\mathrm{Cross} := \frac{1}{k^2}\mathbb{E}_{z_\star}\sum_{i\neq j} a_{i,t}a_{j,t}\operatorname{Cov}(\sigma(h_i), \sigma(h_j) \mid z_\star)$. The first two terms depend on Σ_t only through its diagonal, so, relative to a network with the same per-neuron variances and no cross-neuron structure, the structure changes $2\mathcal{L}_t$ by exactly Cross. To first order (Mehler), $\mathrm{Cross} \approx \frac{1}{k^2}\mathbb{E}_{z_\star}[w^\top(\Sigma_t - \operatorname{diag}\Sigma_t)w]$ with $w_i := a_{i,t}\mathbb{E}[\sigma'(h_i) \mid z_\star]$: it is positive for spikes aligned with w and can be negative for structure orthogonal to w. Since $\operatorname{Var}(\hat{y} \mid z_\star) \geq 0$, we have $\mathrm{Cross} \geq -\mathrm{VarDiag}$, and $\mathrm{VarDiag} \leq \|\sigma\|_\infty^2 B_\tau/k$ on **A6**; hence negative values of Cross are $O(1/k)$, while, to first order, spikes of Σ_t with eigenvalue of order k aligned with w give $\mathrm{Cross} = \Theta(1)$ when $\|w\|_2^2$ is of order k. Likewise, as the ε_i are independent across neurons, $\mathcal{L}_{\mathrm{res},t} \leq \|\sigma\|_\infty^2 B_\tau/k$, and $\mathcal{L}_{\mathrm{str},t}$ differs from the cross-neuron term of $\hat{y}_t^{\mathrm{eff}}$ (its excess over the same per-neuron variances, as above) by at most $\|\sigma\|_\infty^2 B_\tau/k$.

H Proof of Proposition 3.6 (fast alignment of m_t and a_t)

This appendix formalizes the oracle no-spike dynamics and proves Proposition 3.6. We work throughout in the single-index case $k_\star = 1$ (and hence $m_{i,t} \in \mathbb{R}$).

H.1 Additional conditions and oracle no-spike dynamics

The proposition imposes the following additional conditions; **B3** is an idealization.

B1. Odd teacher and odd activation, with nontrivial Hermite-1 coefficients. The teacher $g : \mathbb{R} \to \mathbb{R}$ is bounded, Lipschitz, and odd, and the activation σ is odd (recall that $\sigma \in W^{3,\infty}$ from **A3**). Furthermore,

$$\kappa_g \;:=\; \mathbb{E}[g(Z)Z] \;\neq\; 0, \qquad Z \sim \mathcal{N}(0,1). \tag{65}$$

B2. Non-degenerate average slope.

$$\kappa_\sigma \;:=\; \mathbb{E}[\sigma'(Z)] \;\neq\; 0, \qquad Z \sim \mathcal{N}(0,1). \tag{66}$$

B3. Ideal isotropic complement channel. The oracle dynamics below replace the complement projections by synthetic variables $z_t = (z_{1,t}, \ldots, z_{k,t}) \sim \mathcal{N}(0, I_k)$, independent across t, of $z_{0,t}$ and of $\mathcal{F}_t$, with $(a_0, m_0) = (a_0, F_0\xi)$ initialized as in **A5**. This idealizes the complement channel (unit variances, no cross-neuron covariance); directions U_0 frozen at initialization would instead give covariance $U_0U_0^\top$.

Like Theorem 3.4, the proposition also assumes bounded second-layer initialization: $|a_{i,0}| \leq A$ almost surely for a constant $A < \infty$. This holds, e.g., for Rademacher or uniform $a_{i,0}$, and is standard in rigorous analyses of gradient-based training of two-layer networks [4, 7, 35]. It is used only to bound $\max_i |a_{i,t}|$ by a constant (Appendix I, Step 2, and Step 4 below); under **A5** alone this maximum can grow like $\sqrt{\log k}$ (e.g., for Gaussian $a_{i,0}$). Our simulations use $a_{i,0} \sim \mathcal{N}(0,1)$; rerunning them with bounded $a_{i,0}$ of the same variance gives the same results.

Oracle no-spike dynamics. Let $\Xi = \xi \in \mathbb{R}^n$ and $z_{0,t} := \xi^\top x_t \sim \mathcal{N}(0,1)$. In place of $f_{i,t}^\top x_t$, the oracle uses the synthetic preactivations of **B3**,

$$h_{i,t} := m_{i,t} z_{0,t} + z_{i,t}, \qquad z_{i,t} \sim \mathcal{N}(0,1) \text{ i.i.d. and independent of } z_{0,t}.$$

Define

$$\hat{y}_t = \frac{1}{k}\sum_{\ell=1}^{k} a_{\ell,t}\sigma(h_{\ell,t}), \qquad e_t = \hat{y}_t - g(z_{0,t}).$$

The oracle no-spike dynamics evolves (a_t, m_t) using the conditional drift of the single-sample SGD updates (with ridge term $\lambda \geq 0$ on a consistent with $\frac{\lambda}{2k}\|a\|_2^2$):

$$a_{i,t+1} = a_{i,t} - \frac{\eta_a}{k}\,\mathbb{E}[e_t\,\sigma(h_{i,t}) \,|\, \mathcal{F}_t] - \frac{\eta_a \lambda}{k}\,a_{i,t}, \tag{67}$$

$$m_{i,t+1} = m_{i,t} - \frac{\eta_f}{k}\,\mathbb{E}[e_t\,a_{i,t}\,\sigma'(h_{i,t})\,z_{0,t} \,|\, \mathcal{F}_t]\,. \tag{68}$$

H.2 Proof of Proposition 3.6

Let $\kappa := \kappa_g \kappa_\sigma$ and recall $\cos_t = a_t^\top m_t/(\|a_t\|_2\|m_t\|_2)$.

Step 1: linearization of the teacher contribution. Fix any $m \in \mathbb{R}$. Define the scalar functions

$$\Phi(m) \;:=\; \mathbb{E}[g(Z_0)\,\sigma(mZ_0 + Z)]\,, \qquad \Psi(m) \;:=\; \mathbb{E}[g(Z_0)\,Z_0\,\sigma'(mZ_0 + Z)]\,, \tag{69}$$

where (Z_0, Z) are i.i.d. $\mathcal{N}(0,1)$. Under **A3** and **B1**, Φ is C^1 with $\Phi'(m) = \Psi(m)$ by dominated convergence (bounded σ' and bounded g). Moreover, by independence of Z_0 and Z and the fact that g is odd (**B1**, so $\mathbb{E}[g(Z_0)] = 0$),

$$\Phi(0) = \mathbb{E}[g(Z_0)]\,\mathbb{E}[\sigma(Z)] = 0, \qquad \Psi(0) = \mathbb{E}[g(Z_0)Z_0]\,\mathbb{E}[\sigma'(Z)] = \kappa_g\kappa_\sigma = \kappa.$$

Since σ' is Lipschitz (**A3**) and $g(Z_0)Z_0$ has finite moments (bounded g), Ψ is Lipschitz and hence, for any $|m| \leq \bar{m}$,

$$\Phi(m) = \kappa\, m + R_\Phi(m), \qquad \Psi(m) = \kappa + R_\Psi(m), \qquad |R_\Phi(m)| \leq C\bar{m}\,|m|,\;\; |R_\Psi(m)| \leq C|m|. \tag{70}$$

Here $C < \infty$ depends only on the constants in **A3** and the bound on g.

Step 2: conditional drift is a 2×2 linear system plus a small remainder. Fix t and condition on $\mathcal{F}_t$. Under **B3**, $(z_{0,t}, z_t)$ are independent standard Gaussians. Write the conditional drift terms in (67)–(68) as

$$\mathbb{E}[e_t\sigma(h_{i,t}) \mid \mathcal{F}_t] = \mathbb{E}[\hat{y}_t\,\sigma(h_{i,t}) \mid \mathcal{F}_t] - \mathbb{E}[g(z_{0,t})\sigma(h_{i,t}) \mid \mathcal{F}_t], \tag{71}$$

$$\mathbb{E}[e_t a_{i,t}\sigma'(h_{i,t})z_{0,t} \mid \mathcal{F}_t] = \mathbb{E}[\hat{y}_t\,a_{i,t}\sigma'(h_{i,t})z_{0,t} \mid \mathcal{F}_t] - \mathbb{E}[g(z_{0,t})\,a_{i,t}\sigma'(h_{i,t})z_{0,t} \mid \mathcal{F}_t]. \tag{72}$$

Teacher contributions. Since $h_{i,t} = m_{i,t}z_{0,t} + z_{i,t}$ and $(z_{0,t}, z_{i,t})$ are i.i.d. standard Gaussians,

$$\mathbb{E}[g(z_{0,t})\sigma(h_{i,t}) \mid \mathcal{F}_t] = \Phi(m_{i,t}), \qquad \mathbb{E}[g(z_{0,t})\,\sigma'(h_{i,t})z_{0,t} \mid \mathcal{F}_t] = \Psi(m_{i,t}).$$

Thus on the event $\mathcal{E}_{\bar{m}}(t) := \{\max_{i\leq k}|m_{i,t}| \leq \bar{m}\}$, (70) yields

$$\mathbb{E}[g(z_{0,t})\sigma(h_{i,t}) \mid \mathcal{F}_t] = \kappa\, m_{i,t} + r_{i,t}^{(m)}, \qquad \mathbb{E}[g(z_{0,t})\,a_{i,t}\sigma'(h_{i,t})z_{0,t} \mid \mathcal{F}_t] = \kappa\, a_{i,t} + r_{i,t}^{(a)}, \tag{73}$$

with $|r_{i,t}^{(m)}| \leq C\bar{m}\,|m_{i,t}|$ and $|r_{i,t}^{(a)}| \leq C\bar{m}\,|a_{i,t}|$.

Self-interaction contributions. On $\mathcal{E}_{\bar{m}}(t)$, expanding $\hat{y}_t = \frac{1}{k}\sum_\ell a_\ell\sigma(h_\ell)$ with $h_\ell = m_\ell z_{0,t} + z_{\ell,t}$ and using **B1** (σ odd, so that $\tilde{\sigma}(u) := \mathbb{E}_z[\sigma(u+z)]$ is also odd, with $\tilde{\sigma}(0) = 0$ and Taylor expansion $\tilde{\sigma}(u) = \kappa_\sigma u + O(u^3)$), we obtain by independence of $z_{\ell,t}$ across ℓ,

$$\mathbb{E}[\hat{y}_t^2 \mid \mathcal{F}_t] \;\leq\; \frac{\|\sigma\|_\infty^2}{k^2}\sum_\ell a_{\ell,t}^2 + \frac{\kappa_\sigma^2}{k^2}\,(a_t^\top m_t)^2 + O(\bar{m}^4) \;=\; O\Big(\tfrac{1}{k} + \bar{m}^2\Big),$$

since on $\mathcal{E}_{\bar{m}}$ together with $\max_i|a_{i,t}| \leq \bar{A}$ from Step 4, $|a_t^\top m_t|^2 \leq \|a_t\|^2\|m_t\|^2 \leq k\bar{m}^2 \cdot O(k)$. By Cauchy–Schwarz and bounded σ, σ',

$$\big|\mathbb{E}[\hat{y}_t\,\sigma(h_{i,t}) \mid \mathcal{F}_t]\big| \;\leq\; C\,(\bar{m} + k^{-1/2}), \qquad \big|\mathbb{E}[\hat{y}_t\,a_{i,t}\sigma'(h_{i,t})z_{0,t} \mid \mathcal{F}_t]\big| \;\leq\; C\,|a_{i,t}|\,(\bar{m} + k^{-1/2}). \tag{74}$$

(In the proportional regime the $k^{-1/2}$ term vanishes, so we treat the right-hand sides as $O(\bar{m})$ in what follows.)

Combining (71)–(74) with (73) gives, on $\mathcal{E}_{\bar{m}}(t)$,

$$\mathbb{E}[e_t\sigma(h_{i,t}) \mid \mathcal{F}_t] = -\kappa\, m_{i,t} + \delta_{i,t}^{(a)}, \tag{75}$$

$$\mathbb{E}[e_t a_{i,t}\sigma'(h_{i,t}) z_{0,t} \mid \mathcal{F}_t] = -\kappa\, a_{i,t} + \delta_{i,t}^{(m)}, \tag{76}$$

with bounds

$$|\delta_{i,t}^{(a)}| \le C\,\bar{m}\,(|m_{i,t}| + 1), \qquad |\delta_{i,t}^{(m)}| \le C\,\bar{m}\,|a_{i,t}|. \tag{77}$$

Insert (75)–(76) into (67)–(68). In vector form, for t such that $\mathcal{E}_{\bar{m}}(t)$ holds,

$$\begin{bmatrix} a_{t+1} \\ m_{t+1} \end{bmatrix} = \begin{bmatrix} a_t \\ m_t \end{bmatrix} + \frac{1}{k} \underbrace{\begin{bmatrix} -\eta_a\lambda & \eta_a\kappa \\ \eta_f\kappa & 0 \end{bmatrix}}_{=:G} \begin{bmatrix} a_t \\ m_t \end{bmatrix} - \frac{1}{k} \begin{bmatrix} \eta_a\,\delta_t^{(a)} \\ \eta_f\,\delta_t^{(m)} \end{bmatrix}. \tag{78}$$

Step 3: alignment for the limiting linear flow. Ignore the remainder in (78) and consider the linear ODE on macroscopic time $s = t/k$:

$$\frac{d}{ds}\begin{bmatrix} a(s) \\ m(s) \end{bmatrix} = G \begin{bmatrix} a(s) \\ m(s) \end{bmatrix}, \qquad (a(0), m(0)) = (a_0, m_0). \tag{79}$$

The matrix G has eigenvalues

$$\mu_\pm \;=\; \frac{-\eta_a\lambda \pm \sqrt{\eta_a^2\lambda^2 + 4\eta_a\eta_f\kappa^2}}{2}, \qquad \mathrm{gap} := \mu_+ - \mu_- = \sqrt{\eta_a^2\lambda^2 + 4\eta_a\eta_f\kappa^2} \;>\; 0, \tag{80}$$

since $\kappa \neq 0$ by **B1–B2**. Let $X(s) := [a(s), m(s)] \in \mathbb{R}^{k\times 2}$. Then $X(s) = X(0)e^{sG^\top}$, hence $\det(X(s)^\top X(s)) = e^{2s\,\mathrm{tr}G}\det(X(0)^\top X(0))$ with $\mathrm{tr}\,G = -\eta_a\lambda$. Since $\det(X^\top X) = \|a\|_2^2\|m\|_2^2(1-\cos^2)$,

$$1 - \cos_s^2 = (1 - \cos_0^2)\, e^{-2\eta_a\lambda s}\, \frac{\|a_0\|_2^2\,\|m_0\|_2^2}{\|a(s)\|_2^2\,\|m(s)\|_2^2}. \tag{81}$$

Under **A5**, with probability $1 - o(1)$: $\|a_0\|_2^2 \asymp k$, $\|m_0\|_2^2 = \alpha(1 + o(1))$ and $|a_0^\top m_0| = O(\sqrt{\log n})$. Expanding $m(s) = m_0 + s\eta_f\kappa a_0 + O(s^2)(\|a_0\|_2 + \|m_0\|_2)$ and using $\|a(s)\|_2 \ge \|a_0\|_2 e^{-\|G\|s}$ gives, for $s \le \tau_0'$, with $\tau_0' > 0$ small, $\|m(s)\|_2^2 \ge \|m_0\|_2^2(1 - Cs\sqrt{\log n}) + c_0\, s^2 k$ with $c_0 > 0$ (from the term $s^2\eta_f^2\kappa^2\|a_0\|_2^2$, as $\kappa \neq 0$), hence, with $x_+^{-1} := \infty$ for $x \le 0$,

$$1 - \cos_s^2 \le (1 - \cos_0^2)\, e^{Cs} \min\Big\{ \big(1 - Cs\sqrt{\log n}\big)_+^{-1},\; \frac{C\alpha}{s^2 k} \Big\}.$$

For $s \le k^{-1/3}$ the first term gives $1 - \cos_s^2 \le (1 - \cos_0^2)e^{-cs} + O(k^{-1/3}\sqrt{\log n})$ for any fixed $c > 0$; for $s \ge k^{-1/3}$ the second gives $1 - \cos_s^2 = O(k^{-1/3})$. Hence, with $c := 2\,\mathrm{gap}$,

$$1 - \cos_s^2 \le (1 - \cos_0^2)\, e^{-cs} + o(1) \qquad \text{uniformly on } [0, \tau_0'], \tag{82}$$

and $\|a(s)\|_2 \ge c_1\sqrt{k}$, $\|m(s)\|_2 \ge c_1 \max(\sqrt{\alpha}, s\sqrt{k})$ on $[0, \tau_0']$ for a constant $c_1 > 0$.

Step 4: bootstrap and stability of the Euler scheme. We return to the discrete recursion (78); fix $\varepsilon > 0$.

Bootstrap. The oracle dynamics are deterministic given (a_0, m_0). As in Steps 1–2 of Appendix I, $\frac{1}{k}\|a_t\|_2^2$ stays bounded and $|e_t| \le C$, so $|a_{i,t+1}| \le (1 + C/k)|a_{i,t}| + C/k$ and, since $|a_{i,0}| \le A$, $\max_i |a_{i,t}| \le \bar{A} := e^{C\tau_0'}(A + 1)$ for $t \le \tau_0' k$. The drift (68) is $a_{i,t}$ times a bounded factor, so $|m_{i,t} - m_{i,0}| \le C\eta_f\bar{A}\, t/k$. Since $\max_i |m_{i,0}| = O(\sqrt{\log n/n})$ with probability $1 - o(1)$, $\mathcal{E}_{\bar{m}}(t)$ holds for all $t \le \tau_0 k$ with $\tau_0 := \min\{\tau_0', \bar{m}/(2C\eta_f\bar{A})\}$ (τ_0' from Step 3 above), for every $\bar{m} > 0$ and n large.

Relative error. On this window, let $X^{(k)}(s)$ interpolate the discrete trajectory, let $X(s)$ solve (79), and let $D := X^{(k)} - X$. By (78), D solves the linear equation driven by the remainders (77), up to an Euler error $O(s\|X\|_2/k)$. Since $\|\delta^{(m)}\|_2 \le C\bar{m}\|a\|_2$ and $\|\delta^{(a)}\|_2 \le C\bar{m}(\|m\|_2 + \sqrt{k})$, Grönwall gives $\|D_a(s)\|_2 + \|D_m(s)\|_2 \le C\bar{m}\, s\sqrt{k}$ on $[0, \tau_0]$. With the lower bounds on $\|a(s)\|_2$ and $\|m(s)\|_2$

from Step 3, the relative errors of both directions are $O(\bar{m})$, and since $1-\cos^2$ is Lipschitz in the two unit directions, $|(1-\cos_t^2)-(1-\cos_s^2)|_{s=t/k} \le C'\bar{m}$. Choosing $\bar{m} := \varepsilon/(2C')$ and using (82), whose $o(1)$ term is below $\varepsilon/2$ for n large,

$$1-\cos_t^2 \;\le\; (1-\cos_0^2)\exp\Big(-c\,\frac{t}{k}\Big)+\varepsilon, \qquad \forall\, 0\le t\le \tau_0(\varepsilon)\,k,$$

which is the discrete analogue of (82). Moreover, for $s \ge k^{-1/3}$ Step 3 gives $1-\cos_s^2 = O(k^{-1/3})$ for the linear flow, hence $1-\cos_t^2 \le C'\bar{m} + O(k^{-1/3}) \le \varepsilon$ for all $k^{2/3} \le t \le \tau_0 k$ and n large, which is (21). ∎

I Moment stability under the mean-field scaling

Lemma I.1 (Moment stability: **A6** holds under **A1–A5**). *Assume **A1–A5**. Consider SGD with loss $\ell = \frac{1}{2}(\hat{y}-y)^2 + \frac{\lambda}{2k}\|a\|_2^2$ and updates (3)–(4). Fix $\tau > 0$ and suppose $\eta_a, \eta_f = O(1)$ and $\lambda \ge 0$ are fixed constants. Then there exists $B_\tau < \infty$ — depending only on τ, α, η_a, η_f, λ, $k_\star$, L_g, $\|g\|_\infty$, $\|\sigma\|_\infty$, $\|\sigma'\|_\infty$, $\|\sigma''\|_\infty$, and $K_{\rm sg}$ — such that **A6** holds, i.e.,*

$$\sup_{t\le \tau n} \frac{1}{k}\sum_{i=1}^{k}\Big(a_{i,t}^2 + \|m_{i,t}\|_2^2 + r_{i,t}^2\Big) \;\le\; B_\tau$$

with probability $1-o(1)$ as $n\to\infty$. Moreover, with probability $1-o(1)$, $\max_{i\le k}\sup_{t\le\tau n} a_{i,t}^2 \le A_n^2 := C'_\tau \log k$ (the event $\mathcal{E}_a^{\rm sup}$ of (85)).

Proof. Fix $\tau > 0$ and let $T_{\max} := \lfloor \tau n \rfloor$. Write $\bar{a}_t := (1/k)\|a_t\|_2^2$, $w_t := (1/k)\sum_i \|m_{i,t}\|_2^2$, $u_t := (1/k)\sum_i r_{i,t}^2$.

Step 1 (control of $\bar{a}_t$ and $|e_t|$). Expanding $\|a_{t+1}\|_2^2$ from (4) and using bounded σ ($\|\sigma_t\|_2 \le \|\sigma\|_\infty\sqrt{k}$), bounded teacher ($|y_t| \le \|g\|_\infty$), and Young's inequality on the cross term, one obtains the deterministic recursion

$$\bar{a}_{t+1} \;\le\; (1 + C_1/k)\,\bar{a}_t \;+\; C_1/k, \qquad C_1 = C(\eta_a, \lambda, \|\sigma\|_\infty, \|g\|_\infty). \tag{83}$$

By the law of large numbers and **A5** (sub-Gaussian, hence finite second moment), $\bar{a}_0 = O(1)$ with probability $1-o(1)$. Discrete Grönwall on (83) then yields

$$\sup_{t\le T_{\max}} \bar{a}_t \;\le\; C_\tau, \qquad \sup_{t\le T_{\max}} |e_t| \;\le\; \|\sigma\|_\infty\sqrt{C_\tau} + \|g\|_\infty, \tag{84}$$

with probability $1-o(1)$, with C_τ depending only on $\tau, \alpha, \eta_a, \lambda, \|\sigma\|_\infty, \|g\|_\infty, \mathbb{E}[a_{i,0}^2]$.

Step 2 (per-neuron sup-norm event $\mathcal{E}_a^{\rm sup}$). Let $A_t := \max_{i\le k}|a_{i,t}|$. Since $|e_t| \le \|\sigma\|_\infty A_t + \|g\|_\infty$, the a-update gives, for n large enough that $\eta_a\lambda \le k$ and $n/k \le 2/\alpha$, the deterministic recursion $A_{t+1} \le (1 + C_a/k)A_t + C_a/k$ with $C_a := \eta_a\|\sigma\|_\infty \max(\|\sigma\|_\infty, \|g\|_\infty)$, and Grönwall yields

$$\max_{i\le k}\sup_{t\le T_{\max}} |a_{i,t}| \;\le\; e^{C_a T_{\max}/k}\big(\max_i |a_{i,0}| + 1\big) \;\le\; e^{2C_a\tau/\alpha}\big(\max_i |a_{i,0}| + 1\big).$$

By the sub-Gaussian maximal inequality, $\max_i |a_{i,0}| \le K_{\rm sg}\sqrt{2\log k}\,(1+o(1))$ with probability $1-o(1)$, defining the event

$$\mathcal{E}_a^{\rm sup} \;:=\; \Big\{\max_{i\le k}\sup_{t\le T_{\max}} a_{i,t}^2 \le A_n^2\Big\}, \qquad A_n^2 := C'_\tau \log k, \tag{85}$$

with $\mathbb{P}(\mathcal{E}_a^{\rm sup}) \ge 1-o(1)$.

Step 3 (control of w_t and u_t via drift+martingale). The recursions $m_{i,t+1} = m_{i,t} - (\eta_f/k)v_{i,t}z_{\star,t}$ and $f_{i,t+1}^\perp = f_{i,t}^\perp - (\eta_f/k)v_{i,t}g_t$, with $v_{i,t} = e_t a_{i,t}\sigma'(f_{i,t}^\top x_t)$, $z_{\star,t} \sim \mathcal{N}(0, I_{k_\star})$ and $g_t \sim \mathcal{N}(0, P_\perp)$, expand into per-neuron increments

$$\|m_{i,t+1}\|^2 - \|m_{i,t}\|^2 = -\tfrac{2\eta_f}{k}v_i(m_i^\top z_\star) + \tfrac{\eta_f^2}{k^2}v_i^2\|z_\star\|^2,$$

and analogously for $r_{i,t}^2$ with g_t and $h_i^\perp := \langle f_i^\perp, g_t\rangle$ in place of $z_\star$ and $m_i^\top z_\star$. Note $|v_{i,t}| \le C|a_{i,t}|$ pathwise on Step 1's event.

Drift. By conditional Cauchy–Schwarz ($\mathbb{E}_t[v_i(m_i^\top z_\star)] \le \sqrt{\mathbb{E}_t[v_i^2]}\sqrt{\mathbb{E}_t[(m_i^\top z_\star)^2]} \le C|a_i|\|m_i\|$, and analogously $\mathbb{E}_t[v_i h_i^\perp] \le C|a_i|r_i$), Cauchy in i, and Step 1's bound on $\bar{a}_t$,

$$|\mathbb{E}_t[w_{t+1} - w_t]| \;\le\; \tfrac{1}{k}K_\tau(\sqrt{w_t}+1), \qquad |\mathbb{E}_t[u_{t+1} - u_t]| \;\le\; \tfrac{1}{k}K_\tau(\sqrt{u_t}+1). \tag{86}$$

For the stopping times defined below, $w_s \le B^{(m)}$ for $s < T^{(m)}$ (resp. $u_s \le B^*$ for $s < T^{(r)}$), so the drift accumulated over the steps $s < T^{(m)} \wedge t$ (resp. $s < T^{(r)} \wedge t$) is at most $\tau K_\tau(\sqrt{B}+1)$ for every $t \le T_{\max}$, with $B = B^{(m)}$ (resp. B^*) and K_τ enlarged to absorb the factor $n/k \to 1/\alpha$.

Martingale variance via $\mathcal{E}_a^{\sup}$. On $\mathcal{E}_a^{\sup}$ and Step 1's event, $|v_i| \le C|a_i| \le CA_n$ pointwise, so $\sum_i v_i^4 \le CA_n^2 \sum_i a_i^2 \le CA_n^2 k$. Using $\mathbb{E}_t[(m_i^\top z_\star)^2] = \|m_i\|^2$, $\mathbb{E}_t[(h_i^\perp)^2] = r_i^2$, $\mathbb{E}[\|z_\star\|^4] = O(1)$, and $\mathbb{E}[\|g_t\|^4] = O(n^2)$, expanding $(\Delta\|m_i\|^2)^2$ and $(\Delta r_i^2)^2$ and summing in i:

$$\sum_i \mathbb{E}_t[(\Delta\|m_i\|^2)^2] \le \frac{C\,A_n^2}{k}(w_t+1), \qquad \sum_i \mathbb{E}_t[(\Delta r_i^2)^2] \le \frac{C\,A_n^2}{k}(u_t+1),$$

hence, by Cauchy–Schwarz in i, the centered martingale increments $\xi_t^{(w)} := (w_{t+1}-w_t) - \mathbb{E}_t[w_{t+1}-w_t]$ and $\xi_t^{(u)} := (u_{t+1}-u_t) - \mathbb{E}_t[u_{t+1}-u_t]$ satisfy

$$\mathbb{E}_t[\xi_t^{(w)2}] \le \frac{C\,A_n^2(w_t+1)}{k^2}, \qquad \mathbb{E}_t[\xi_t^{(u)2}] \le \frac{C\,A_n^2(u_t+1)}{k^2}. \tag{87}$$

Let $M_t^{(w)} := \sum_{s<t}\xi_s^{(w)}$ and $M_t^{(u)} := \sum_{s<t}\xi_s^{(u)}$.

Doob localization and bootstrap. Define stopping times $T^{(m)} := \inf\{t \le T_{\max} : w_t > B^{(m)}\}$, $T^{(r)} := \inf\{t \le T_{\max} : u_t > B^*\}$, and $T^\circ := \inf\{t \le T_{\max} :$ Step 1's event or $\mathcal{E}_a^{\sup}$ fails at $t\}$, all $\wedge T_{\max}$, and let $T^{*,m} := T^{(m)} \wedge T^\circ$, $T^{*,r} := T^{(r)} \wedge T^\circ$. On $\{s < T^{*,m}\}$ (resp. $\{s < T^{*,r}\}$), (87) holds with $w_s \le B^{(m)}$ (resp. $u_s \le B^*$). Doob's L^2 maximal inequality applied to the stopped martingales yields

$$\mathbb{E}\big[\sup_t (M^{(w)}_{T^{*,m}\wedge t})^2\big],\ \mathbb{E}\big[\sup_t (M^{(u)}_{T^{*,r}\wedge t})^2\big] \;=\; O\Big(\frac{(B+1)\,\tau\log n}{n}\Big) \;\to\; 0,$$

where B stands for $B^{(m)}$ or B^*. By Markov at fixed $\delta > 0$, $\sup_t |M^{(w)}_{T^{*,m}\wedge t}| = o_p(1)$ and $\sup_t |M^{(u)}_{T^{*,r}\wedge t}| = o_p(1)$, and a union bound with $\mathbb{P}(T^\circ < T_{\max}) = o(1)$ gives the same conclusion with T°-stopping removed.

The self-consistency

$$1 + \tau K_\tau(\sqrt{B^{(m)}}+1) + \delta \le B^{(m)}, \qquad 2 + \tau K_\tau(\sqrt{B^*}+1) + \delta \le B^*, \tag{88}$$

is solvable at constants $B^{(m)}, B^* = O((\tau K_\tau)^2 + \delta^2 + 1)$ (LHS grows only as $\sqrt{B}$). Telescoping the w- and u-recursions and bounding the martingale by δ on the high-probability event yields $w_{T^{(m)}\wedge t} \le B^{(m)}$ and $u_{T^{(r)}\wedge t} \le B^*$ for all $t \le T_{\max}$. Since $w_{T^{(m)}} > B^{(m)}$ if $T^{(m)} < T_{\max}$, and likewise for u, this gives $T^{(m)} = T^{(r)} = T_{\max}$. With $w_0 = O(1/n)$ from $\|m_{i,0}\|^2 \sim \chi^2_{k_\star}/n$ and $u_0 = 1 + o(1)$ from $\|f_{i,0}^\perp\|^2 \sim \chi^2_{n-k_\star}/n$, $w_0 \le 1$ and $u_0 \le 2$ with probability $1 - o(1)$.

Conclusion. Setting $B_\tau := C_\tau + B^{(m)} + B^*$ — a constant depending only on $(\tau, \alpha, \eta_a, \eta_f, \lambda, k_\star, L_g, \|g\|_\infty, \|\sigma\|_\infty, \|\sigma'\|_\infty, \|\sigma''\|_\infty, K_{\mathrm{sg}})$ — proves the lemma, establishing the moment-stability event **A6** under **A1**–**A5**. ∎

I.1 Coupled second-moment bounds

The Sobolev bound in Lemma E.3 of Appendix E controls $\sum_i \mathbb{E}_{z_\star}[\|\nabla_{z_\star}\mu_{i,t}(z_\star)\|_2^2 \mid \mathcal{F}_t]$, which decomposes (via the chain rule) into terms of order $\sum_i a_{i,t}^2$ and the coupled sum $\sum_i a_{i,t}^2\|m_{i,t}\|_2^2$. The first is bounded by Lemma I.1; we now bound the coupled sum directly using the per-neuron sup-norm event $\mathcal{E}_a^{\sup}$ of (85).

Lemma I.2 (Coupled second-moment bounds)**.** *Assume **A1–A5**. Fix $\tau > 0$. There exists a sequence $B'_\tau(n) = O(\log n)$, with constants depending only on the same parameters as B_τ in Lemma I.1, such that with probability $1 - o(1)$ as $n \to \infty$,*

$$\sup_{t \le \tau n} \tfrac{1}{k}\sum_{i=1}^k a_{i,t}^2\|m_{i,t}\|_2^2 \;\le\; B'_\tau(n), \qquad \sup_{t \le \tau n} \tfrac{1}{k}\sum_{i=1}^k a_{i,t}^2 r_{i,t}^2 \;\le\; B'_\tau(n). \tag{89}$$

One may take $B'_\tau(n) := A_n^2 B_\tau = O(\log n)$*;* (89) *then holds on the intersection of* $\mathcal{E}_a^{\sup}$ *with the event* ***A6****.*

Proof. On $\mathcal{E}_a^{\sup}$ (eq. (85)), $\sup_{i,t} a_{i,t}^2 \le A_n^2 = O(\log n)$. Hence

$$\tfrac{1}{k}\sum_i a_{i,t}^2 \|m_{i,t}\|_2^2 \;\le\; A_n^2 \cdot \tfrac{1}{k}\sum_i \|m_{i,t}\|_2^2 \;\le\; A_n^2 \cdot B_\tau,$$

on the event **A6** of Lemma I.1, and similarly for $\sum_i a_{i,t}^2 r_{i,t}^2$. This event has probability $1 - o(1)$. ■

Consequence for the bulk-plus-spike approximation. Combining (89) with the Sobolev bound of Lemma E.3 (which is established in Appendix E), Theorems E.4 and E.6 hold with rate $G^2(n)/(D+1)$ where $G^2(n) = O(\log n)$ via Lemma I.2. The Hermite truncation of S_t and of the undamped sum $\sum_{s<t} S_s$ underlying the bulk-plus-spike approximation $n_\perp G_t \approx \hat{c}_t I + B_t^{\mathrm{acc}}(B_t^{\mathrm{acc}})^\top$ used in Proposition 3.5 is therefore established at rate $O(\log n/(D+1))$ in normalized Frobenius norm; choosing the truncation degree $D = D(n)$ with $\log n/(D+1) \to 0$ (e.g., $D = \lceil (\log n)^2 \rceil$) makes this $o(1)$. The same rate holds for the damped accumulation $\sum_s \Pi_{s,t} S_s \Pi_{s,t}^\top$ behind B_t^{acc} whenever the propagators are uniformly bounded, $\sup_{s<t\le \tau n} \|\Pi_{s,t}\|_2 = O(1)$.

J Extensions: multi-index teacher and beyond smooth activations

J.1 From single-index ($k_\star = 1$) to multi-index ($k_\star > 1$)

Replace $z_0 = \xi^\top x \in \mathbb{R}$ by $z_\star = \Xi^\top x \in \mathbb{R}^{k_\star}$ and $m_{i,t} \in \mathbb{R}$ by $m_{i,t} = \Xi^\top f_{i,t} \in \mathbb{R}^{k_\star}$. All arguments (except for Proposition 3.6) extend with the same structure: $g_t = P_\perp x_t$ remains Gaussian and independent of $z_{\star,t}$, and the conditional Gaussianity of test-time projections $F_t^\perp g$ holds with $P_\perp = I - \Xi\Xi^\top$. Result by result:

- **Complement update.** (7) holds verbatim, with $n_\perp = \mathrm{Tr}(P_\perp) = n - k_\star$.
- **Lemma 3.1.** Set $\mu_s := \mathbb{E}_s[v_s \mid z_{\star,s}]$ and $\mathcal{G}_s := \sigma(\mathcal{F}_s, z_{\star,s})$. Given $\mathcal{G}_s$, $h_{i,s} = m_{i,s}^\top z_{\star,s} + h_{i,s}^\perp$ with $h_s^\perp \sim \mathcal{N}(0, \Sigma_s)$, as for $k_\star = 1$, so the decoupling bound of Appendix B (with $m_{i,t} z_{0,t}$ replaced by $m_{i,t}^\top z_{\star,t}$) and the martingale arguments of Appendix C carry over unchanged.
- **Lemma 3.2.** Stein's identity is applied over $h_t^\perp$ given $(\mathcal{F}_t, z_{\star,t})$ (Appendix D); J_t, K_t and the error bound are unchanged.
- **Finite-mode structure.** Appendix E is written for general $k_\star$. The truncation keeps the $N_D = \binom{k_\star + D}{D}$ Hermite modes of total degree at most D, rather than $D+1$, so a multi-index teacher can inject a richer set of coherent modes.
- **Theorem 3.4.** Steps 1–3 of Appendix F are written in terms of $z_{\star,t}$ and $M_t \in \mathbb{R}^{k \times k_\star}$, and the test loss still depends on the complement only through Σ_t. Lemma F.1 is proved for $M_t \in \mathbb{R}^{k \times k_\star}$ (the norm $\|\cdot\|_\star$ uses $\|M\|_F$, and $k_\star$ enters its constants only through moments of $\|z_\star\|_2$), and Step 4 is unchanged.
- **Proposition 3.5** is stated for $z_\star \in \mathbb{R}^{k_\star}$ and needs no change; neither do the moment bounds of Appendix I.
- **Proposition 3.6** does not extend verbatim. For $k_\star > 1$ the signal state is the matrix $M_t \in \mathbb{R}^{k \times k_\star}$ rather than a vector, and its coupling to a_t depends on the direction-dependent Hermite coefficients of g; a multi-index alignment result requires a separate matrix-valued analysis.

J.2 Beyond smooth activations (e.g., ReLU)

Assumption **A3** ($\sigma \in W^{3,\infty}$) is dictated by the proof technique rather than by the underlying phenomena. Boundedness of σ'' enters our analysis via three ingredients: (i) Stein's identity for J_t in Lemma 3.2 and Step 3 of Appendix F, and Lemma F.1 in its Step 4, which also uses bounded σ'''; (ii) the chain rule for $\sigma'(h_{i,t})$ in Lemma E.3 and the Lipschitz bound on Ψ in Appendix H; and (iii) the Gaussian-interpolation covariance bound of Lemma B.1, used for Lemma 3.1. (Proposition 3.6 additionally requires an odd σ, **B1**, which ReLU violates.) In Figure 2 (ReLU student in panels (a)–(b) and in one curve of (c)), Decomposed Dynamics closely tracks SGD, suggesting that the smoothness requirement is technical rather than essential.

Several standard routes can extend the present analysis to non-smooth activations such as ReLU:

1. **Mollification.** Replace σ by a smooth surrogate $\sigma^{(\delta)} := \sigma_M * \phi_\delta$ for a smooth mollifier ϕ_δ with bandwidth $\delta > 0$, with σ_M the truncation of item 2 for unbounded σ (and $\sigma_M := \sigma$ otherwise), so that $\sigma^{(\delta)} \in W^{3,\infty}$ and the present analysis applies for each $\delta > 0$. Quantitative tracking of the resulting bounds as $\delta \to 0$ and, for unbounded σ, $M \to \infty$ (with $\delta = \delta(n) \to 0$ and $M = M(n) \to \infty$ slowly) recovers the result for σ in the limit, provided the σ- and $\sigma^{(\delta)}$-trajectories can be compared over $t \le \tau n$. For Theorem 3.4, $\delta(n)$ must also decay slowly enough for the bound of Step 4 of Appendix F to vanish, since the constants of Lemma F.1 depend on $\|\sigma^{(\delta)}\|_{W^{3,\infty}}$, which for ReLU grows like δ^{-2}.
2. **Truncation and uniform integrability.** For unbounded σ (e.g., ReLU) replace it by $\sigma_M := \psi_M \circ \sigma$ at level $M \to \infty$, where $\psi_M \in C^\infty$ satisfies $\psi_M(u) = u$ for $|u| \le M$ and is constant on $(-\infty, -M-1]$ and on $[M+1, \infty)$, then propagate uniform moment bounds on the preactivations through the truncation error.
3. **Distributional Stein's identity.** For Lipschitz σ with weak derivative $\sigma' \in L^\infty$, the Gaussian integration-by-parts identity $\mathbb{E}[Z\,\phi(Z)] = \mathbb{E}[\phi'(Z)]$ holds in the distributional sense, and extends to integrals of $\sigma'(h)$ via Stein's lemma without requiring pointwise σ''. The chain-rule expansions in Lemmas 3.2 and E.3 can be replaced by their Sobolev/Stein–Malliavin counterparts [51], at the cost of additional bookkeeping. For ReLU, Lemma E.3 then needs a lower bound on the complement norms $r_{i,t}$, and Lemmas B.1 and F.1 need separate arguments, since they use L^2 bounds on $\partial_i F_i$ and bounded σ''', respectively. The bound on $\|J_t\|_2$ likewise requires a lower bound on $\min_i \|f_{i,t}\|_2$, and for unbounded σ this route must be combined with item 2.
4. **Hermite truncation.** Decompose σ into its Hermite expansion $\sigma = \sum_{j \ge 0} \sigma_j H_j$ on $L^2(\gamma)$ and truncate at finite degree N. The truncated activation $\sigma^{(\le N)}$ is a polynomial (hence smooth, but unbounded, so it must also be combined with item 2), and for Lipschitz σ the truncation error vanishes as $N \to \infty$ at rate $\sigma - \sigma^{(\le N)} = O(1/\sqrt{N})$ in $L^2(\gamma)$; since the preactivation variances $q_{i,t} := \|f_{i,t}\|_2^2$ differ from 1 and vary, the expansion must be taken in $L^2(\mathcal{N}(0, \bar{q}))$ with $\bar{q} \ge \max_{i,t} q_{i,t}$, which calls for uniform upper and lower bounds on $q_{i,t}$.

Carrying out any of these routes for ReLU, with full quantitative tracking through every proof step, falls outside the scope of the present paper. The agreement observed for ReLU (Figure 2) suggests that the equivalence of Theorem 3.4 extends, and we leave a complete extension as an open direction.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions listed in the abstract and introduction correspond to the formal statements in Section 3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Assumptions **A1**–**A6** are stated explicitly in Section 2.1; the main text covers the single-index case, with the multi-index extension deferred to the appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All formal statements state their assumptions; complete proofs are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Simulations use synthetic Gaussian inputs and the SGD updates specified in Section 2; hyperparameters are reported in the figure captions.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code reproducing all figures is available at `https://github.com/KU-MLIP/Decomposing-SGD-Dynamics-in-Neural-Networks`; data are synthetic and generated from the model in Section 2.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The optimizer, learning rates, regularization, dimensions, time horizon, and choices of g and σ are reported in the figure captions; initialization follows Assumption **A5**.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Figures illustrate qualitative agreement with the asymptotic theoretical predictions; error bars are not reported.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Compute resources are not documented; the simulations are small-scale and run on a standard CPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Theoretical work with synthetic-data simulations; no human subjects, scraped data, or released models.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Foundational theoretical research on SGD dynamics with no direct societal impact.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No models or datasets with misuse risk are released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: No external datasets, models, or code packages are used.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code (see the answer on open access to data and code) comes with a README describing how to reproduce each figure.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLM-based assistants were used to support writing and to assist with checking mathematical derivations and proofs. All content was reviewed by the authors, who take full responsibility for the paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.