High-Dimensional Learning Dynamics of Attention-Indexed Models

Yizhou Xu
Information, Learning & Physics Laboratory
Statistical Physics of Computation Laboratory
EPFL, Switzerland

Margarita Sagitova
Statistical Physics of Computation Laboratory
EPFL, Switzerland

Lenka Zdeborová
Statistical Physics of Computation Laboratory
EPFL, Switzerland

Florent Krzakala
Information, Learning & Physics Laboratory
EPFL, Switzerland

Abstract

Attention mechanisms are central to modern foundation models, yet their training dynamics remain poorly understood, especially when the attention matrices have extensive rank. In this work, we study attention-indexed models, a broad framework that can represent multi-layer and multi-head attention architectures. First, we show that, in a suitable high-dimensional limit, the population-loss landscape is characterized by a finite set of order parameters. In contrast, online stochastic gradient descent (SGD) is governed by an infinite hierarchy of matrix moments, which we show can be exponentially well-approximated by a finite truncated system. Second, this framework reveals that attention parameterization itself can act as an architectural implicit bias. Direct optimization of an attention matrix $S \in \mathbb{R}^{d\times d}$ can remain trapped in an uninformative state. Tied attention ($S = WW^\top$) induces an automatic symmetry-breaking mechanism and yields weak recovery in $\Theta(d^2 \log d)$ samples. For untied attention, $S = UV^\top$, we uncover a fast-slow mechanism: the pre-activation mean first evolves on a fast timescale, while the overlaps evolve on a slower one. Weak recovery on the $\Theta(d^2 \log d)$ scale occurs when the state selected by the fast dynamics breaks the initial symmetry.

1 Introduction

Attention mechanisms [29] are the central building blocks of modern foundation models, but their training dynamics remain poorly understood. This is particularly true in the high-dimensional feature-learning regime with extensive-rank attention matrices, whose ranks diverge as the embedding dimension grows. Understanding this regime raises two related but distinct questions. First, what is the correct macroscopic description of learning when the attention matrices have extensive rank? Second, once such a description is available, how does the parameterization of an attention matrix affect whether features can be learned?

Existing theoretical work has made substantial progress on several aspects of attention. One line studies the mechanisms of in-context learning [10, 17, 23], while another analyzes training from scratch in simplified transformer models [9, 11, 30]. For feature-learning dynamics, however, most existing analyses either do not take a high-dimensional limit [12, 16, 24, 25, 30] or focus on finite-rank attention matrices, corresponding to sequence-indexed models [1, 6–8, 13, 27]. Such low-rank descriptions do not capture the extensive-rank regime naturally associated with attention matrices in practice, whose rank scales with the embedding dimension.

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

In this work, we study this regime through attention-indexed models, building on the formulation introduced in [4, 5]. The model class describes losses depending on collections of quadratic token interaction $\frac{1}{\sqrt{d}} x_i^T S_{ij} x_k$ with the attention matrix $S_{ij} \in \mathbb{R}^{d\times d}$ of extensive rank. The formulation is broad: multi-layer and multi-head attention-only architectures can be expressed in terms of finitely many such quadratic forms; we give the construction in Section 2 and Appendix B. Previous analyses of attention-indexed models focused on static Bayes-optimal or empirical-risk-minimization [4, 5], with tractable analysis restricted to a single-layer, single-head tied setting. Dynamical analyses of extensive-rank matrices have so far been available mainly for substantially simpler quadratic models [14, 15, 31].

Our analysis yields two primary theoretical contributions:

- A high-dimensional dynamical framework for extensive-rank attention. We formulate a broad class of attention-indexed models and show that their population loss admits a finite-dimensional characterization. For the tied and untied parameterizations, we derive high-dimensional limits for online SGD. The resulting dynamics involve an infinite hierarchy of matrix moments, for which we establish well-posedness, convergence, and exponentially accurate truncations.
- Tied attention induces automatic symmetry breaking and achieves weak recovery in $\Theta(d^2 \log d)$ samples. Untied attention learns through a fast–slow mechanism, which leads to a weak-recovery dichotomy: recovery on the $\Theta(d^2 \log d)$ scale occurs when the fast process breaks the relevant symmetry, whereas the dynamics stalls on this scale when it does not.

The remainder of the paper is organized as follows. Section 2 introduces attention-indexed models and establishes the finite-dimensional characterization of their population-loss landscape. Section 3 studies tied attention, derives its infinite-dimensional macroscopic dynamics and truncation theory, and identifies the symmetry-breaking mechanism responsible for weak recovery. Section 4 turns to untied attention and derives its two-timescale dynamics and associated weak-recovery dichotomy. The appendices contain the proofs, additional examples of attention-indexed models, and the analysis of direct optimization over the attention matrix S.

2 Attention-indexed models and their population loss landscape

The first goal of this paper is to identify a macroscopic description that remains tractable when the ranks of the attention matrices diverge with the embedding dimension. Let $L = \Theta(1)$ represent the sequence length (number of tokens), and $K = \Theta(1)$ denote the number of indices, such as attention heads or layers ($K = 2$ in [4] with one student and one teacher index). Let the data vectors $x_1, \dots, x_L \in \mathbb{R}^d$ be jointly Gaussian with zero mean and covariance given by $\mathbb{E}[x_{ia} x_{jb}] = \mathcal{C}_{ij}\delta_{ab}$, where $C \in \mathbb{R}^{L\times L}$ is a PSD matrix. Let $S_{ij}^k \in \mathbb{R}^{d\times d}$ for $i, j = 1, \dots, L$ and $k = 1, \dots, K$ be a set of deterministic weight matrices. The assumption $\mathbb{E}[x_{ia} x_{jb}] = \mathcal{C}_{ij}\delta_{ab}$ can be directly generalized when we can absorb the covariance structure into the weight matrix S.

We consider losses of the form

$$\mathcal{L}\left(\left\{\frac{x_i^T S_{ij}^k x_j}{\sqrt{d}}\right\}_{i,j,k=1}^{L,L,K}\right).$$

under the following assumption.

Assumption 1. *In the limit as $d \to \infty$ with $K, L = \Theta(1)$, the following conditions hold for all i, j, k, i', j', k':*

1. *The matrices satisfy* $\lim_{d\to\infty} \frac{||S_{ij}^k||_{op}}{||S_{ij}^k||_F} = 0$.
2. *The scaled traces have finite limits:* $\lim_{d\to\infty} \frac{1}{\sqrt{d}} \mathrm{Tr}(S_{ij}^k) = \mu_{ijk}$, $\lim_{d\to\infty} \frac{1}{d} \mathrm{Tr}(S_{ij}^k (S_{i'j'}^{k'})^T) = \Omega_{ijk,i'j'k'}$ *and* $\lim_{d\to\infty} \frac{1}{d} \mathrm{Tr}(S_{ij}^k S_{i'j'}^{k'}) = \Psi_{ijk,i'j'k'}$.
3. *The loss $\mathcal{L} : \mathbb{R}^{L\times L\times K} \to \mathbb{R}$ is continuous with at most polynomial growth.*

Remark. *While Assumptions 1.2 and 1.3 are primarily regularity conditions, Assumption 1.1 identifies the extensive-rank regime of interest here. We argue that this assumption is highly relevant in practice. In standard multi-head attention, for example, the rank of an individual attention matrix can grow with the embedding dimension when the number of heads remains fixed. This marks a critical difference from existing literature on low-rank attention (sequence-indexed models) [1, 7, 8, 27].*

In this section, we do not impose any structural assumptions on the attention matrices S_{ij}^k. We will specifically analyze both tied and untied attention mechanisms in subsequent sections. Remarkably, the attention-indexed model inherently covers standard multi-layer, multi-head attention networks (i.e., attention-only transformers).

Example (Multi-layer multi-head attention)**.** *Consider an M-layer, H-head attention network:*

$$X_m = X_{m-1} + \sum_{h=1}^{H} \textit{softmax}\left(A_m^{(h)}(X_{m-1})\right) X_{m-1} V_{m-1}^{(h)} \in \mathbb{R}^{L\times d}, \quad m = 1, \dots, M \tag{1}$$

where the pre-softmax attention matrix is given by:

$$A_m^{(h)}(X_{m-1}) = \frac{1}{\sqrt{d}} X_{m-1} K_{m-1}^{(h)} (Q_{m-1}^{(h)})^T X_{m-1}^T \in \mathbb{R}^{L\times L}. \tag{2}$$

The matrices $K_m^{(h)}, Q_m^{(h)}$, and $V_m^{(h)}$ represent the key, query, and value weights of the $m-$th layer, respectively. The final output is $y = \sigma\left(\frac{1}{\sqrt{d}} X_M K_M Q_M^T X_M^T\right)$ for some activation function σ.

Consider a teacher-student setup with input data $X_0 \sim \mathcal{N}(0, \mathcal{C} \otimes I_d)$ and a loss function $\mathcal{L}(y, \hat{y})$, where both teacher and student networks employ multi-layer multi-head architecture. This setup naturally falls within our general framework (see Appendix B for detailed derivations). Notably, we do not require the teacher weights to be Gaussian, rendering the existence of such a teacher model a relatively plausible assumption.

In Appendix B, we introduce several additional models that fit this framework.

Although the matrices S_{ij}^k contain $\mathcal{O}(d^2)$ degrees of freedom, the loss does not retain this full complexity in the high-dimensional limit. The reason is that, under the extensive-rank condition, the finite collection of quadratic forms entering the loss becomes jointly Gaussian. Its limiting distribution is therefore completely specified by the first- and second-order trace statistics in Assumption 1. The following theorem makes this reduction precise.

Theorem 1. *Under Assumption 1, the joint distribution of the quadratic forms converges weakly to a multivariate Gaussian as follows:*

$$\lim_{d\to\infty} \mathbb{E}_x \left[\mathcal{L}\left(\left\{ \frac{x_i^T S_{ij}^k x_j}{\sqrt{d}} \right\}_{i,j,k=1}^{L,L,K} \right) \right] = \mathbb{E}\left[\mathcal{L}(\{G_{ijk}\}_{i,j,k=1}^{L,L,K}) \right], \tag{3}$$

where $\{G_{ijk}\}_{i,j,k=1}^{L,L,K}$ is a multivariate Gaussian variable whose mean and covariance are given by:

$$\mathbb{E}[G_{ijk}] = \mathcal{C}_{ij}\mu_{ijk}, \ \textit{Cov}(G_{ijk}, G_{i'j'k'}) = \mathcal{C}_{ii'}\mathcal{C}_{jj'}\Omega_{ijk,i'j'k'} + \mathcal{C}_{ij'}\mathcal{C}_{ji'}\Psi_{ijk,i'j'k'}. \tag{4}$$

Theorem 1 gives the basic reduction underlying the rest of the paper. In the high-dimensional limit, the population loss depends only on the finite collection of order parameters $\{\mu, \Omega, \Psi\}$. Thus, although the parameter space grows with d, the limiting loss landscape is finite-dimensional. The Gaussian limit in Theorem 1 can be viewed as a consequence of the fourth-moment theorem for Wiener chaos [20]; for completeness, we provide a direct proof in Appendix C.1. A centered version needed for tied attention is given in Appendix C.2. Appendix G provides numerics beyond Gaussian data.

Theorem 1 characterizes the loss at any fixed state. It also implies a stronger optimization-level statement: subject to an appropriate realizability condition, optimizing over the high-dimensional matrices is asymptotically equivalent to optimizing over their finite set of limiting order parameters.

Corollary 1. *Let $\mathcal{K} \subseteq \{1, 2, \dots, K\}$ denote the index set of the learnable weights, while the remaining weights $\{S_{ij}^k\}_{k\notin\mathcal{K}}$ are fixed and satisfy the condition $\lim_{d\to\infty} \|S_{ij}^k\|_{op}/\|S_{ij}^k\|_F = 0$. Define the order parameter set Q as $Q := \left\{ \left\{ \frac{1}{\sqrt{d}} \mathrm{Tr}[S_{ij}^k] \right\}_{i,j,k}, \left\{ \frac{1}{d} \mathrm{Tr}[S_{ij}^k S_{i'j'}^{k'}], \frac{1}{d} \mathrm{Tr}[S_{ij}^k (S_{i'j'}^{k'})^T] \right\}_{i,j,k,i',j',k'} \right\}$. Let $\mathcal{Q}$ denote the feasible domain of Q realizable by matrix sequences satisfying Assumption 1 in the limit $d \to \infty$, subject to the constraints imposed by the fixed weights $\{S_{ij}^k\}_{k\notin\mathcal{K}}$. Let $\mathcal{R}(Q) := \mathbb{E}_G[\mathcal{L}(\{G_{ijk}\}_{i,j,k=1}^{L,L,K})]$ denote the expected population loss in Theorem 1 specified by Q and the covariance $\mathcal{C}$. Under the regularity assumption specified in Appendix C.3,*

$$\lim_{d\to\infty} \inf_{\{S_{ij}^k\}_{k\in\mathcal{K}}} \mathbb{E}_x \left[\mathcal{L}\left(\left\{ \frac{x_i^T S_{ij}^k x_j}{\sqrt{d}} \right\}_{i,j,k=1}^{L,L,K} \right) \right] = \inf_{Q\in\mathcal{Q}} \mathcal{R}(Q). \tag{5}$$

Corollary 1, proven in Appendix C.3, makes the finite-dimensional nature of the limiting landscape explicit: The optimal asymptotic population loss is determined by an optimization over the finite-dimensional feasible set $\mathcal{Q}$.

This does not imply that training can be described by gradient descent directly on Q. Whether the order parameters form a closed dynamical system depends on how the attention matrices are parameterized. Direct optimization over S does close at the level of finitely many first- and second-order quantities (Appendix E), whereas the tied and untied factorizations studied in Sections 3 and 4 generate additional matrix moments. This distinction has important consequences for weak recovery.

3 Online SGD of tied attention

We now turn to a question that cannot be answered from the loss landscape alone: does the parameterization of the attention matrix change the way SGD learns features? In this section, we first derive the macroscopic dynamics of tied attention and show that, despite the finite-dimensional population loss, the factorized flow requires an infinite hierarchy of matrix moments. We then use these dynamics to characterize weak recovery and contrast it with direct optimization over S.

3.1 Macroscopic dynamics

For tied attention, the trace of $S_k = W_k W_k^\top$ is typically of order d. The pre-activation $d^{-1/2} x_i^\top S_k x_j$ therefore contains a diverging contribution. Following the centered formulation of Appendix C.2, we consider the regularized loss

$$\mathcal{L}\left(\left\{\frac{\mathrm{Tr}[W_k W_k^T (x_j x_i^T - \mathcal{C}_{ij} I_d)]}{\sqrt{d}}\right\}_{i,j,k=1}^{L,L,K}\right) + \frac{\gamma}{2d}\sum_{k=1}^{K} ||W_k||_F^2,$$

where the input data $x_1, \dots, x_L \sim \mathcal{N}(0, \mathcal{C} \otimes I_d)$ and $\gamma = \Theta(1) > 0$ denotes the weight decay strength. For $k = 1, \dots, K$, let $W_k \in \mathbb{R}^{d \times d_k}$ and $S_k = W_k W_k^\top \succeq 0$. As before, let $\mathcal{K} \subseteq \{1, \dots, K\}$ denote the trainable indices. For $k \in \mathcal{K}$, W_k evolves during training, whereas for $k \notin \mathcal{K}$ it remains fixed and may, for example, represent a teacher matrix.

At iteration n, online SGD uses a fresh sample $x^{(n)} \sim \mathcal{N}(0, C \otimes I_d)$ and updates, for $k \in \mathcal{K}$,

$$W_k^{(n+1)} = W_k^{(n)} - \alpha_d \nabla_{W_k} \mathcal{L}(G^{(n)}) - \frac{\alpha_d \gamma}{d} W_k^{(n)}, \tag{6}$$

where we denote $G_{ijk}^{(n)} := \frac{1}{\sqrt{d}} \mathrm{Tr}[W_k W_k^T (x_j^{(n)} (x_i^{(n)})^T - \mathcal{C}_{ij} I_d)]$ and α_d is the step size. Define $S_k^{(n)} = W_k^{(n)} (W_k^{(n)})^\top$ and, for any multi-index $\alpha = (k_1, \dots, k_w)$, $\mu_\alpha^{(n)} = \frac{1}{d} \mathrm{Tr}\left[S_{k_1}^{(n)} S_{k_2}^{(n)} \cdots S_{k_w}^{(n)}\right]$. In particular, $q_{kl} = \mu_{(k,l)}$ denotes the second-order moment matrix. We introduce the continuous time $t = n\tau_d$, $\tau_d = \frac{4\alpha_d}{d}$ and let $\tilde{\mu}^{(d)}(t)$ denote the piecewise-constant interpolation of these moments:

$$\tilde{\mu}^{(d)}(t) = \mu^{(n)}, \quad \text{for } t \in [n\tau_d, (n+1)\tau_d). \tag{7}$$

Define the potential as the limiting population loss in Theorem 1: $\Phi(q) := \mathbb{E}_G[\mathcal{L}(G)]$ for $q \in \mathbb{S}_+^K$, where $G \in \mathbb{R}^{L \times L \times K}$ is a zero-mean Gaussian tensor with covariance $\mathrm{Cov}\left((G)_{ijk}, (G)_{i'j'l}\right) = (\mathcal{C}_{ii'}\mathcal{C}_{jj'} + \mathcal{C}_{ij'}\mathcal{C}_{ji'}) q_{kl}$. We define its matrix derivative as $\nabla\Phi(q)$[1] for $q \succeq 0$.

Assumption 2. *1. The loss function $\mathcal{L} : \mathbb{R}^{L \times L \times K} \to \mathbb{R}$ is four times continuously differentiable. $\mathcal{L}$ and its derivatives up to the fourth order have at most polynomial growth.*

2. At initialization, the weights $W_k^{(0)}$ satisfy $\|W_k^{(0)}\|_F = \Theta(\sqrt{d})$. Additionally, there exists a constant $C_0 > 0$ independent of d such that $\sup_{1 \le k \le K} \|W_k^{(0)}\|_{op} \le C_0$ almost surely.

3. $\lim_{d \to \infty} |\tilde{\mu}_\alpha^{(d)}(0) - \bar{\mu}_\alpha(0)| = 0$ in probability for every multi-index α.

4. The regularization is sufficiently large, i.e., $\gamma > \gamma_*(C_0, \mathcal{C}, \mathcal{L}, K, L)$ for some γ_* independent of d.

Remark. *Assumption 2.4 is used only to guarantee uniform spectral control and prevent finite-time blow-up. Equivalently, one may impose an a priori uniform bound on the operator norms of the weights; numerically, our trajectories remain bounded even without weight decay.*

[1] More rigorously speaking, it should be defined as $\lim_{\epsilon \to 0^+} \nabla\Phi(q + \epsilon I)$.

While the potential Φ depends only on the finite-dimensional matrix q, the evolution of q is not closed: second moments depend on third moments, third moments on fourth moments, and so on. The tied factorization therefore converts a finite-dimensional loss landscape into an infinite-dimensional dynamical hierarchy. The following theorem shows that this hierarchy nevertheless has a deterministic and well-posed high-dimensional limit.

Theorem 2. *Under Assumption 2, the online SGD trajectory is governed by an infinite-dimensional ODE driven by the gradients of Φ:*

$$\frac{d\bar{\mu}_{(k_1,\ldots,k_w)}}{dt} = -\sum_{\substack{i=1\\ k_i\in\mathcal{K}}}^{w}\sum_{l}[\nabla\Phi(\bar{q})]_{k_i l}\left(\bar{\mu}_{(k_1,\ldots,k_{i-1},l,k_i,\ldots,k_w)} + \bar{\mu}_{(k_1,\ldots,k_i,l,k_{i+1},\ldots,k_w)}\right) - \frac{\gamma}{2}|\alpha|_{\mathcal{K}}\bar{\mu}_{(k_1,\ldots,k_w)}, \tag{8}$$

where $|\alpha|_{\mathcal{K}} \le w$ is the number of trainable indices in the multi-index α of length w. (8) *has a unique solution* [2] *and the empirical trajectory uniformly converges to it in probability.*

Specifically, if the width scales such that $\max_{1\le k\le K} d_k \le d^{c_w}$ for some fixed $c_w < +\infty$ and the learning rate scales such that $\alpha_d \le c_0(d\log d)^{-1}$ and $\alpha_d = \Omega(d^{-\iota})$ for some constants $c_0 > 0$ small enough and $\iota > 0$, then the empirical SGD trajectory satisfies:

$$\lim_{d\to\infty}\mathbb{P}\left(\sup_{t\in[0,T]}\left|\tilde{\mu}_{\alpha}^{(d)}(t) - \bar{\mu}_{\alpha}(t)\right| > \epsilon\right) = 0,\ \forall\epsilon > 0 \tag{9}$$

for every multi-index α and for any fixed $T > 0$.

Theorem 2 gives the macroscopic dynamical counterpart of the static reduction in Section 2. The same finite-dimensional potential $\Phi(q)$ determines the coefficients of the flow, but it does not determine a closed ODE for q. Instead, the geometry of the factorization $S = WW^\top$ propagates information through the entire hierarchy of joint moments. This infinite-dimensionality is not merely an artifact of our choice. Appendix D.3 gives an exactly solvable quadratic example. In that example, no finite-dimensional ODE with a continuous initialization map can reproduce all possible trajectories.

The proof, given in Appendix D.1, combines four ingredients: local well-posedness in a scale of weighted moment spaces, uniform spectral bounds obtained through a stopping-time argument and matrix Freedman inequalities, tightness and convergence of the empirical moment process, and global well-posedness of the resulting infinite-dimensional trajectory.

The infinite hierarchy is required for an exact description, but this does not make the theory numerically intractable. Higher-order moments can be truncated with exponentially small error. For $M \ge 2$, we define the degree-M truncated trajectory $\mu^{(M)}(t)$ by

$$\frac{d}{dt}\mu_{\alpha}^{(M)}(t) = [V(\mu^{(M)}(t))]_{\alpha}, \qquad |\alpha| \le M, \tag{10}$$

with the initial condition $\mu_{\alpha}^{(M)}(0) = \bar{\mu}_{\alpha}(0)$ for $|\alpha| \le M$ and the boundary condition $\mu_{\alpha}^{(M)}(t) \equiv 0$ for $|\alpha| > M$. Here V denotes the right side of (8)[3]. The following corollary guarantees that $\mu^{(M)}(t)$ can approximate the true dynamics $\bar{\mu}(t)$ to arbitrary precision.

Corollary 2. *Under the conditions of Theorem 2, there exist constants $s_0 > C_*^2$ and $c < \infty$ independent of M, α and T such that, for all sufficiently large M, the truncated ODE* (10) *has a unique solution and for every multi-index α, $\sup_{t\in[0,T]}\left|\mu_{\alpha}^{(M)}(t) - \bar{\mu}_{\alpha}(t)\right| \le c\left(\frac{C_*^2}{s_0}\right)^{M+1} s_0^{|\alpha|}$.*

Corollary 2 suggests that while the limiting trajectory is intrinsically infinite-dimensional, any fixed collection of low-order observables can be approximated to arbitrary accuracy by increasing the truncation degree. In the exactly solvable quadratic setting (quadratic networks [3, 14, 15]) of Appendix D.3, these finite-order systems coincide with the Taylor expansion of the exact matrix flow.

[2] A solution of (8) on $[0,T]$ is a trajectory μ for which there exist $0 < s < s' < \infty$ such that $\mu \in C([0,T]; X_s^0)$, and $\mu(t) = \mu(0) + \int_0^t V(\mu(r))\,dr$ holds as an identity in $X_{s'}^0$ for every $t \in [0,T]$. Here V refers to the right side of (8) and $X_s^0 := \left\{\mu : \lim_{w\to\infty}\sup_{|\alpha|=w}\frac{|\mu_\alpha|}{s^w} = 0\right\}$.

[3] More specifically, V refers to the extended vector field defined in (58) because the truncated moment sequence might not satisfy $q \succeq 0$.

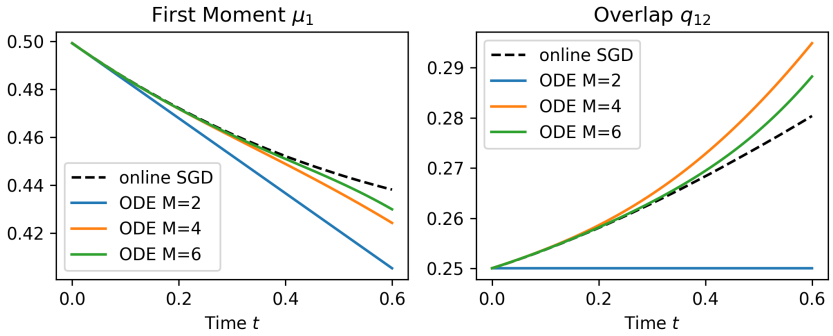


Figure 1: Finite truncations accurately approximate the infinite moment dynamics. Comparison of online SGD with degree-M truncations of (10) for $L = 2$, softmax activation, and MSE loss. Increasing M improves the agreement with the empirical trajectory. We use $d = 400$, learning rate 0.01, batch size 512, and no weight decay.

Corollary 2 is proven in Appendix D.2. As a side remark, Assumption 2.4 is also used in the proof to obtain an error estimate uniform in time. Figure 1 verifies the same phenomenon beyond the exactly solvable setting. For $L = 2$, softmax activation, and MSE loss, increasing the truncation degree systematically improves agreement between the deterministic theory and online SGD.

3.2 Weak recovery and its sample complexity

We now use the macroscopic dynamics to address the main learning question of this section: can the tied factorization escape an initially uninformative state, and does its behavior differ from direct optimization over S? The relevant notion of information is not the raw overlap $\frac{1}{d}\operatorname{Tr}(S_1 S_2)$. Because tied matrices have a nonzero trace, this quantity can be $\mathcal{O}(1)$ even when the student contains no information about the teacher.

For clarity, consider one trainable student $S_1 = W_1 W_1^T$ and one fixed teacher S_2. Define $t_k^{(d)} := \frac{1}{d}\operatorname{Tr}[S_k]$, $\mathring{S}_k := S_k - t_k^{(d)} I_d$ and the structural overlaps $p_{kl}^{(d)} := \frac{1}{d}\operatorname{Tr}[\mathring{S}_k \mathring{S}_l] = q_{kl}^{(d)} - t_k^{(d)} t_l^{(d)}$. Their deterministic limits are denoted by $\bar{t}_k, \bar{q}_{kl}$, and $\bar{p}_{kl}$. Thus $\bar{p}_{12}$, rather than $\bar{q}_{12}$, measures alignment between the student and teacher. We impose the following uninformative initialization.

Assumption 3. *As $d \to \infty$, the initial student $S_1(0)$ and the teacher S_2 satisfy $\bar{p}_{12}(0) = 0$, $\lim_{d\to\infty} \frac{1}{d}\operatorname{Tr}[\mathring{S}_1(0)^2 \mathring{S}_2] = 0$ and $\lim_{d\to\infty} \frac{1}{d}\operatorname{Tr}[\mathring{S}_1(0)\mathring{S}_2^2] = 0$. Moreover, we assume that $\bar{t}_1(0) > 0$ and $\bar{p}_{22} > 0$.*

Assumption 3 places the student on an uninformative manifold. An independently initialized Gaussian $W_1(0)$ satisfies these conditions. For a fixed recovery threshold $c > 0$, define the weak-recovery sample complexity by $N_{\mathrm{wr}}^{(d)}(c) := \inf\left\{n \geq 0 : |p_{12}^{(d,n)}| \geq c\right\}$, where $p_{12}^{(d,n)}$ denotes the structural overlap after n steps of online SGD (6).

Corollary 3. *Under Assumption 3 and the conditions of Theorem 2, suppose additionally that*

$$[\nabla\Phi(\bar{q}(0))]_{12} \neq 0, \tag{11}$$

Then there exist constants $c > 0$ and $0 < T_- < T_+ < \infty$ independent of d, such that $\lim_{d\to\infty} \mathbb{P}\left(\left\lfloor \frac{T_- d}{4\alpha_d} \right\rfloor < N_{\mathrm{wr}}^{(d)}(c) \leq \left\lceil \frac{T_+ d}{4\alpha_d} \right\rceil\right) = 1$. In particular, choosing the largest learning-rate scaling $\alpha_d = \Theta\left(\frac{1}{d \log d}\right)$ covered by Theorem 2 gives the weak recovery sample complexity $N_{\mathrm{wr}}^{(d)}(c) = \Theta_{\mathbb{P}}(d^2 \log d)$.

Remark. *The condition (11) is mild in a broad class of teacher-student losses. To see this, consider the MSE loss $\mathcal{L}(G_1, G_2) = \frac{1}{2}\|F(G_1) - F(G_2)\|^2$ with the same activation F for student and teacher. Then $[\nabla\Phi(\bar{q})]_{12} = -\frac{1}{2}\mathbb{E}\left[\left\langle \nabla F(G_1)B^{1/2}, \nabla F(G_2)B^{1/2}\right\rangle_F\right]$, where $B_{(ij),(i'j')} = \mathcal{C}_{ii'}\mathcal{C}_{jj'} + \mathcal{C}_{ij'}\mathcal{C}_{ji'}$. Suppose additionally that the initialization scale is chosen such that $q_{11}(0) = q_{22} > 0$. A Gaussian Hermite expansion then gives $-[\nabla\Phi(\bar{q}(0))]_{12} = \frac{1}{2}\sum_\nu \rho^{|\nu|}\nu!\|H_\nu\|_F^2$, where $\{H_\nu\}$ are the Hermite coefficients of $\nabla F(\sqrt{q_{22}}B^{1/2}z)B^{1/2}$ and $\rho = \frac{\bar{t}_1(0)\bar{t}_2}{\sqrt{\bar{q}_{11}(0)\bar{q}_{22}}} > 0$. Hence $[\nabla\Phi(\bar{q}(0))]_{12} < 0$ whenever $\nabla F(\sqrt{q_{22}}B^{1/2}Z)B^{1/2}$ is not constantly zero.*

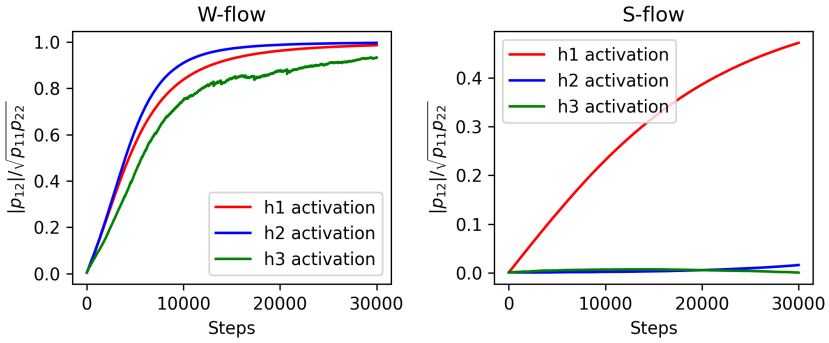


Figure 2: The parameterization of the attention matrix qualitatively changes weak recovery. We compare online SGD on the tied factor W, with $S = WW^\top$ (**Left**), and direct online SGD on S (**Right**), starting from uninformative initializations in the same teacher–student setting. For the linear activation both parameterizations recover the teacher on the displayed scale, whereas for the higher-order activations the direct S-flow remains close to the uninformative state while the tied W-flow develops substantial overlap. We use $d = 400$, learning rate 0.005, batch size 512, and no weight decay. The activations are chosen from: $h_1(x) = x$, $h_2(x) = \frac{x^2-1}{\sqrt{2}}$ and $h_3(x) = \frac{x^3-3x}{\sqrt{6}}$, respectively ($L = 1$). The input data is $\mathcal{N}(0, I_d)$.

The significance of Corollary 3 is not only the $d^2 \log d$ sample scale, but the mechanism that produces it. The student begins with zero structural overlap with the teacher, yet the tied factorization makes this uninformative state non-invariant. This behavior should be contrasted with direct optimization over S. As shown in Appendix E, the direct S-flow closes at the level of finitely many order parameters. More importantly, when the relevant gradients vanish, weak recovery does not occur on the $d^2 \log d$ sample scale. The tied factorization therefore acts as an architectural implicit bias. Figure 2 illustrates this distinction. For a linear activation, both parameterizations can develop overlap. When we do online SGD directly over S on higher-order activations (h_2, h_3), however, it remains uninformative on the displayed sample scale, whereas the tied flow escapes and learns the teacher.

4 Online SGD of untied attention

The symmetry breaking described in Section 3 relies crucially on the PSD geometry $S = WW^\top$. This raises the natural next question: what remains of the mechanism when S is not PSD? This section answers this question for the untied parameterization $S = UV^T$.

The mechanism is qualitatively different from the tied case. For untied attention, the pre-activation has a nontrivial mean $m_k = \frac{1}{\sqrt{d}} \operatorname{Tr}(U_k V_k^\top)$, and this mean evolves faster than the matrix moments. The remainder of this section makes this fast–slow picture rigorous. We first derive the two-timescale macroscopic limit and then use it to characterize a corresponding dichotomy for weak recovery.

4.1 Macroscopic dynamics

For each index $k = 1, \dots, K$, let $U_k, V_k \in \mathbb{R}^{d \times d_k}$ and define the associated matrices $M_k := U_k V_k^\top$, $A_k := U_k U_k^\top$ and $B_k := V_k V_k^\top$. We consider the regularized loss

$$\mathcal{L}\left(\left\{ \frac{1}{\sqrt{d}} x_i^T U_k V_k^T x_j \right\}_{i,j,k=1}^{L,L,K} \right) + \frac{\gamma}{2d} \sum_{k \in \mathcal{K}} \left(\|U_k\|_F^2 + \|V_k\|_F^2 \right), \tag{12}$$

where $x_1, \dots, x_L \sim \mathcal{N}(0, \mathcal{C} \otimes I_d)$, and $\mathcal{K} \subseteq \{1, \dots, K\}$ denotes the set of trainable indices. At iteration n, a fresh sample $x^{(n)} \sim \mathcal{N}(0, \mathcal{C} \otimes I_d)$ is drawn, and for $k \in \mathcal{K}$ the online SGD updates are

$$U_k^{(n+1)} = U_k^{(n)} - \alpha_d \nabla_{U_k} \mathcal{L}(G^{(n)}) - \frac{\alpha_d \gamma}{d} U_k^{(n)}, \; V_k^{(n+1)} = V_k^{(n)} - \alpha_d \nabla_{V_k} \mathcal{L}(G^{(n)}) - \frac{\alpha_d \gamma}{d} V_k^{(n)}, \tag{13}$$

where $G_{ijk}^{(n)} := \frac{1}{\sqrt{d}} (x_i^{(n)})^T U_k^{(n)} (V_k^{(n)})^T x_j^{(n)}$ and α_d is the learning rate.

Order parameters. Unlike the centered tied model of Section 3, the untied pre-activations can have an $\mathcal{O}(1)$ limiting mean. We therefore need to track separately $m_k^{(d)} := \frac{1}{\sqrt{d}} \operatorname{Tr}(M_k)$.

The remaining macroscopic state is described by normalized joint moments of the matrices $M_k, M_k^\top, A_k, B_k$. Introduce the alphabet $\mathscr{Z} := \{M_k, M_k^T, A_k, B_k : 1 \le k \le K\}$ and for a word $\alpha = (Z_1, \dots, Z_w) \in \mathscr{Z}^w$, define $\mu_\alpha^{(d)} := \frac{1}{d}\mathrm{Tr}[Z_1 \cdots Z_w]$. In particular, we define the traces $t_k^A := \mu_{(A_k)}$, $t_k^B := \mu_{(B_k)}$ and the overlaps $q_{kl}^{(1)} := \mu_{(M_k, M_l^T)}$, $q_{kl}^{(2)} := \mu_{(M_k, M_l)}$.

The potential is defined as the limiting population loss in Theorem 1: $\Phi(m, q^{(1)}, q^{(2)}) := \mathbb{E}\big[\mathcal{L}(G)\big]$, where $G \in \mathbb{R}^{L\times L\times K}$ is Gaussian with $\mathbb{E}[G_{ijk}] = \mathcal{C}_{ij} m_k$ and $\mathrm{Cov}(G_{ijk}, G_{i'j'l}) = \mathcal{C}_{ii'}\mathcal{C}_{jj'} q_{kl}^{(1)} + \mathcal{C}_{ij'}\mathcal{C}_{ji'} q_{kl}^{(2)}$. We denote by $\nabla_m \Phi$ and $\nabla_{q^{(r)}}\Phi$, $r = 1, 2$, the vector and matrix gradients of the potential, respectively.

For later use, let $M_l^{(1)} := M_l$ and $M_l^{(2)} := M_l^T$ for $\bar{1} := 2, \quad \bar{2} := 1$. Define

$$\psi_k(m, \mu) := 2\sum_{l=1}^{K}\sum_{r=1}^{2}\left[\nabla_{q^{(r)}}\Phi(m, q^{(1)}, q^{(2)})\right]_{kl}\left[\mu_{(M_l^{(r)}, B_k)} + \mu_{(A_k, M_l^{(r)})}\right]. \tag{14}$$

The new point is that m and μ evolve on different scales. A single SGD step changes m at order α_d, whereas the moments change only at order α_d/d. This produces two timescales below.

Fast learning phase. On the fast timescale $t = n\alpha_d$, the normalized moments remain asymptotically frozen, while the mean vector evolves at order one. Denoting the limiting fast trajectory by $\widetilde{m}(t)$, the trainable components satisfy

$$\frac{d\tilde{m}_k}{dt} = -\left(\bar{t}_k^A(0) + \bar{t}_k^B(0)\right)\frac{\partial\Phi}{\partial m_k}\left(\tilde{m}(t), \bar{q}^{(1)}(0), \bar{q}^{(2)}(0)\right), \qquad k \in \mathcal{K}. \tag{15}$$

The non-trainable components remain fixed. Thus, to leading order, the fast learning phase performs gradient flow in the mean variables while seeing the covariance structure as frozen.

Slow learning phase. We now introduce the slow timescale $\tau = \frac{t}{d} = \frac{n\alpha_d}{d}$. Denote $m_{\mathcal{K}}^\star = m_{\mathcal{K}}^\star(q^{(1)}, q^{(2)})$ satisfying $\nabla_{m_{\mathcal{K}}}\Phi\left(m^\star(q^{(1)}, q^{(2)}), q^{(1)}, q^{(2)}\right) = 0$ and $\mathcal{M}_0 := \left\{(m, \mu) : \nabla_{m_{\mathcal{K}}}\Phi(m, q^{(1)}, q^{(2)}) = 0\right\}$.After the initial boundary layer, the trainable means are asymptotically slaved as:

$$\bar{m}_{\mathcal{K}}(\tau) = m_{\mathcal{K}}^\star\left(\bar{q}^{(1)}(\tau), \bar{q}^{(2)}(\tau)\right), \qquad (\bar{m}(\tau), \bar{\mu}(\tau)) \in \mathcal{M}_0. \tag{16}$$

The small residual of the fast variables contributes during the slow timescale. This produces the effective bias

$$\bar{\beta}_k(\bar{m}, \bar{\mu}) := -\frac{\psi_k(\bar{m}, \bar{\mu})}{\bar{t}_k^A + \bar{t}_k^B}. \tag{17}$$

with ψ_k defined in (14). For a word $\alpha = (Z_1, \dots, Z_w)$, let k_i denote the index carried by the letter Z_i. For each trainable position i, define the substitution operators

$$\mathcal{T}_i^{\mathrm{bias}}(\mu) = \begin{cases} \mu_{(\dots, A_k + B_k, \dots)}, & Z_i \in \{M_k, M_k^T\}, \\ \mu_{(\dots, M_k + M_k^T, \dots)}, & Z_i \in \{A_k, B_k\}, \end{cases} \quad \mathcal{T}_{i,l,r}^{\mathrm{diff}}(\mu) = \begin{cases} \mu_{(\dots, M_l^{(r)} B_k + A_k M_l^{(r)}, \dots)}, & Z_i = M_k, \\ \mu_{(\dots, B_k M_l^{(\bar{r})} + M_l^{(\bar{r})} A_k, \dots)}, & Z_i = M_k^T, \\ \mu_{(\dots, M_l^{(r)} M_k^T + M_k M_l^{(\bar{r})}, \dots)}, & Z_i = A_k, \\ \mu_{(\dots, M_l^{(\bar{r})} M_k + M_k^T M_l^{(r)}, \dots)}, & Z_i = B_k. \end{cases} \tag{18}$$

The notation is understood linearly; for example, $\mu_{(\dots, A_k + B_k, \dots)} = \mu_{(\dots, A_k, \dots)} + \mu_{(\dots, B_k, \dots)}$.

With these operators, the slow evolution of all normalized matrix moments closes as an infinite-dimensional hierarchy

$$\frac{d\bar{\mu}_\alpha}{d\tau} = -\sum_{\substack{i=1 \\ k_i \in \mathcal{K}}}^{w} \bar{\beta}_{k_i}(\bar{m}, \bar{\mu})\mathcal{T}_i^{\mathrm{bias}}(\bar{\mu}) - 2\sum_{\substack{i=1 \\ k_i \in \mathcal{K}}}^{w}\sum_{l=1}^{K}\sum_{r=1}^{2}\frac{\partial\Phi}{\partial q_{k_i l}^{(r)}}\left(\bar{m}, \bar{q}^{(1)}, \bar{q}^{(2)}\right)\mathcal{T}_{i,l,r}^{\mathrm{diff}}(\bar{\mu}) - 2\gamma|\alpha|_{\mathcal{K}}\bar{\mu}_\alpha \tag{19}$$

for $0 \le \tau \le T$. Together with (16), this determines the limiting trajectory in the slow learning phase.

We now define the empirical interpolations. On the fast timescale, $\tilde{m}^{(d)}(t) := m^{(n)}$ for $t \in [n\alpha_d, (n+1)\alpha_d)$, while on the slow timescale, $\tilde{\mu}^{(d)}(\tau) = \mu^{(n)}$ for $\tau \in \left[\frac{n\alpha_d}{d}, \frac{(n+1)\alpha_d}{d}\right)$. The following theorem rigorously justifies the separation into a fast boundary layer and a slow infinite-dimensional flow.

Theorem 3. *Suppose that assumptions in Appendix F.1 hold for a fixed $T > 0$. Assume furthermore that $\max_{1\le k\le K} d_k \le d^{c_w}$ for some fixed $c_w < \infty$, and that the learning rate satisfies $\alpha_d \le \frac{c_0}{d \log d}$ and $\alpha_d = \Omega(d^{-\iota})$ for some constants $\iota > 0$ and sufficiently small $c_0 > 0$ independent of d. Then the online SGD trajectory exhibits the following two-timescale limit. Fix any $\epsilon > 0$. For every finite fast-time horizon $\widetilde{T} > 0$,*

$$\lim_{d\to\infty} \mathbb{P}\left(\sup_{t\in[0,\widetilde{T}]} \left\| \tilde{m}^{(d)}(t) - \tilde{m}(t) \right\|_\infty > \epsilon \right) = 0, \tag{20}$$

where $\tilde{m}$ is the unique solution of (15)*. On the slow timescale, let $(\bar{m}(\tau), \bar{\mu}(\tau))$ denote the unique solution of* (16)–(19)*. Then for every $0 < T_1 < T_2 \le T$ and every word α,*

$$\lim_{d\to\infty} \mathbb{P}\left(\sup_{\tau\in[T_1,T_2]} \left\| \tilde{m}^{(d)}(\tau d) - \bar{m}(\tau) \right\|_\infty > \epsilon \right) = 0, \tag{21}$$

$$\lim_{d\to\infty} \mathbb{P}\left(\sup_{\tau\in[0,T_2]} \left| \tilde{\mu}_\alpha^{(d)}(\tau) - \bar{\mu}_\alpha(\tau) \right| > \epsilon \right) = 0. \tag{22}$$

Theorem 3 shows that the untied factorization introduces a dynamical structure that has no analogue in the tied model. On the fast scale $t = n\alpha_d$, only the pre-activation means undergo $\mathcal{O}(1)$ evolution. On the slow scale $\tau = n\alpha_d/d$, these means have already equilibrated and follow the evolving covariance, while the matrix moments undergo their $\mathcal{O}(1)$ evolution. The proof is given in Appendix F. Its main additional difficulty relative to the tied case is the singular separation of timescales. We first establish convergence of the fast mean dynamics while the normalized moments remain frozen. We then obtain a refined $\mathcal{O}(d^{-1/2})$ control of the residual mean gradient after the boundary layer; this residual accumulates into the effective bias β_k on the slow scale. Uniform spectral bounds, tightness of the moment process, and uniqueness then yield the limiting slow trajectory. As in the tied case, the slow infinite-dimensional hierarchy admits finite-order approximations. Appendix F.3 proves an exponentially decaying truncation error.

4.2 Weak recovery and its sample complexity

We now ask what the fast–slow structure implies for feature learning. To make this mechanism explicit, consider one trainable student $M_1 = U_1 V_1^\top$ and one fixed teacher M_2. Since $\frac{1}{d}\operatorname{Tr}(M_k) = \frac{m_k^{(d)}}{\sqrt{d}} \to 0$, feature learning is naturally measured through the matrix overlaps $q_{12}^{(1,d)} := \frac{1}{d}\operatorname{Tr}[M_1 M_2^T]$ and $q_{12}^{(2,d)} := \frac{1}{d}\operatorname{Tr}[M_1 M_2]$. We write $\mathbf{q}_{12}^{(d)} := \begin{pmatrix} q_{12}^{(1,d)} \\ q_{12}^{(2,d)} \end{pmatrix}$ and $\bar{\mathbf{q}}_{12} := \begin{pmatrix} \bar{q}_{12}^{(1)} \\ \bar{q}_{12}^{(2)} \end{pmatrix}$. We impose the following uninformative initialization.

Assumption 4. *At the initial time, the matrix algebra generated by $\{M_1, M_1^T, A_1, B_1\}$ is asymptotically free from the algebra generated by $\{M_2, M_2^T, A_2, B_2\}$.*

Assumption 4 is the untied analogue of an uninformative initialization. Define the teacher covariance as $\mathsf{Q}_2 := \begin{pmatrix} \bar{q}_{22}^{(1)} & \bar{q}_{22}^{(2)} \\ \bar{q}_{22}^{(2)} & \bar{q}_{22}^{(1)} \end{pmatrix}$. For $c > 0$, define the weak-recovery sample complexity by $N_{\mathrm{wr}}^{(d)}(c) := \inf\left\{ n \ge 0 : \left\| \mathbf{q}_{12}^{(d,n)} \right\|_\infty \ge c \right\}$. Define the gradient at the slow-time initial state by $g_{12}^\star := \left(\frac{\partial \Phi}{\partial q_{12}^{(1)}}, \frac{\partial \Phi}{\partial q_{12}^{(2)}} \right)^T$ evaluated at $\left(\bar{m}^\star, \bar{q}^{(1)}(0), \bar{q}^{(2)}(0)\right)$.

Corollary 4. *Suppose that Assumption 4 and the conditions of Theorem 3 hold.*

(i) *If* $\mathsf{Q}_2 g_{12}^\star \neq 0$*, then there exist constants* $c > 0$ *and* $0 < T_- < T_+ < \infty$*, independent of* d*, such that* $\lim_{d\to\infty} \mathbb{P}\left(\left\lfloor \frac{T_- d}{\alpha_d} \right\rfloor < N_{\mathrm{wr}}^{(d)}(c) \le \left\lceil \frac{T_+ d}{\alpha_d} \right\rceil\right) = 1$*. Consequently, choosing the largest learning-rate scaling* $\alpha_d = \Theta\left(\frac{1}{d \log d}\right)$ *in Theorem 3 gives* $N_{\mathrm{wr}}^{(d)}(c) = \Theta_{\mathbb{P}}(d^2 \log d)$*.*

(ii) *Let* $\mathcal{M}_{\mathrm{free}}$ *denote the set of moment sequences for which the algebra generated by* $\{M_1, M_1^T, A_1, B_1\}$ *is free from the algebra generated by* $\{M_2, M_2^T, A_2, B_2\}$*. Suppose that* $\left[\nabla_{q^{(r)}} \Phi\left(m^\star(q), q^{(1)}, q^{(2)}\right)\right]_{12} = 0$ *for* $r = 1, 2$ *throughout* $\mathcal{M}_{\mathrm{free}}$*. Then for every* $c > 0$ *and every fixed* $T > 0$*,* $\lim_{d\to\infty} \mathbb{P}\left(N_{\mathrm{wr}}^{(d)}(c) \le \left\lfloor \frac{Td}{\alpha_d} \right\rfloor\right) = 0$*. Thus, for* $\alpha_d = \Theta((d \log d)^{-1})$*, weak recovery does not occur on the* $d^2 \log d$ *sample scale.*

Corollary 4 identifies a dichotomy. Part (i) states that if the state selected by the fast dynamics exposes a nonzero gradient, weak recovery occurs after $\Theta(d^2 \log d)$ samples. Part (ii) gives that if the relevant gradients vanish throughout the uninformative manifold, then weak recovery does not occur on the same sample scale. The contrast with tied attention is therefore sharp. In the tied case, the PSD geometry itself generates a nonzero escape direction from the original initialization. In the untied case, the fast phase first selects a new mean state, and it is the geometry of the loss at this selected state that determines whether the subsequent slow dynamics can escape.

The mechanism is particularly transparent for $L = 1$ and the MSE loss $\mathcal{L}(z_1, z_2) = \frac{1}{2}\big(\sigma(z_1) - \sigma_\star(z_2)\big)^2$. At an uninformative state $q_{12} = 0$, the student and teacher pre-activations $Z_1 \sim \mathcal{N}\left(\bar{m}_1^\star, \bar{q}_{11}^{(1)} + \bar{q}_{11}^{(2)}\right)$ and $Z_2 \sim \mathcal{N}\left(\bar{m}_2, \bar{q}_{22}^{(1)} + \bar{q}_{22}^{(2)}\right)$ are independent. Price's theorem gives $\frac{\partial \Phi}{\partial q_{12}^{(1)}} = \frac{\partial \Phi}{\partial q_{12}^{(2)}} = \mathbb{E}\left[\frac{\partial^2 \mathcal{L}}{\partial z_1 \partial z_2}\right] = -\mathbb{E}[\sigma'(Z_1)]\mathbb{E}[\sigma_*'(Z_2)]$. Hence a sufficient condition for weak recovery is $\mathbb{E}[\sigma'(Z_1)]\mathbb{E}[\sigma_*'(Z_2)] \neq 0$. The two factors play different roles. The teacher-side term $\mathbb{E}[\sigma_\star'(Z_2)]$ determines whether a first-order correlation is available. The fast dynamics cannot change this factor. By contrast, the student-side term $\mathbb{E}[\sigma'(Z_1)]$ is evaluated at the shifted mean $\bar{m}_1^\star$ selected by the fast phase. The fast learning phase can therefore repair a student-side symmetry: even if $\mathbb{E}[\sigma'(Z_1)]$ vanishes initially, it may become nonzero after the mean shifts. When this occurs and the teacher-side first-order signal is nonzero, weak recovery takes $\Theta(d/\alpha_d)$ samples. If $\bar{m}_1^*$ fails to break this symmetry, the network stalls.

A numerical illustration is presented in Figure 3, which verifies that when $\mathbb{E}_{z_2}\left[\sigma_*'(Z_2)\right] = 0$, weak recovery takes substantially more samples. Figure 3 also suggests that the weak recovery time depends on the information exponent of the activation functions, which is left for future work.

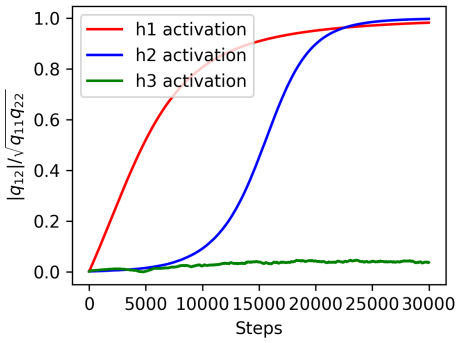


Figure 3: Weak recovery becomes delayed for higher-order activations. Evolution of the symmetric student–teacher overlap (the overlap between $M_1 + M_1^T$ and $M_2 + M_2^T$) for untied attention in a single-token teacher–student model with MSE loss. The linear activation $h_1(x) = x$ develops overlap rapidly, whereas the higher-order activations $h_2(x) = \frac{x^2-1}{\sqrt{2}}$ and $h_3(x) = \frac{x^3-3x}{\sqrt{6}}$ exhibit increasingly slow escape from the uninformative state, whose precise sample complexity is left for future work. We use $d = 800$, learning rate 0.01, batch size 4096, standard Gaussian initialization, and no weight decay. The input data is $\mathcal{N}(0, I_d)$.

5 Conclusion and future work

We develop a high-dimensional framework for attention-indexed models, where the population loss is characterized by finitely many order parameters, while online SGD is governed by an infinite moment hierarchy. We further show that attention parameterization fundamentally affects feature learning. Tied attention $S = WW^\top$ induces automatic symmetry breaking and achieves weak recovery in $\Theta(d^2 \log d)$ samples, even when direct optimization over S remains trapped. Untied attention $S = UV^\top$ instead exhibits fast–slow dynamics, where recovery depends on the state selected by the fast phase.

Future directions include extending the analysis to deep transformers, characterizing the dependence of weak recovery on information exponents, and studying finite-sample optimization. Alternative formulations, such as resolvent-based approaches, may also simplify the infinite moment hierarchy.

Acknowledgements

We acknowledge funding from the Swiss National Science Foundation grants SNSF SMArtNet (grant number 212049), OperaGOST (grant number 200021 200390), and DSGIANGO (grant number 225837). This work was supported by the Simons Collaboration on the Physics of Learning and Neural Computation via the Simons Foundation grant (#1257412 (FK) and #1257413 (LZ)). This work was also supported by the EPFL AI Center PhD Fellowship Program 2026 (YX).

References

[1] Luca Arnaboldi, Bruno Loureiro, Ludovic Stephan, Florent Krzakala, and Lenka Zdeborova. Asymptotics of sgd in sequence-single index models and single-layer attention networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[2] Jean Barbier, Francesco Camilli, Justin Ko, and Koki Okajima. Phase diagram of extensive-rank symmetric matrix denoising beyond rotational invariance. *Physical Review X*, 15(2):021085, 2025.

[3] Antoine Bodin and Nicolas Macris. Gradient flow on extensive-rank positive semi-definite matrix denoising. In *2023 IEEE Information Theory Workshop (ITW)*, pages 365–370. IEEE, 2023.

[4] Fabrizio Boncoraglio, Vittorio Erba, Emanuele Troiani, Yizhou Xu, Florent Krzakala, and Lenka Zdeborová. Single-head attention in high dimensions: A theory of generalization, weights spectra, and scaling laws. In *Proceedings of the 43rd International Conference on Machine Learning*, pages 9058–9093. PMLR, 2026.

[5] Fabrizio Boncoraglio, Emanuele Troiani, Vittorio Erba, and Lenka Zdeborova. Bayes optimal learning of attention-indexed models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[6] Hugo Cui. High-dimensional learning of narrow neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(2):023402, 2025.

[7] Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborová. A phase transition between positional and semantic learning in a solvable model of dot-product attention. *Advances in Neural Information Processing Systems*, 37:36342–36389, 2024.

[8] Odilon Duranthon, Pierre Marion, Claire Boyer, Bruno Loureiro, and Lenka Zdeborová. Statistical advantage of softmax attention: insights from single-location regression. *arXiv preprint arXiv:2509.21936*, 2025.

[9] Ruiquan Huang, Yingbin Liang, and Jing Yang. How transformers learn regular language recognition: A theoretical study on training dynamics and implicit bias. In *International Conference on Machine Learning*, pages 25516–25549. PMLR, 2025.

[10] Juno Kim and Taiji Suzuki. Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. In *Forty-first International Conference on Machine Learning*, 2024.

[11] Daniel Kunin, Giovanni Luca Marchetti, Feng Chen, Dhruva Karkada, James B Simon, Michael R DeWeese, Surya Ganguli, and Nina Miolane. Alternating gradient flows: A theory of feature learning in two-layer neural networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[12] Ashok V Makkuva, Marco Bondaschi, Chanakya Ekbote, Adway Girish, Alliot Nagle, Hyeji Kim, and Michael Gastpar. Local to global: Learning dynamics and effect of initialization for transformers. *Advances in Neural Information Processing Systems*, 37:86243–86308, 2024.

[13] Pierre Marion, Raphaël Berthier, Gérard Biau, and Claire Boyer. Attention layers provably solve single-location regression. In *ICLR 2025-Thirteenth International Conference on Learning Representations*, 2025.

[14] Simon Martin, Francis Bach, and Giulio Biroli. On the impact of overparameterization on the training of a shallow neural network in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 3655–3663. PMLR, 2024.

[15] Simon Martin, Giulio Biroli, and Francis Bach. High-dimensional analysis of gradient flow for extensive-width quadratic neural networks. *arXiv preprint arXiv:2601.10483*, 2026.

[16] Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. In *Proceedings of the 41st International Conference on Machine Learning*, pages 38018–38070, 2024.

[17] Naoki Nishikawa, Yujin Song, Kazusato Oko, Denny Wu, and Taiji Suzuki. Nonlinear transformers can perform inference-time feature learning. In *Forty-second International Conference on Machine Learning*, 2025.

[18] Salim Noreddine and Ivan Nourdin. On the gaussian approximation of vector-valued multiple integrals. *Journal of multivariate analysis*, 102(6):1008–1017, 2011.

[19] Ivan Nourdin, Giovanni Peccati, and Anthony Réveillac. Multivariate normal approximation using stein's method and malliavin calculus. In *Annales de l'IHP Probabilités et statistiques*, volume 46, pages 45–58, 2010.

[20] David Nualart and Giovanni Peccati. Central limit theorems for sequences of multiple stochastic integrals. *The Annals of Probability*, 33(1):177–193, 1 2005.

[21] Farzad Pourkamali, Jean Barbier, and Nicolas Macris. Matrix inference in growing rank regimes. *IEEE Transactions on Information Theory*, 70(11):8133–8163, 2024.

[22] Farzad Pourkamali and Nicolas Macris. Rectangular rotational invariant estimator for high-rank matrix estimation. *Information and Inference: A Journal of the IMA*, 14(3):iaaf015, 2025.

[23] Chen Siyu, Sheen Heejune, Wang Tianhao, and Yang Zhuoran. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4573–4573. PMLR, 2024.

[24] Bingqing Song, Boran Han, Shuai Zhang, Jie Ding, and Mingyi Hong. Unraveling the gradient descent dynamics of transformers. *Advances in Neural Information Processing Systems*, 37:92317–92351, 2024.

[25] Yuandong Tian, Yiping Wang, Beidi Chen, and Simon S Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in neural information processing systems*, 36:71911–71947, 2023.

[26] Francois Treves. Ovsyannikov theorem and hyperdifferential operators. *Notas de Matematica*, 46, 1968.

[27] Emanuele Troiani, Hugo Cui, Yatin Dandi, Florent Krzakala, and Lenka Zdeborová. Fundamental limits of learning in sequence multi-index models and deep attention networks: High-dimensional asymptotics and sharp thresholds. In *Forty-second International Conference on Machine Learning*, 2025.

[28] Emanuele Troiani, Vittorio Erba, Florent Krzakala, Antoine Maillard, and Lenka Zdeborová. Optimal denoising of rotationally invariant rectangular matrices. In *Mathematical and Scientific Machine Learning*, pages 97–112. PMLR, 2022.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[30] Hongru Yang, Bhavya Kailkhura, Zhangyang Wang, Yingbin Liang, et al. Training dynamics of transformers to recognize word co-occurrence via gradient flow analysis. *Advances in Neural Information Processing Systems*, 37:46047–46117, 2024.

[31] Liu Ziyin, Yizhou Xu, Tomaso Poggio, and Isaac Chuang. Neural quadratic forms: A unified minimal model for sudden learning and scaling laws. *arXiv preprint arXiv:2608.13335*, 2026.

Contents

1 Introduction **1**

2 Attention-indexed models and their population loss landscape **2**

3 Online SGD of tied attention **4**

3.1 Macroscopic dynamics 4

3.2 Weak recovery and its sample complexity 6

4 Online SGD of untied attention **7**

4.1 Macroscopic dynamics 7

4.2 Weak recovery and its sample complexity 9

5 Conclusion and future work **10**

A Notation **14**

B Examples of attention-indexed models **15**

B.1 Example 1: Multi-layer multi-head attention 15

B.2 Example 2: Multiple-location regression 16

B.3 Example 3: Extensive-rank matrix denoising 16

C Theory in Section 2 **16**

C.1 Proof of Theorem 1 16

C.2 Tied attention 18

C.3 Proof of Corollary 1 19

D Theory in Section 3 **20**

D.1 Proof of Theorem 2 20

D.1.1 Local well-posedness 20

D.1.2 Uniform spectral bounds 22

D.1.3 Moment expansion 28

D.1.4 Compactness of the empirical trajectories 33

D.1.5 Global well-posedness 34

D.1.6 Final proof of Theorem 2 38

D.2 Proof of Corollary 2 39

D.3 An exactly solvable example 41

D.4 Proof of Corollary 3 42

D.5 Convergence rate of strong recovery 44

E Online SGD over S **46**

E.1 Proof of Theorem 4 48

E.1.1 Uniform spectral bounds . . . 48
E.1.2 Moment expansion . . . 51
E.1.3 Final proof of Theorem 4 . . . 53
E.2 Proof of Corollary 7 . . . 55

F Theory in Section 4 **57**
F.1 Assumptions for Theorem 3 . . . 57
F.2 Proof of Theorem 3 . . . 58
F.2.1 A stronger version of Theorem 1 . . . 58
F.2.2 Local well-posedness . . . 58
F.2.3 Uniform spectral bounds . . . 59
F.2.4 Fast-dynamics convergence . . . 73
F.2.5 Moment expansion . . . 75
F.2.6 Limiting slow dynamics . . . 84
F.2.7 Final proof of Theorem 3 . . . 90
F.3 Truncation to finite orders . . . 91
F.4 Proof of Corollary 4 . . . 94
F.5 Convergence rate of strong recovery . . . 95
F.6 Proof of Lemma 10 . . . 97

G Additional Experiments **98**

H Declaration of LLM usage **99**

A Notation

We write $[K] := \{1, \dots, K\}$ and let $\mathcal{K} \subseteq [K]$ denote the set of trainable indices. For a multi-index or word α, $|\alpha|$ denotes its length, while $|\alpha|_{\mathcal{K}}$ denotes the number of positions whose associated index belongs to $\mathcal{K}$. The underlying finite alphabet of α will be clear from context.

For matrices, $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_{\mathrm{F}}$ denote the operator and Frobenius norms, respectively, and $\langle A, B\rangle_F := \mathrm{Tr}(A^\top B)$. We write $\mathbb{S}^K$ for the space of real symmetric $K \times K$ matrices and $\mathbb{S}^K_+$ for the space of positive-semidefinite matrices. For symmetric matrices, $A \succeq B$ means that $A - B$ is positive semidefinite. For finite-dimensional vectors, $\|\cdot\|_\infty$ and $\|\cdot\|$ denote the ℓ_∞ and Euclidean norms.

We write $\mathbf{1}_E$ for the indicator of an event E, $a \wedge b := \min\{a, b\}$, and $[a]_+ := \max\{a, 0\}$. For sets $A \subseteq B$, the notation $A \Subset B$ means that $\overline{A}$ is a compact subset of B.

We use the standard asymptotic notation $O(\cdot)$, $o(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ as $d \to \infty$. For random variables, $X_d = O_{\mathbb{P}}(a_d)$ means that X_d/a_d is bounded in probability, $X_d = o_{\mathbb{P}}(a_d)$ means that $X_d/a_d \xrightarrow{\mathbb{P}} 0$, and $X_d = \Theta_{\mathbb{P}}(a_d)$ means that both $X_d = O_{\mathbb{P}}(a_d)$ and $a_d = O_{\mathbb{P}}(X_d)$. We write $X_d \xrightarrow{\mathbb{P}} X$ for convergence in probability.

For $s > 0$, let X_s denote the Banach space of moment sequences $\mu = (\mu_\alpha)_{|\alpha|\geq 1}$ equipped with the norm $\|\mu\|_s := \sup_{w\geq 1} \sup_{|\alpha|=w} \frac{|\mu_\alpha|}{s^w}$ and define $X_s^0 := \Big\{\mu \in X_s : \lim_{w\to\infty} \sup_{|\alpha|=w} \frac{|\mu_\alpha|}{s^w} = 0\Big\}$. Equipped with the norm inherited from X_s, X_s^0 is a separable Banach space. $C([0,T]; X_s^0)$ denotes the space of continuous functions from $[0,T]$ to X_s^0. $D([0,T]; X_s^0)$ represents the space of all right-continuous functions with left limits mapping from $[0,T]$ into X_s^0, equipped with the Skorokhod J_1 topology.

Throughout the paper, a superscript (d) denotes a finite-dimensional quantity, a bar denotes its deterministic high-dimensional limit, and a tilde denotes a piecewise-constant time interpolation, unless specified otherwise.

B Examples of attention-indexed models

B.1 Example 1: Multi-layer multi-head attention

In this example, we demonstrate that the standard multi-layer multi-head attention network (attention-only transformer) falls within our general theoretical framework.

Let the input to the model be Gaussian $X_0 \sim \mathcal{N}(0, \mathcal{C} \otimes I_d)$. For layer $m = 1, \ldots, M$ with H attention heads, the pre-softmax attention matrix for the h-th head is given by

$$A_m^{(h)}(X_{m-1}) = \frac{1}{\sqrt{d}} X_{m-1} S_{m-1}^{(h)} X_{m-1}^T \in \mathbb{R}^{L \times L} \tag{23}$$

where $S_{m-1}^{(h)} := K_{m-1}^{(h)} (Q_{m-1}^{(h)})^T \in \mathbb{R}^{d \times d}$ is the combined key-query weight matrix. The intermediate representations follow the standard multi-head attention mechanism:

$$X_m = X_{m-1} + \sum_{h=1}^{H} \text{softmax}\left(A_m^{(h)}(X_{m-1}) \right) X_{m-1} V_{m-1}^{(h)} \in \mathbb{R}^{L \times d}, \tag{24}$$

where $V_{m-1}^{(h)} \in \mathbb{R}^{d \times d}$ represents the combined value and output projection matrix for head h. Unrolling the recurrence relation reveals that this network can be expressed as a composite function of the quadratic forms $G_k := \frac{1}{\sqrt{d}} X_0 \tilde{S}_k X_0^T$.

For the first layer ($m = 1$), the pre-softmax matrix constitutes a bilinear form:

$$A_1^{(h)}(X_0) = \frac{1}{\sqrt{d}} X_0 S_0^{(h)} X_0^T := G_0^{(h)}. \tag{25}$$

The output of the first layer is thus $X_1 = X_0 + \sum_{h=1}^{H} \text{softmax}(G_0^{(h)}) X_0 V_0^{(h)}$.

For the second layer ($m = 2$), substituting X_1 into the pre-softmax attention matrix $A_2^{(h)}(X_1) = \frac{1}{\sqrt{d}} X_1 S_1^{(h)} X_1^T$ yields a sum of four types of cross terms:

$$\begin{aligned} A_2^{(h)}(X_1) =& \frac{1}{\sqrt{d}} X_0 S_1^{(h)} X_0^T + \sum_{i=1}^{H} \text{softmax}(G_0^{(i)}) \left(\frac{1}{\sqrt{d}} X_0 \left(V_0^{(i)} S_1^{(h)} \right) X_0^T \right) \\ &+ \sum_{j=1}^{H} \left(\frac{1}{\sqrt{d}} X_0 \left(S_1^{(h)} V_0^{(j)T} \right) X_0^T \right) \text{softmax}(G_0^{(j)})^T \\ &+ \sum_{i=1}^{H} \sum_{j=1}^{H} \text{softmax}(G_0^{(i)}) \left(\frac{1}{\sqrt{d}} X_0 \left(V_0^{(i)} S_1^{(h)} V_0^{(j)T} \right) X_0^T \right) \text{softmax}(G_0^{(j)})^T. \end{aligned} \tag{26}$$

Every inner term encapsulated by parentheses in the expansion above takes the form $\frac{1}{\sqrt{d}} X_0 \tilde{S} X_0^T$. For instance, the attention outputs cross term forms a new bilinear form defined by the effective weight matrix $\tilde{S} := V_0^{(i)} S_1^{(h)} V_0^{(j)T}$.

By induction, any attention matrix A_m, and consequently the final output representation, can be written as a finite composite function of a set of base bilinear forms $\{G_k\}_{k=1}^K$ generated by equivalent effective weight matrices $\tilde{S}_k$. These effective matrices $\tilde{S}_k$ are finite products of the original network weights (S, V), which falls within the scope of attention-indexed models.

Furthermore, because the softmax function is continuous and bounded, the composite function remains continuous and exhibits at most polynomial growth. Therefore, when we assume that both the teacher and the student models are the multi-layer multi-head attention architecture, the final loss satisfies Assumption 1. Note that we do not require the teacher weights to be Gaussian, making the existence of such a teacher model a weak assumption.

B.2 Example 2: Multiple-location regression

This example considers a classification task (e.g., sentiment analysis), where the goal is to predict the label $y \in \{+1, -1\}$ based on the interaction between tokens (e.g., detecting negation or intensification). We model the student predictor as a one-layer attention:

$$\hat{y} = \text{sign}\left(\sum_{i,j} \alpha_{ij} \sigma\left(\{x_i^T W^h x_j\}_{ij}\right)\right), \tag{27}$$

where the model learns α_{ij} to weight different token pairs. Moreover, we assume that the true label is generated by a "bag-of-interactions" model:

$$y = \text{sign}\left(\sum_{i,j} S_{ij} \cdot \epsilon_{ij}\right), \tag{28}$$

where $S_{ij} \in \{+1, -1\}$ represents the inherent sentiment polarity of the word pair (i, j) (e.g., $+1$ for "very good", -1 for "not good"), and ϵ_{ij} indicates whether a meaningful grammatical connection exists between them:

$$\epsilon_{ij} = \begin{cases} 1, & \text{w.p. } g(\{x_i^T W^* x_j\}_{ij}) \\ 0, & \text{otherwise.} \end{cases} \tag{29}$$

In this setting, ϵ_{ij} can be interpreted as the ground-truth dependency graph (e.g., i modifies j), which is activated probabilistically by the underlying semantic matching W^*.

B.3 Example 3: Extensive-rank matrix denoising

Consider the observation model:

$$Y = S^* + Z \in \mathbb{R}^{d \times d}, \tag{30}$$

where S^* is an unknown deterministic symmetric matrix with an extensive rank (i.e., $\lim_{d\to\infty} \|S^*\|_{\text{op}}/\|S^*\|_F = 0$), and Z is the noise matrix drawn from the Gaussian Orthogonal Ensemble (GOE), satisfying $\mathbb{E}[Z_{ab}] = 0$ and $\mathbb{E}[Z_{ab}Z_{cd}] = \frac{1}{d}(\delta_{ac}\delta_{bd} + \delta_{ad}\delta_{bc})$. While prior work has extensively studied rotationally invariant priors [21, 22, 28], the general extensive-rank setting remains largely open [2].

A standard approach for inferring S^* from Y is via:

$$\min_S \mathcal{L}\left(\frac{1}{d}\text{Tr}[YS]\right) + \lambda f(S),$$

where $\mathcal{L} : \mathbb{R} \to \mathbb{R}$ is the loss function and $f(S)$ represents a regularizer. The population loss can be expressed as an expectation over a standard scalar Gaussian variable $g \sim \mathcal{N}(0, 1)$:

$$\mathbb{E}_Z \mathcal{L}\left(\frac{1}{d}\text{Tr}[YS]\right) = \mathbb{E}_{g\sim\mathcal{N}(0,1)}\left[\mathcal{L}\left(\frac{1}{d}\text{Tr}[S^*S] + \frac{\sqrt{2}}{d^{3/2}}\|S\|_F g\right)\right], \tag{31}$$

which can be regarded as a single-token single-index attention-indexed model as a special case of Theorem 1 and Corollary 5.

C Theory in Section 2

C.1 Proof of Theorem 1

The proof relies on the method of joint cumulants. If all joint cumulants of order $m \geq 3$ for a set of random variables vanish in the limit, the variables converge in distribution to a multivariate Gaussian (satisfying Carleman's condition).

Let $Z_{ijk} = \frac{1}{\sqrt{d}} x_i^T S_{ij}^k x_j$. We concatenate all data vectors into a single vector $X = [x_1^T, x_2^T, \ldots, x_L^T]^T \in \mathbb{R}^{dL}$. Given the covariance structure $\mathbb{E}[x_{ia}x_{jb}] = \mathcal{C}_{ij}\delta_{ab}$, we have $X \sim$

$\mathcal{N}(0, \Sigma)$, where the covariance matrix is the Kronecker product $\Sigma = \mathcal{C} \otimes I_d$. We can rewrite Z_{ijk} as a quadratic form:

$$Z_{ijk} = X^T A^{(ijk)} X, \tag{32}$$

where $A^{(ijk)}$ is a $dL \times dL$ symmetric block matrix. Its (i, j)-th block is $\frac{1}{2\sqrt{d}} S_{ij}^k$, its (j, i)-th block is $\frac{1}{2\sqrt{d}} (S_{ij}^k)^T$, and all other blocks are zero.

By the standard expectation of Gaussian quadratic forms, the first moment is:

$$\mathbb{E}[Z_{ijk}] = \mathrm{Tr}(A^{(ijk)}\Sigma) = \frac{1}{\sqrt{d}} \sum_{a=1}^{d} \mathbb{E}[x_{ia} x_{ja}] (S_{ij}^k)_{aa} = \frac{\mathcal{C}_{ij}}{\sqrt{d}} \mathrm{Tr}(S_{ij}^k). \tag{33}$$

By Assumption 1.2, as $d \to \infty$, this converges to $\mathbb{E}[G_{ijk}] = \mathcal{C}_{ij}\mu_{ijk}$.

To calculate the covariance, we use Wick's Theorem:

$$\mathrm{Cov}(Z_{ijk}, Z_{i'j'k'}) = 2\mathrm{Tr}(A^{(ijk)}\Sigma A^{(i'j'k')}\Sigma). \tag{34}$$

Expanding this into block components yields:

$$\begin{aligned} \mathrm{Cov}(Z_{ijk}, Z_{i'j'k'}) &= \frac{1}{d} \sum_{a,b,c,e} (S_{ij}^k)_{ab} (S_{i'j'}^{k'})_{ce} \left(\mathcal{C}_{ii'}\mathcal{C}_{jj'}\delta_{ac}\delta_{be} + \mathcal{C}_{ij'}\mathcal{C}_{ji'}\delta_{ae}\delta_{bc} \right) \\ &= \mathcal{C}_{ii'}\mathcal{C}_{jj'} \frac{\mathrm{Tr}(S_{ij}^k (S_{i'j'}^{k'})^T)}{d} + \mathcal{C}_{ij'}\mathcal{C}_{ji'} \frac{\mathrm{Tr}(S_{ij}^k S_{i'j'}^{k'})}{d}. \end{aligned} \tag{35}$$

By Assumption 1.2, as $d \to \infty$, this converges to the finite constant $\mathcal{C}_{ii'}\mathcal{C}_{jj'}\Omega_{ijk,i'j'k'} + \mathcal{C}_{ij'}\mathcal{C}_{ji'}\Psi_{ijk,i'j'k'}$.

Now we calculate the joint cumulant $\kappa_m := \kappa(Q_1, \cdots, Q_m)$ for $Q_r = X^T A_r X$, where we denote $A_r := A^{(i_r j_r k_r)}$ for $r \in \{1, \ldots, m\}$. The joint cumulant generating function reads

$$K(t_1, \ldots, t_m) = \log \mathbb{E}\left[\exp\left(\sum_{r=1}^{m} t_r Q_r \right) \right] = \log \mathbb{E}\left[\exp(X^T \tilde{A} X) \right], \tag{36}$$

where $\tilde{A} = \sum_{r=1}^{m} t_r A_r$. Using the standard Gaussian integral for quadratic forms in the neighborhood of $(t_1, \ldots, t_m) = 0$, the expectation can be calculated as

$$K(t_1, \ldots, t_m) = -\frac{1}{2} \log\det(I - 2\Sigma\tilde{A}). \tag{37}$$

Using the identity $\log\det(I - M) = -\sum_{n=1}^{\infty} \frac{1}{n} \mathrm{Tr}(M^n)$, the right side can be expanded into a series of traces:

$$K(t_1, \ldots, t_m) = \sum_{n=1}^{\infty} \frac{2^{n-1}}{n} \mathrm{Tr}\left((\Sigma\tilde{A})^n \right) = \sum_{n=1}^{\infty} \frac{2^{n-1}}{n} \mathrm{Tr}\left(\left(\sum_{r=1}^{m} t_r \Sigma A_r \right)^n \right). \tag{38}$$

Then the joint cumulant is given by $\kappa_m := \left. \frac{\partial^m K}{\partial t_1 \cdots \partial t_m} \right|_{t_1 = \cdots = t_m = 0}$. The only term in the series that contributes to this cross-derivative is $n = m$. Therefore, after expanding the m-th power, κ_m is a linear combination of traces over all permutations π in the symmetric group $\mathcal{S}_m$:

$$\kappa_m = c_m \sum_{\pi \in \mathcal{S}_m} \mathrm{Tr}\left(A_{\pi(1)} \Sigma A_{\pi(2)} \Sigma \ldots A_{\pi(m)} \Sigma \right), \tag{39}$$

where c_m is a constant depending only on m. Since m is fixed, the sum consists of a finite number of terms. We can bound the absolute value of each individual trace in the sum using the matrix norm inequality $|\mathrm{Tr}(B_1 B_2 \ldots B_m)| \leq \left(\prod_{r=1}^{m-2} ||B_r||_{op} \right) ||B_{m-1}||_F ||B_m||_F$.

Let $B_r = A^{(i_r j_r k_r)}\Sigma$ and let us analyze the asymptotic scales of these norms: $||\Sigma||_{op} = \lambda_{\max}(\mathcal{C} \otimes I_d) = \lambda_{\max}(\mathcal{C}) = \Theta(1)$. $||A^{(ijk)}||_F = \frac{1}{\sqrt{2d}} ||S_{ij}^k||_F = \mathcal{O}(1)$, based on Assumption 1.2. From Assumption 1.1, $||S_{ij}^k||_{op} = o(||S_{ij}^k||_F) = o(\sqrt{d})$. Therefore, $||A^{(ijk)}||_{op} = \frac{1}{2\sqrt{d}} ||S_{ij}^k||_{op} = o(1)$.

Thus, for B_r, the spectral norm is bounded by $||B_r||_{op} \leq ||A^{(i_r j_r k_r)}||_{op}||\Sigma||_{op} = o(1)\cdot\mathcal{O}(1) = o(1)$, and the Frobenius norm is $\mathcal{O}(1)$. Applying this to the trace inequality for any $m \geq 3$:

$$\kappa_m = O\left([o(1)]^{m-2} \cdot \mathcal{O}(1) \cdot \mathcal{O}(1)\right) = o(1). \tag{40}$$

Therefore, all joint cumulants of order 3 and higher vanish as $d \to \infty$.

Because all higher-order cumulants vanish while the first two cumulants converge to finite limits, the vector of quadratic forms $\{Z_{ijk}\}_{i,j,k=1}^{L,L,K}$ converges weakly in distribution to the multivariate Gaussian vector $\{G_{ijk}\}_{i,j,k=1}^{L,L,K}$.

Finally, to pass the limit inside the expectation, we need to show that the sequence $\mathcal{L}(\{Z_{ijk}\}_{i,j,k=1}^{L,L,K})$ is uniformly integrable. By Assumption 1.3, $\mathcal{L}$ is a continuous function with at most polynomial growth, meaning there exist constants $c > 0$ and $p > 0$ such that $|\mathcal{L}(Z)| \leq c(1 + ||Z||^p)$. Therefore, it suffices to show that the random tensor $Z = \{Z_{ijk}\}_{i,j,k=1}^{L,L,K}$ has bounded joint moments of all orders (up to $p+1$) as $d \to \infty$.

Recall that any joint moment of a set of random variables can be expressed as a finite polynomial of their joint cumulants. From our previous derivations, the joint cumulants κ_m converge to finite limits (specifically, bounded constants for $m \leq 2$ and 0 for $m \geq 3$). Because all joint cumulants of Z are uniformly bounded with respect to d, it immediately follows that all joint moments of Z (up to $p+1$) are also uniformly bounded.

This uniform boundedness of all moments guarantees that $\mathcal{L}(\{Z_{ijk}\}_{i,j,k=1}^{L,L,K})$ is uniformly integrable. Thus, we can exchange the limit and the expectation, yielding:

$$\lim_{d\to\infty} \mathbb{E}_x[\mathcal{L}(\{Z_{ijk}\}_{i,j,k=1}^{L,L,K})] = \mathbb{E}[\mathcal{L}(\{G_{ijk}\}_{i,j,k=1}^{L,L,K})]. \tag{41}$$

This completes the proof.

C.2 Tied attention

The assumption $\lim_{d\to\infty} \frac{1}{\sqrt{d}}\mathrm{Tr}(S_{ij}^k) = \mu_{ijk}$ in Theorem 1 might not be satisfied for certain architectures, particularly those with tied weights. For instance, if the attention mechanism uses symmetric tied weights (e.g., $S = WW^T$), its trace scales proportionally to the dimension d, causing the term $\frac{1}{\sqrt{d}}\mathrm{Tr}(S)$ to diverge as $d \to \infty$. To accommodate such cases, we introduce a centered variant of the attention-indexed model.

Corollary 5. *Under the Assumption 1 except that the condition* $\lim_{d\to\infty} \frac{1}{\sqrt{d}}\mathrm{Tr}(S_{ij}^k) = \mu_{ijk}$ *is dropped, we have*

$$\lim_{d\to\infty} \mathbb{E}_x \left[\mathcal{L}\left(\left\{ \frac{\mathrm{Tr}[S_{ij}^k(x_j x_i^T - \mathcal{C}_{ij} I_d)]}{\sqrt{d}} \right\}_{i,j,k=1}^{L,L,K} \right) \right] = \mathbb{E}\left[\mathcal{L}(\{G_{ijk}\}_{i,j,k=1}^{L,L,K}) \right] \tag{42}$$

where $\{G_{ijk}\}_{i,j,k=1}^{L,L,K}$ *is a multivariate Gaussian vector whose mean and covariance are given by:*

$$\mathbb{E}[G_{ijk}] = 0,\ \mathit{Cov}(G_{ijk}, G_{i'j'k'}) = \mathcal{C}_{ii'}\mathcal{C}_{jj'}\Omega_{ijk,i'j'k'} + \mathcal{C}_{ij'}\mathcal{C}_{ji'}\Psi_{ijk,i'j'k'}. \tag{43}$$

Proof. Let $Y_{ijk} = \frac{1}{\sqrt{d}}\mathrm{Tr}[S_{ij}^k(x_j x_i^T - \mathcal{C}_{ij} I_d)]$. We can rewrite this as a centered quadratic form:

$$Y_{ijk} = \frac{1}{\sqrt{d}}(x_i^T S_{ij}^k x_j - \mathcal{C}_{ij}\mathrm{Tr}(S_{ij}^k)) = Z_{ijk} - \mathbb{E}[Z_{ijk}], \tag{44}$$

where $Z_{ijk} = \frac{1}{\sqrt{d}} x_i^T S_{ij}^k x_j$ as defined in the proof of Theorem 1.

By construction, the first moment of Y_{ijk} is zero for any d: $\mathbb{E}[Y_{ijk}] = 0$. Because Y_{ijk} is a deterministic shift of Z_{ijk}, all joint cumulants of order $m \geq 2$ are identical to those of Z_{ijk}. Specifically, the covariance structure remains:

$$\lim_{d\to\infty} \mathrm{Cov}(Y_{ijk}, Y_{i'j'k'}) = \lim_{d\to\infty} \mathrm{Cov}(Z_{ijk}, Z_{i'j'k'}) = \mathcal{C}_{ii'}\mathcal{C}_{jj'}\Omega_{ijk,i'j'k'} + \mathcal{C}_{ij'}\mathcal{C}_{ji'}\Psi_{ijk,i'j'k'}. \tag{45}$$

Furthermore, for any order $m \geq 3$, the joint cumulants of $\{Y_{ijk}\}$ match those of $\{Z_{ijk}\}$ and thus vanish as $o(1)$, following Theorem 1.

Since the first two cumulants converge to finite limits and all higher-order cumulants vanish, the sequence of random vectors $\{Y_{ijk}\}_{i,j,k=1}^{L,L,K}$ converges weakly in distribution to $\{G_{ijk}\}_{i,j,k=1}^{L,L,K}$, which is a multivariate Gaussian vector with mean zero and the asymptotic covariance structure derived above.

Because the loss function $\mathcal{L}$ is continuous with at most polynomial growth, and $\{Y_{ijk}\}$ has bounded moments (due to the bounds on the spectral and Frobenius norms of S_{ij}^k by Assumption 1), the sequence $\mathcal{L}(\{Y_{ijk}\}_{i,j,k=1}^{L,L,K})$ is uniformly integrable. This allows us to pass the limit inside the expectation:

$$\lim_{d\to\infty} \mathbb{E}_x \left[\mathcal{L}\left(\{Y_{ijk}\}_{i,j,k=1}^{L,L,K} \right) \right] = \mathbb{E}\left[\mathcal{L}(\{G_{ijk}\}_{i,j,k=1}^{L,L,K}) \right]. \tag{46}$$

This completes the proof. □

C.3 Proof of Corollary 1

Corollary 1 is proven under the following regularity assumption.

Assumption 5. *For a given dimension d, let $\hat{S}^{(d)} = \{\hat{S}_{ij}^{k,(d)}\}_{i,j=1,k\in\mathcal{K}}^{L,L}$ be a sequence of minimizers of the population loss with respect to the learnable weights. Assume that as $d \to \infty$:*

1. *The minimizers satisfy the spectral norm condition uniformly: $\lim_{d\to\infty} \|\hat{S}_{ij}^{k,(d)}\|_{op}/\|\hat{S}_{ij}^{k,(d)}\|_F = 0$ for all $k \in \mathcal{K}$.*
2. *The order parameters associated with both the minimizers and the fixed weights, denoted as $Q(\hat{S}^{(d)})$, converge to a valid point $\hat{Q} \in \mathcal{Q}$.*

Now we prove Corollary 1.

Let $S_{\mathcal{K}}$ denote the collection of learnable weights $\{S_{ij}^k\}_{k\in\mathcal{K}}$. Let the population loss in dimension d be denoted by:

$$F_d(S_{\mathcal{K}}) := \mathbb{E}_x \left[\mathcal{L}\left(\left\{ \frac{x_i^T S_{ij}^k x_j}{\sqrt{d}} \right\}_{i,j,k=1}^{L,L,K} \right) \right]. \tag{47}$$

We aim to prove that $\lim_{d\to\infty} \inf_{S_{\mathcal{K}}} F_d(S_{\mathcal{K}}) = \inf_{Q\in\mathcal{Q}} \mathcal{R}(Q)$ by establishing matching upper and lower bounds.

For any arbitrary $\epsilon > 0$, by the definition of the infimum, there exists $Q_\epsilon \in \mathcal{Q}$ such that:

$$\mathcal{R}(Q_\epsilon) \leq \inf_{Q\in\mathcal{Q}} \mathcal{R}(Q) + \epsilon. \tag{48}$$

Because Q_ϵ lies within the feasible domain $\mathcal{Q}$, there exists a sequence of learnable matrices $\tilde{S}_{\mathcal{K}}^{(d)}$ such that, combined with the fixed matrices, the joint order parameters satisfy $\lim_{d\to\infty} Q(\tilde{S}_{\mathcal{K}}^{(d)}) = Q_\epsilon$, and the learnable matrices satisfy the norm condition $\lim_{d\to\infty} \|\tilde{S}_{ij}^{k,(d)}\|_{\text{op}}/\|\tilde{S}_{ij}^{k,(d)}\|_F = 0$.

Since the infimum over all possible learnable weights $S_{\mathcal{K}}$ must be less than or equal to the loss evaluated at this specific sequence $\tilde{S}_{\mathcal{K}}^{(d)}$, we have:

$$\inf_{S_{\mathcal{K}}} F_d(S_{\mathcal{K}}) \leq F_d(\tilde{S}_{\mathcal{K}}^{(d)}). \tag{49}$$

Taking the limit superior as $d \to \infty$ on both sides, and noting that the complete set of matrices (both $\tilde{S}_{\mathcal{K}}^{(d)}$ and the fixed matrices) satisfies all conditions of Theorem 1, we can apply Theorem 1 to the right-hand side:

$$\limsup_{d\to\infty} \inf_{S_{\mathcal{K}}} F_d(S_{\mathcal{K}}) \leq \lim_{d\to\infty} F_d(\tilde{S}_{\mathcal{K}}^{(d)}) = \mathcal{R}(Q_\epsilon) \leq \inf_{Q\in\mathcal{Q}} \mathcal{R}(Q) + \epsilon. \tag{50}$$

Taking the limit as $\epsilon \to 0$ yields the upper bound:

$$\limsup_{d\to\infty} \inf_{S_{\mathcal{K}}} F_d(S_{\mathcal{K}}) \leq \inf_{Q\in\mathcal{Q}} \mathcal{R}(Q). \tag{51}$$

Let $\hat{S}_{\mathcal{K}}^{(d)}$ be the sequence of minimizers such that $\inf_{S_{\mathcal{K}}} F_d(S_{\mathcal{K}}) = F_d(\hat{S}_{\mathcal{K}}^{(d)})$. According to Assumption 5.1, the sequence $\hat{S}_{\mathcal{K}}^{(d)}$ uniformly satisfies the norm ratio condition, and the combined order parameters converge to some state $\hat{Q} \in \mathcal{Q}$. We can then apply Theorem 1 to this joint sequence of minimizers and fixed weights:

$$\lim_{d\to\infty} F_d(\hat{S}_{\mathcal{K}}^{(d)}) = \mathcal{R}(\hat{Q}). \tag{52}$$

Furthermore, since $\hat{Q}$ is an element of the feasible domain $\mathcal{Q}$, we have $\mathcal{R}(\hat{Q}) \geq \inf_{Q\in\mathcal{Q}} \mathcal{R}(Q)$. Therefore, taking the limit inferior, we obtain:

$$\liminf_{d\to\infty} \inf_{S_{\mathcal{K}}} F_d(S_{\mathcal{K}}) = \lim_{d\to\infty} F_d(\hat{S}_{\mathcal{K}}^{(d)}) = \mathcal{R}(\hat{Q}) \geq \inf_{Q\in\mathcal{Q}} \mathcal{R}(Q). \tag{53}$$

Combining the upper bound (51) and the lower bound (53), we get:

$$\lim_{d\to\infty} \inf_{S_{\mathcal{K}}} F_d(S_{\mathcal{K}}) = \inf_{Q\in\mathcal{Q}} \mathcal{R}(Q). \tag{54}$$

This concludes the proof.

D Theory in Section 3

D.1 Proof of Theorem 2

D.1.1 Local well-posedness

We first define a scale of the Banach spaces X_s for $s > 0$, consisting of infinite-dimensional moment sequences $\mu = (\mu_\alpha)_{|\alpha|\geq 1}$ equipped with the norm $\|\mu\|_s := \sup_{n\geq 1} \sup_{|\alpha|=n} |\mu_\alpha|/s^n$.

By Price's theorem, the matrix derivative of the potential can be written as $\nabla\Phi(q) = \Gamma_{kl}(q)$ with

$$\Gamma_{kl}(q) := \frac{1}{2} \sum_{i,j,i',j'=1}^{L} (\mathcal{C}_{ii'}\mathcal{C}_{jj'} + \mathcal{C}_{ij'}\mathcal{C}_{ji'})\, \mathbb{E}\left[\partial^2_{(ijk),(i'j'l)}\mathcal{L}(G)\right], \qquad 1 \leq k, l \leq K. \tag{55}$$

See Lemma 2 for more detailed derivation. A technical issue is that the potential $\Phi(q)$, and hence the gradient Γ (55), are well-defined only for $q \succeq 0$, whereas an arbitrary element of X_s need not have a PSD second-moment matrix. To use the Ovsyannikov argument, we extend Γ outside the PSD cone. Define

$$\mathscr{P}(q) := \Pi_{\mathbb{S}_+^K}\left(\frac{q+q^T}{2}\right), \tag{56}$$

where $\Pi_{\mathbb{S}_+^K}$ denotes the orthogonal projection onto the closed convex cone of PSD matrices, and set

$$\widetilde{\Gamma}(q) := \Gamma(\mathscr{P}(q)), \qquad q \in \mathbb{R}^{K\times K}. \tag{57}$$

For $q \succeq 0$, this extension agrees with the generalized gradient: $\widetilde{\Gamma}(q) = \Gamma(q)$.

Now we denote the flow as

$$\frac{d}{dt}\bar{\mu}(t) = V(\bar{\mu}(t)) := F(\bar{\mu}(t)) + G(\bar{\mu}(t)), \tag{58}$$

where

$$[F(\mu)]_{(k_1,\ldots,k_w)} := -\sum_{\substack{i=1\\ k_i\in\mathcal{K}}}^{w} \sum_{l=1}^{K} \widetilde{\Gamma}(q)_{k_i l} \left(\mu_{(k_1,\ldots,k_{i-1},l,k_i,\ldots,k_w)} + \mu_{(k_1,\ldots,k_i,l,k_{i+1},\ldots,k_w)}\right). \tag{59}$$

and

$$[G(\bar{\mu})]_\alpha := -\frac{\gamma}{2}|\alpha|_{\mathcal{K}}\bar{\mu}_\alpha \tag{60}$$

are two unbounded operators in the Banach space X_s.

Lemma 1. *Given an initial condition $\bar{\mu}(0) \in X_{s_0}$ for some $s_0 > 0$, there exist $T_{\mathrm{local}} > 0$ and a strictly increasing continuous function $s(t)$ with $s(0) = s_0$ such that* (58) *admits a unique solution $\bar{\mu}(t) \in X_{s(t)}$ for $t \in [0, T_{\mathrm{local}})$.*

Proof. For every bounded subset of $\mathbb{S}_+^K$, $\Gamma(q) \in \mathbb{S}^K$ is bounded and Lipschitz continuous by definition (55) and Assumption 2.1 (using Price's theorem). Therefore, for every $R > 0$ there exist constants $C_\Gamma(R)$ and $L_\Gamma(R)$ such that

$$\sup_{q \succeq 0, \|q\|_F \leq R} \|\Gamma(q)\|_F \leq C_\Gamma(R), \tag{61}$$

and

$$\|\Gamma(q) - \Gamma(q')\|_F \leq L_\Gamma(R)\|q - q'\|_F \tag{62}$$

whenever $q, q' \succeq 0$ and $\|q\|_F, \|q'\|_F \leq R$. Since the orthogonal projection onto the closed convex cone $\mathbb{S}_+^K$ satisfies $\|\mathscr{P}(q) - \mathscr{P}(q')\|_F \leq \|q - q'\|_F$, $\widetilde{\Gamma}(q) = \Gamma(\mathscr{P}(q))$ is locally bounded and locally Lipschitz on $\mathbb{R}^{K \times K}$.

Now fix $0 < s < s'$ and $R > 0$, and let $B_R(X_s) := \{\mu \in X_s : \|\mu\|_s \leq R\}$. For $\mu, \nu \in B_R(X_s)$,

$$|q_{kl}(\mu)| \leq Rs^2, \qquad |q_{kl}(\mu) - q_{kl}(\nu)| \leq s^2\|\mu - \nu\|_s, \tag{63}$$

where we denote $q_{kl}(\mu) = \mu_{(k,l)}$ Hence, because K is fixed, we have

$$|\widetilde{\Gamma}(q(\mu))|_F \leq C_{R,s}, \tag{64}$$

and

$$\|\widetilde{\Gamma}(q(\mu)) - \widetilde{\Gamma}(q(\nu))\|_F \leq L_{R,s}\|\mu - \nu\|_s. \tag{65}$$

for constants $C_{R,s}$ and $L_{R,s}$.

Now we prove that the vector field $V : X_s \to X_{s'}$ is bounded and locally Lipschitz. For the linear term G, we have $|[G(\mu)]_\alpha - [G(\nu)]_\alpha| \leq \frac{\gamma}{2} w |\mu_\alpha - \nu_\alpha|$, and thus

$$\|G(\mu) - G(\nu)\|_{s'} = \sup_{|\alpha| = w \geq 1} \frac{\gamma w |\mu_\alpha - \nu_\alpha|}{2(s')^w} \leq \frac{\gamma}{2}\|\mu - \nu\|_s \sup_{w \geq 1} \left[w \left(\frac{s}{s'} \right)^w \right]. \tag{66}$$

Using the inequality $\sup_{w \geq 1} w x^w \leq \frac{1}{e|\log x|} \leq \frac{1}{1-x}$ for $x \in (0,1)$, we obtain:

$$\sup_{w \geq 1} w \left(\frac{s}{s'} \right)^w \leq \frac{s'}{s' - s}. \tag{67}$$

Thus, $\|G(\mu) - G(\nu)\|_{s'} \leq \frac{\gamma s'}{2(s'-s)}\|\mu - \nu\|_s$.

For the non-linear term $F(\mu)$, the difference $[F(\mu)]_\alpha - [F(\nu)]_\alpha$ for a multi-index α of length w is bounded by

$$|[F(\mu)]_\alpha - [F(\nu)]_\alpha| \tag{68}$$

$$\leq \sum_{\substack{i=1 \\ k_i \in \mathcal{K}}}^{w} \sum_l \left(\left| \widetilde{\Gamma}(q(\mu))_{k_i l} - \widetilde{\Gamma}(q(\nu))_{k_i l} \right| \left| \mu_{w+1}^{(1)} + \mu_{w+1}^{(2)} \right| + \left| \widetilde{\Gamma}(q(\nu))_{k_i l} \right| \left| \left(\mu_{w+1}^{(1)} + \mu_{w+1}^{(2)} \right) - \left(\nu_{w+1}^{(1)} + \nu_{w+1}^{(2)} \right) \right| \right). \tag{69}$$

Since the moments of length $w+1$ are bounded by $\|\mu\|_s s^{w+1} \leq Rs^{w+1}$ and their differences are bounded by $\|\mu - \nu\|_s s^{w+1}$, we can obtain:

$$|[F(\mu)]_\alpha - [F(\nu)]_\alpha| \leq wK \left[L_{R,s}\|\mu - \nu\|_s \left(2Rs^{w+1} \right) + C_{R,s} \left(2\|\mu - \nu\|_s s^{w+1} \right) \right]. \tag{70}$$

Computing the norm in $X_{s'}$, we obtain:

$$\|F(\mu) - F(\nu)\|_{s'} = \sup_{|\alpha| = w \geq 1} \frac{|[F(\mu)]_\alpha - [F(\nu)]_\alpha|}{(s')^w} \leq 2K \left(RL_{R,s} + C_{R,s} \right) s \|\mu - \nu\|_s \sup_{w \geq 1} \left[w \left(\frac{s}{s'} \right)^w \right]. \tag{71}$$

Using the previously established inequality $\sup_{w \geq 1} w \left(\frac{s}{s'} \right)^w \leq \frac{s'}{s' - s}$, we consequently arrive at:

$$\|F(\mu) - F(\nu)\|_{s'} \leq \frac{2K \left(RL_{R,s} + C_{R,s} \right) s s'}{s' - s} \|\mu - \nu\|_s. \tag{72}$$

Combining F and G, the vector field $V : X_s \to X_{s'}$ is locally Lipschitz and bounded (using $V(0) = 0$). Then Lemma 1 is proven by the Ovsyannikov theorem [26]. □

D.1.2 Uniform spectral bounds

Let $\mathcal{F}_n = \sigma(x^{(0)}, x^{(1)}, \dots, x^{(n-1)})$ be the filtration generated by the data samples up to step $n-1$. The weights $W_k^{(n)}$ are $\mathcal{F}_n$-measurable. In the online SGD, the update at step n relies on a fresh sample $x^{(n)}$. Consider the stochastic gradient:

$$g_k^{(n)}(x^{(n)}, W) := \nabla_{W_k} \mathcal{L}(G^{(n)}(x^{(n)}, W)), \tag{73}$$

and the expected gradient $\bar{g}_k^{(n)} = \mathbb{E}_x \left[\sum_{i,j=1}^L \partial_{(i,j,k)} \mathcal{L}(G^{(n)}) X_{ij} W_k \right]$ conditioned on $\mathcal{F}_n$.

Lemma 2. *Fix a constant $C_* > 0$. Suppose that, at step n,*

$$\max_{1 \le l \le K} \|W_l^{(n)}\|_{\mathrm{op}} \le C_*. \tag{74}$$

The conditional expected gradient $\bar{g}_k^{(n)}$ satisfies $\bar{g}_k^{(n)} = \frac{1}{d} H_k^{(n)} W_k^{(n)} + \mathcal{E}_{\mathrm{Stein},k}^{(n)}$, where

$$\|H_k^{(n)}\|_{\mathrm{op}} \le C_H(C_*), \qquad \|\mathcal{E}_{\mathrm{Stein},k}^{(n)}\|_{\mathrm{op}} \le C_E(C_*)\, d^{-3/2}, \tag{75}$$

for constants $C_H(C_*)$ and $C_E(C_*)$ independent of d and n.

Proof. Throughout the proof we condition on $\mathcal{F}_n$ and suppress the superscript (n) whenever no confusion can arise. Conditional on $\mathcal{F}_n$, the matrices $\{W_l, S_l\}_{l=1}^K$ are deterministic, while $x = (x_1, \dots, x_L) \sim \mathcal{N}(0, \mathcal{C} \otimes I_d)$.

Recall that

$$G_{ijk} := \frac{1}{\sqrt{d}} \mathrm{Tr}\big[S_k (x_j x_i^T - \mathcal{C}_{ij} I_d)\big]. \tag{76}$$

Since $S_k = S_k^T$, we have $\nabla_{W_k} G_{ijk} = X_{ij} W_k$ with $X_{ij} := \frac{1}{\sqrt{d}} \big(x_j x_i^T + x_i x_j^T - 2\mathcal{C}_{ij} I_d\big)$. Consequently,

$$g_k = \sum_{i,j=1}^L f_{ijk}(x) X_{ij} W_k, \qquad f_{ijk}(x) := \partial_{(ijk)} \mathcal{L}(G(x)), \tag{77}$$

and hence

$$\bar{g}_k = \sum_{i,j=1}^L \mathbb{E}_x[f_{ijk}(x) X_{ij}] W_k. \tag{78}$$

We evaluate $\mathbb{E}_x[f_{ijk}(x) X_{ij}]$ using the second-order Stein identity. For any $f \in C^2(\mathbb{R}^{dL})$ whose derivatives up to second order have at most polynomial growth[4], and for a centered Gaussian vector x with covariance matrix Σ,

$$\mathbb{E}\big[f(x)(xx^T - \Sigma)\big] = \Sigma\, \mathbb{E}[\nabla_x^2 f(x)]\, \Sigma. \tag{79}$$

Taking the (j,i) block of this identity for $\Sigma = \mathcal{C} \otimes I_d$ gives

$$\mathbb{E}_x \left[f_{ijk}(x) \left(x_j x_i^T - \mathcal{C}_{ji} I_d \right) \right] = \sum_{u,v=1}^L \mathcal{C}_{ju} \mathcal{C}_{vi}\, \mathbb{E}_x \left[\nabla_{x_u} \nabla_{x_v} f_{ijk}(x) \right]. \tag{80}$$

Using the definition of X_{ij}:

$$\mathbb{E}_x[f_{ijk}(x) X_{ij}] = \frac{1}{\sqrt{d}} \sum_{u,v=1}^L \left(\mathcal{C}_{ju} \mathcal{C}_{vi} + \mathcal{C}_{iu} \mathcal{C}_{vj} \right) \mathbb{E}_x \left[\nabla_{x_u} \nabla_{x_v} f_{ijk}(x) \right]. \tag{81}$$

We next compute the Hessian of f_{ijk}. Introduce the shorthand $a = (i,j,k)$ and let $b = (i',j',l)$ and $c = (i'',j'',m)$ as coordinates of the tensor G. By the chain rule,

$$\nabla_{x_u} \nabla_{x_v} f_a = \sum_b \partial_{a,b}^2 \mathcal{L}(G)\, \nabla_{x_u} \nabla_{x_v} G_b + R_{uv,a}, \tag{82}$$

[4]This is satisfied under Assumption 2.1.

where

$$R_{uv,a} := \sum_{b,c} \partial^3_{a,b,c}\mathcal{L}(G)\, \nabla_{x_u} G_b \left(\nabla_{x_v} G_c\right)^T . \tag{83}$$

For $b = (i', j', l)$,

$$\nabla_{x_u} G_{i'j'l} = \frac{1}{\sqrt{d}} \left(\delta_{ui'} S_l x_{j'} + \delta_{uj'} S_l x_{i'}\right), \tag{84}$$

and

$$\nabla_{x_u}\nabla_{x_v} G_{i'j'l} = \frac{1}{\sqrt{d}} S_l \left(\delta_{ui'}\delta_{vj'} + \delta_{uj'}\delta_{vi'}\right). \tag{85}$$

Substituting the first term in (82) into (81), and then using (85), gives

$$\begin{aligned}\frac{1}{\sqrt{d}} \sum_{u,v=1}^{L} \left(\mathcal{C}_{ju}\mathcal{C}_{vi} + \mathcal{C}_{iu}\mathcal{C}_{vj}\right) \sum_{i',j',l} \mathbb{E}_x \left[\partial^2_{(ijk),(i'j'l)}\mathcal{L}(G)\right] \nabla_{x_u}\nabla_{x_v} G_{i'j'l} \\ = \frac{2}{d} \sum_{i',j',l} \left(\mathcal{C}_{ii'}\mathcal{C}_{jj'} + \mathcal{C}_{ij'}\mathcal{C}_{ji'}\right) \mathbb{E}_x \left[\partial^2_{(ijk),(i'j'l)}\mathcal{L}(G)\right] S_l.\end{aligned} \tag{86}$$

The factor 2 comes from the two Kronecker-delta contributions in (85).

It remains to control the remainder $R_{uv,a}$. Under the spectral bound $\max_l \|W_l\|_{\mathrm{op}} \le C_*$, we have

$$\|S_l\|_{\mathrm{op}} \le C_*^2, \qquad \|S_l\|_F \le \sqrt{d}\, C_*^2. \tag{87}$$

Consequently, each G_{ijk} satisfies

$$\sup_d \max_{i,j,k} \mathbb{E}_x |G_{ijk}|^p \le C_p(C_*) \tag{88}$$

for every fixed $p < \infty$. Since L and K are fixed, the same is true for any polynomial function of the finite-dimensional tensor G. By Assumption 2.1, the third derivatives of $\mathcal{L}$ have at most polynomial growth. Hence, for every collection of indices,

$$\mathbb{E}_x \left|\partial^3_{a,b,c}\mathcal{L}(G)\right|^2 \le C(C_*). \tag{89}$$

Now fix arbitrary unit vectors $p, q \in \mathbb{R}^d$. Using (83) and (84), every term in $p^T \mathbb{E}_x[R_{uv,a}]q$ is, up to constants depending only on L, K, and $\mathcal{C}$, bounded by $\frac{1}{d}\mathbb{E}_x \left[\left|\partial^3_{a,b,c}\mathcal{L}(G)\right| \left|p^T S_l x_r\right| \left|q^T S_m x_s\right|\right]$ for some r, s, l, m. By Hölder's inequality,

$$\mathbb{E}_x \left[\left|\partial^3_{a,b,c}\mathcal{L}(G)\right| \left|p^T S_l x_r\right| \left|q^T S_m x_s\right|\right] \le \left(\mathbb{E}_x \left|\partial^3_{a,b,c}\mathcal{L}(G)\right|^2\right)^{1/2} \left(\mathbb{E}_x |p^T S_l x_r|^4\right)^{1/4} \left(\mathbb{E}_x |q^T S_m x_s|^4\right)^{1/4}. \tag{90}$$

The first factor is bounded by (89). Since x_r and x_s are Gaussian and $\|S_l\|_{\mathrm{op}}, \|S_m\|_{\mathrm{op}} \le C_*^2$, the remaining two factors are also uniformly bounded. Therefore,

$$\left\|\mathbb{E}_x[R_{uv,a}]\right\|_{\mathrm{op}} \le \frac{C(C_*)}{d}. \tag{91}$$

Combining (81), (86), and (91), we obtain

$$\mathbb{E}_x[f_{ijk}(x)X_{ij}] = \frac{2}{d} \sum_{i',j',l} \left(\mathcal{C}_{ii'}\mathcal{C}_{jj'} + \mathcal{C}_{ij'}\mathcal{C}_{ji'}\right) \mathbb{E}_x \left[\partial^2_{(ijk),(i'j'l)}\mathcal{L}(G)\right] S_l + E_{ijk}, \tag{92}$$

where, uniformly over (74),

$$\|E_{ijk}\|_{\mathrm{op}} \le C(C_*) d^{-3/2}. \tag{93}$$

Substituting (92) into (78), and defining H_k as

$$H_k^{(n)} := 2 \sum_{i,j,i',j'=1}^{L} \sum_{l=1}^{K} \left(\mathcal{C}_{ii'}\mathcal{C}_{jj'} + \mathcal{C}_{ij'}\mathcal{C}_{ji'}\right) \mathbb{E}_x \left[\partial^2_{(ijk),(i'j'l)}\mathcal{L}(G^{(n)})\right] S_l^{(n)}, \tag{94}$$

gives

$$\bar{g}_k = \frac{1}{d} H_k W_k + \mathcal{E}_{\mathrm{Stein},k}, \tag{95}$$

where $\mathcal{E}_{\mathrm{Stein},k} := \sum_{i,j=1}^{L} E_{ijk} W_k$. By (74) and (93) we have

$$\|\mathcal{E}_{\mathrm{Stein},k}\|_{\mathrm{op}} \le C_E(C_*) d^{-3/2}. \tag{96}$$

Finally, Assumption 2.1 and the uniform moment bound (88) imply

$$\sup_{i,j,k,i',j',l} \mathbb{E}_x \left| \partial^2_{(ijk),(i'j'l)} \mathcal{L}(G) \right| \le C(C_*). \tag{97}$$

Using (87), together with the fact that K and L are fixed, we therefore obtain

$$\|H_k\|_{\mathrm{op}} \le C_H(C_*). \tag{98}$$

Restoring the superscript (n) completes the proof. □

To ensure the trajectory does not blow up, we must prove that the spectral norm of the weights remains uniformly bounded across all training steps.

Lemma 3. *Under the conditions of Theorem 2, there exist constants $C_* > C_0$ independent of d and T, such that for every fixed $T > 0$,*

$$\lim_{d\to\infty} \mathbb{P}\left(\sup_{0\le n\le N_T} \max_{1\le k\le K} \|W_k^{(n)}\|_{\mathrm{op}} \le C_* \right) = 1, \tag{99}$$

where

$$N_T := \left\lfloor \frac{T}{\tau_d} \right\rfloor = \left\lfloor \frac{Td}{4\alpha_d} \right\rfloor. \tag{100}$$

Proof. The weights with $k \notin \mathcal{K}$ remain fixed throughout training and satisfy $\|W_k^{(n)}\|_{\mathrm{op}} \le C_0$. It therefore suffices to control $k \in \mathcal{K}$.

Fix a constant $C_* > C_0$, to be chosen later, and define the stopping time

$$J := \inf\left\{ n \ge 0 : \max_{1\le k\le K} \|W_k^{(n)}\|_{\mathrm{op}} > C_* \right\}. \tag{101}$$

Since $\max_k \|W_k^{(0)}\|_{\mathrm{op}} \le C_0 < C_*$ almost surely, $J \ge 1$. On the event $\{n < J\}$,

$$\max_k \|W_k^{(n)}\|_{\mathrm{op}} \le C_*, \qquad \max_k \|S_k^{(n)}\|_{\mathrm{op}} \le C_*^2. \tag{102}$$

Step 1: Truncation of the fresh samples. For every step n, define

$$\mathcal{A}_n := \left\{ \|x^{(n)}\|^2 \le c_x d \right\} \cap \left\{ \max_{i,j,k} |G_{ijk}^{(n)}| \le c_G \log d \right\}, \tag{103}$$

where c_x and c_G are sufficiently large constants independent of d and n.

Conditional on $\mathcal{F}_n$, the state $W^{(n)}$ is deterministic. On $\{n < J\}$, (102) holds. Standard Gaussian concentration gives

$$\mathbf{1}_{\{n<J\}} \mathbb{P}\left(\|x^{(n)}\|^2 > c_x d \,\middle|\, \mathcal{F}_n \right) \le e^{-c_1 d}. \tag{104}$$

Moreover, conditional on $\mathcal{F}_n$, each $G_{ijk}^{(n)}$ is a centered quadratic form of $x^{(n)}$. To make it explicit, let $x^{(n)} = (\mathcal{C}^{1/2} \otimes I_d) z^{(n)}$, where $z^{(n)} \sim \mathcal{N}(0, I_{Ld})$ is independent of $\mathcal{F}_n$. For $i, j \in \{1, \dots, L\}$, define $B_{ij} := \frac{1}{2}(e_i e_j^T + e_j e_i^T)$, where $\{e_i\}_{i=1}^L$ is the canonical basis of $\mathbb{R}^L$. Then from (76) we can write

$$G_{ijk}^{(n)} = (z^{(n)})^T A_{ijk}^{(n)} z^{(n)} - \mathrm{Tr}(A_{ijk}^{(n)}), \tag{105}$$

where

$$A_{ijk}^{(n)} := \frac{1}{\sqrt{d}} (\mathcal{C}^{1/2} \otimes I_d)(B_{ij} \otimes S_k^{(n)})(\mathcal{C}^{1/2} \otimes I_d). \tag{106}$$

Notice that $A_{ijk}^{(n)}$ is an $Ld \times Ld$ symmetric matrix. On the event $\{n < J\}$,

$$\|S_k^{(n)}\|_{\mathrm{op}} \le C_*^2, \qquad \|S_k^{(n)}\|_F \le \sqrt{d}\, C_*^2. \tag{107}$$

Since $L = \Theta(1)$ and $\|B_{ij}\|_{\mathrm{op}} \le 1$, $\|B_{ij}\|_F \le 1$, it follows that

$$\|A_{ijk}^{(n)}\|_{\mathrm{op}} \le \frac{\|\mathcal{C}\|_{\mathrm{op}}}{\sqrt{d}} \|B_{ij}\|_{\mathrm{op}} \|S_k^{(n)}\|_{\mathrm{op}} \le \frac{\|\mathcal{C}\|_{\mathrm{op}} C_*^2}{\sqrt{d}}, \tag{108}$$

$$\|A_{ijk}^{(n)}\|_F \le \frac{\|\mathcal{C}\|_{\mathrm{op}}}{\sqrt{d}} \|B_{ij} \otimes S_k^{(n)}\|_F = \frac{\|\mathcal{C}\|_{\mathrm{op}}}{\sqrt{d}} \|B_{ij}\|_F \|S_k^{(n)}\|_F \le \|\mathcal{C}\|_{\mathrm{op}} C_*^2. \tag{109}$$

Therefore, uniformly on $\{n < J\}$,

$$\|A_{ijk}^{(n)}\|_F \le C, \qquad \|A_{ijk}^{(n)}\|_{\mathrm{op}} \le \frac{C}{\sqrt{d}}, \tag{110}$$

where C depends only on C_* and $\mathcal{C}$.

Applying the Hanson–Wright inequality to (105), we obtain

$$\mathbf{1}_{\{n<J\}} \mathbb{P}\left(|G_{ijk}^{(n)}| > t \,\middle|\, \mathcal{F}_n \right) \le 2 \exp\left[-c \min\left\{ \frac{t^2}{\|A_{ijk}^{(n)}\|_F^2}, \frac{t}{\|A_{ijk}^{(n)}\|_{\mathrm{op}}} \right\} \right]. \tag{111}$$

Hence, taking $t = c_G \log d$ and using (110),

$$\mathbf{1}_{\{n<J\}} \mathbb{P}\left(|G_{ijk}^{(n)}| > c_G \log d \,\middle|\, \mathcal{F}_n \right) \le 2 \exp\left[-c \min\left\{ (\log d)^2, \sqrt{d}\, \log d \right\} \right] \le 2 e^{-c' (\log d)^2}. \tag{112}$$

Since K and L are fixed, a union bound over (i, j, k) gives

$$\mathbf{1}_{\{n<J\}} \mathbb{P}\left(\max_{i,j,k} |G_{ijk}^{(n)}| > c_G \log d \,\middle|\, \mathcal{F}_n \right) \le 2L^2 K\, e^{-c' (\log d)^2}. \tag{113}$$

Therefore there exists $p_d \le C\left(e^{-c_1 d} + e^{-c_2 (\log d)^2} \right)$ such that

$$\mathbf{1}_{\{n<J\}} \mathbb{P}(\mathcal{A}_n^c \mid \mathcal{F}_n) \le p_d. \tag{114}$$

Define the event that no truncation failure occurs before the stopping time,

$$\mathcal{A}_{\mathrm{data}} := \bigcap_{n=0}^{N_T - 1} \left(\{n \ge J\} \cup \mathcal{A}_n \right). \tag{115}$$

Using (114),

$$\mathbb{P}(\mathcal{A}_{\mathrm{data}}^c) \le \sum_{n=0}^{N_T - 1} \mathbb{E}\left[\mathbf{1}_{\{n<J\}} \mathbb{P}(\mathcal{A}_n^c \mid \mathcal{F}_n) \right] \le N_T p_d. \tag{116}$$

Since $\alpha_d = \Omega(d^{-\iota})$, $N_T \le C_T d^{1+\iota}$, and hence

$$\mathbb{P}(\mathcal{A}_{\mathrm{data}}^c) \to 0. \tag{117}$$

Step 2: Stopped martingale decomposition and its bounds. For $k \in \mathcal{K}$, define the stopped and truncated stochastic gradient

$$\widehat{g}_k^{(n)} := \mathbf{1}_{\{n<J\}} g_k^{(n)} \mathbf{1}_{\mathcal{A}_n}, \tag{118}$$

its conditional mean

$$\widehat{\bar{g}}_k^{(n)} := \mathbb{E}[\widehat{g}_k^{(n)} \mid \mathcal{F}_n], \tag{119}$$

and the martingale difference

$$\widehat{\xi}_k^{(n)} := \widehat{g}_k^{(n)} - \widehat{\bar{g}}_k^{(n)}. \tag{120}$$

Then $\mathbb{E}[\widehat{\xi}_k^{(n)} \mid \mathcal{F}_n] = 0$.

Let

$$\bar{g}_k^{(n)} := \mathbb{E}[g_k^{(n)} \mid \mathcal{F}_n]. \tag{121}$$

Since $\{n < J\} \in \mathcal{F}_n$,

$$\widehat{\bar{g}}_k^{(n)} = \mathbf{1}_{\{n<J\}} \left(\bar{g}_k^{(n)} - \mathcal{T}_k^{(n)} \right), \tag{122}$$

where

$$\mathcal{T}_k^{(n)} := \mathbb{E}\left[g_k^{(n)} \mathbf{1}_{\mathcal{A}_n^c} \,\middle|\, \mathcal{F}_n \right]. \tag{123}$$

We next record a uniform second-moment estimate for the stochastic gradient. On $\{n < J\}$,

$$\max\left\{ \left\| \mathbb{E}[g_k^{(n)} (g_k^{(n)})^T \mid \mathcal{F}_n] \right\|_{\mathrm{op}}, \left\| \mathbb{E}[(g_k^{(n)})^T g_k^{(n)} \mid \mathcal{F}_n] \right\|_{\mathrm{op}} \right\} \le C_g(C_*). \tag{124}$$

Indeed, for any unit vectors $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^{d_k}$, the two quadratic forms in (124) are respectively bounded by finite sums of expectations of the form

$$\mathbb{E}\left[|\partial_{(ijk)} \mathcal{L}(G^{(n)})|^2 \, \|(W_k^{(n)})^T (X_{ij}^{(n)})^T u\|^2 \,\middle|\, \mathcal{F}_n \right]$$

and

$$\mathbb{E}\left[|\partial_{(ijk)} \mathcal{L}(G^{(n)})|^2 \, \|X_{ij}^{(n)} W_k^{(n)} v\|^2 \,\middle|\, \mathcal{F}_n \right].$$

Similarly to the estimates used in the proof of Lemma 2, Assumption 2.1 together with the spectral bound (102) implies uniform bounds on all fixed moments of $G^{(n)}$ and hence on the corresponding polynomially growing derivatives of $\mathcal{L}$. In particular, by Hölder's inequality, for every unit vector $u \in \mathbb{R}^d$,

$$\begin{aligned} &\mathbb{E}\left[|\partial_{(ijk)} \mathcal{L}(G^{(n)})|^2 \, \|(W_k^{(n)})^T (X_{ij}^{(n)})^T u\|^2 \,\middle|\, \mathcal{F}_n \right] \\ &\quad \le \left(\mathbb{E}\left[|\partial_{(ijk)} \mathcal{L}(G^{(n)})|^4 \,\middle|\, \mathcal{F}_n \right] \right)^{1/2} \left(\mathbb{E}\left[\|(W_k^{(n)})^T (X_{ij}^{(n)})^T u\|^4 \,\middle|\, \mathcal{F}_n \right] \right)^{1/2} = \mathcal{O}(1), \end{aligned} \tag{125}$$

uniformly on $\{n < J\}$. The same argument applies to the other quadratic variation in (124).

Consequently,

$$\max\left\{ \left\| \mathbb{E}[\widehat{\xi}_k^{(n)} (\widehat{\xi}_k^{(n)})^T \mid \mathcal{F}_n] \right\|_{\mathrm{op}}, \left\| \mathbb{E}[(\widehat{\xi}_k^{(n)})^T \widehat{\xi}_k^{(n)} \mid \mathcal{F}_n] \right\|_{\mathrm{op}} \right\} \le C_g(C_*). \tag{126}$$

On $\mathcal{A}_n \cap \{n < J\}$, $\|X_{ij}^{(n)}\|_{\mathrm{op}} \le C\sqrt{d}$, the first derivatives of $\mathcal{L}$ have polynomial growth and $\max_{i,j,k} |G_{ijk}^{(n)}| \le c_G \log d$. Therefore, for some fixed integer r,

$$\|\widehat{g}_k^{(n)}\|_{\mathrm{op}} \le C\sqrt{d}\,(1 + (\log d)^r) =: B_d. \tag{127}$$

The same bound holds for $\|\widehat{\widehat{g}}_k^{(n)}\|_{\mathrm{op}}$, and hence

$$\|\widehat{\xi}_k^{(n)}\|_{\mathrm{op}} \le 2B_d. \tag{128}$$

Step 3: Uniform control of the martingale noise. Set

$$a_d := 1 - \frac{\alpha_d \gamma}{d} = 1 - \frac{\tau_d \gamma}{4}. \tag{129}$$

For sufficiently large d, $a_d \in (0,1)$. Define

$$D_{k,n} := -\alpha_d \widehat{\xi}_k^{(n)}. \tag{130}$$

For every terminal time $m \ge 1$, consider the discounted martingale sum

$$Y_{k,m} := \sum_{n=0}^{m-1} a_d^{m-1-n} D_{k,n}. \tag{131}$$

For each fixed m, the summands in (131) form a rectangular matrix martingale difference sequence. By (128),

$$\|a_d^{m-1-n} D_{k,n}\|_{\mathrm{op}} \le 2\alpha_d B_d =: R_d. \tag{132}$$

Moreover, by (126), the predictable quadratic variation in both rectangular directions is bounded by

$$\sigma_d^2 \le C_g \alpha_d^2 \sum_{r=0}^{\infty} a_d^{2r} = \frac{C_g \alpha_d^2}{1 - a_d^2} \le C \frac{d\alpha_d}{\gamma}. \tag{133}$$

Under $\alpha_d \le \frac{c_0}{d \log d}$, we have

$$\sigma_d^2 \le \frac{C c_0}{\gamma \log d}, \qquad R_d = o((\log d)^{-1}). \tag{134}$$

The Matrix Freedman inequality therefore gives for every fixed $\delta > 0$,

$$\mathbb{P}(\|Y_{k,m}\|_{\mathrm{op}} \ge \delta) \le (d + d_k) \exp\left[-\frac{\delta^2/2}{\sigma_d^2 + R_d \delta/3}\right]. \tag{135}$$

For sufficiently large d, (134) implies

$$\mathbb{P}(\|Y_{k,m}\|_{\mathrm{op}} \ge \delta) \le (d + d_k) d^{-c\delta^2/c_0} \tag{136}$$

for some constant $c > 0$ independent of m, T, and d. Choose $\delta := \frac{C_* - C_0}{4} > 0$ and set $\beta := \max\{1, c_w\}$. Since $\max_k d_k \le d^{c_w}$, for all sufficiently large d, $d + d_k \le 2d^\beta$. Moreover, since $\alpha_d = \Omega(d^{-\iota})$, we have $N_T \le C_T d^{1+\iota}$. Therefore, by (136) and a union bound over $k \in \mathcal{K}$ and $1 \le m \le N_T$,

$$\mathbb{P}\left(\max_{k \in \mathcal{K}} \max_{1 \le m \le N_T} \|Y_{k,m}\|_{\mathrm{op}} > \delta\right) \le 2K C_T d^{1+\iota+\beta-c\delta^2/c_0}. \tag{137}$$

Choose $c_0 > 0$ sufficiently small so that

$$\frac{c\delta^2}{c_0} > 1 + \iota + \beta. \tag{138}$$

Then the exponent in (137) is negative, and hence

$$\mathbb{P}\left(\max_{k \in \mathcal{K}} \max_{1 \le m \le N_T} \|Y_{k,m}\|_{\mathrm{op}} > \delta\right) \to 0. \tag{139}$$

Denote the complementary high-probability event by $\mathcal{A}_{\mathrm{noise}}$.

We also control the truncation tail (123). On $\{n < J\}$, the stochastic gradient $g_k^{(n)}$ has at most polynomial growth in $x^{(n)}$. Hence, for some finite constant q,

$$\mathbb{E}\left[\|g_k^{(n)}\|_{\mathrm{op}}^2 \,\middle|\, \mathcal{F}_n\right] \le d^q. \tag{140}$$

Combining Cauchy–Schwarz with (114),

$$\mathbf{1}_{\{n<J\}} \|\mathcal{T}_k^{(n)}\|_{\mathrm{op}} \le \mathbf{1}_{\{n<J\}} \left(\mathbb{E}[\|g_k^{(n)}\|_{\mathrm{op}}^2 \mid \mathcal{F}_n]\right)^{1/2} \left(\mathbb{P}(\mathcal{A}_n^c \mid \mathcal{F}_n)\right)^{1/2} \le r_d, \tag{141}$$

where

$$r_d \le d^{q/2} p_d^{1/2} = o(d^{-M}) \tag{142}$$

for every fixed $M > 0$.

Step 4: Closing the stopping-time argument. On $\mathcal{A}_{\mathrm{data}}$, for every $m \le J$ and every $n < m$, we have $n < J$ and $\mathcal{A}_n$ occurs. Hence

$$g_k^{(n)} = \widehat{g}_k^{(n)} = \bar{g}_k^{(n)} - \mathcal{T}_k^{(n)} + \widehat{\xi}_k^{(n)}. \tag{143}$$

The SGD (6) therefore gives, for $m \le J$,

$$W_k^{(m)} = a_d^m W_k^{(0)} - \alpha_d \sum_{n=0}^{m-1} a_d^{m-1-n} \bar{g}_k^{(n)} + \alpha_d \sum_{n=0}^{m-1} a_d^{m-1-n} \mathcal{T}_k^{(n)} + Y_{k,m}. \tag{144}$$

By Lemma 2, uniformly on $\{n < J\}$,

$$\bar{g}_k^{(n)} = \frac{1}{d} H_k^{(n)} W_k^{(n)} + \mathcal{E}_{\mathrm{Stein},k}^{(n)}, \tag{145}$$

with $\|H_k^{(n)}\|_{\mathrm{op}} \le C_H(C_*)$ and $\|\mathcal{E}_{\mathrm{Stein},k}^{(n)}\|_{\mathrm{op}} \le C_E(C_*) d^{-3/2}$. Therefore,

$$\|\bar{g}_k^{(n)}\|_{\mathrm{op}} \le \frac{C_H(C_*) C_*}{d} + C_E(C_*) d^{-3/2}. \tag{146}$$

Since $1 - a_d = \frac{\alpha_d \gamma}{d}$, we have

$$\alpha_d \sum_{r=0}^{\infty} a_d^r = \frac{d}{\gamma}. \tag{147}$$

Combining (144)–(147), on $\mathcal{A}_{\text{data}} \cap \mathcal{A}_{\text{noise}}$ and for every $m \le J \wedge N_T$,

$$\|W_k^{(m)}\|_{\text{op}} \le \|W_k^{(0)}\|_{\text{op}} + \frac{C_H(C_*)C_*}{\gamma} + \frac{C_E(C_*)}{\gamma\sqrt{d}} + \frac{dr_d}{\gamma} + \delta \le C_0 + \frac{C_H(C_*)C_*}{\gamma} + \delta + o(1). \tag{148}$$

We now fix $C_* := 2C_0 + 1$ and choose γ sufficiently large (by Assumption 2.4) so that

$$\frac{C_H(C_*)C_*}{\gamma} \le \frac{C_* - C_0}{4}. \tag{149}$$

For all sufficiently large d, (148) then gives

$$\max_k \|W_k^{(m)}\|_{\text{op}} < C_* \qquad \text{for every } m \le J \wedge N_T \tag{150}$$

on $\mathcal{A}_{\text{data}} \cap \mathcal{A}_{\text{noise}}$. Suppose, on this event, that $J \le N_T$. Taking $m = J$ in (148) yields

$$\max_k \|W_k^{(J)}\|_{\text{op}} < C_*, \tag{151}$$

contradicting the definition (101). Therefore, $J > N_T$ on $\mathcal{A}_{\text{data}} \cap \mathcal{A}_{\text{noise}}$ for all sufficiently large d. Finally,

$$\mathbb{P}(J \le N_T) \le \mathbb{P}(\mathcal{A}_{\text{data}}^c) + \mathbb{P}(\mathcal{A}_{\text{noise}}^c) \to 0 \tag{152}$$

by (117) and (139). Hence

$$\mathbb{P}\left(\sup_{0 \le n \le N_T} \max_k \|W_k^{(n)}\|_{\text{op}} \le C_* \right) \longrightarrow 1, \tag{153}$$

which proves the lemma. □

D.1.3 Moment expansion

Let C_* be the constant in Lemma 3, and fix $T > 0$ and $\rho > C_*^2$. Denote

$$\mathcal{E}_d(T) := \left\{ \sup_{0 \le n \le N_T} \max_{1 \le k \le K} \|W_k^{(n)}\|_{\text{op}} \le C_* \right\}, \qquad N_T := \left\lfloor \frac{T}{\tau_d} \right\rfloor. \tag{154}$$

Then $\mathbb{P}(\mathcal{E}_d(T)) \to 1$ by Lemma 3.

For the compactness argument later on, it is convenient to introduce the separable closed subspace

$$X_\rho^0 := \left\{ \mu \in X_\rho : \lim_{w \to \infty} \sup_{|\alpha| = w} \frac{|\mu_\alpha|}{\rho^w} = 0 \right\}. \tag{155}$$

Equipped with the norm inherited from X_ρ, X_ρ^0 is a separable Banach space.

Since the empirical moment sequence need not belong to X_ρ^0 on the exceptional event $\mathcal{E}_d(T)^c$, we introduce an auxiliary spectrally bounded process. For the fixed time horizon T, define

$$\check{W}_k^{(n)} := \begin{cases} W_k^{(n)}, & \text{on } \mathcal{E}_d(T), \\ W_k^{(0)}, & \text{on } \mathcal{E}_d(T)^c, \end{cases} \qquad 0 \le n \le N_T, \tag{156}$$

and let

$$\check{S}_k^{(n)} := \check{W}_k^{(n)} (\check{W}_k^{(n)})^T. \tag{157}$$

For a multi-index $\alpha = (k_1, \dots, k_w)$, define

$$\check{\mu}_\alpha^{(n)} := \frac{1}{d} \text{Tr}\left[\check{S}_{k_1}^{(n)} \cdots \check{S}_{k_w}^{(n)} \right], \tag{158}$$

and let $\check{\mu}^{(d)}(t)$ denote its piecewise-constant interpolation,

$$\check{\mu}^{(d)}(t) = \check{\mu}^{(n)}, \qquad t \in [n\tau_d, (n+1)\tau_d). \tag{159}$$

Since $C_* > C_0$, we have almost surely

$$\sup_{0\le n\le N_T} \max_{1\le k\le K} \|\check{W}_k^{(n)}\|_{\mathrm{op}} \le C_*. \tag{160}$$

Consequently, for every $|\alpha| = w$,

$$|\check{\mu}_\alpha^{(d)}(t)| \le C_*^{2w}, \qquad t \in [0,T], \tag{161}$$

and hence, almost surely (w.r.t. the initialization),

$$\check{\mu}^{(d)}(t) \in X_\rho^0 \qquad \text{for all } t \in [0,T]. \tag{162}$$

Moreover,

$$\mathbb{P}\left(\check{\mu}^{(d)}(t) = \tilde{\mu}^{(d)}(t) \text{ for all } t \in [0,T]\right) \ge \mathbb{P}(\mathcal{E}_d(T)) \longrightarrow 1. \tag{163}$$

All compactness arguments below will be carried out for $\check{\mu}^{(d)}$, and the resulting convergence will subsequently be transferred to the original empirical trajectory $\tilde{\mu}^{(d)}$ using (163).

We define empirical vector field $V_d(\tilde{\mu}^{(d)}(t))$ as

$$[V_d(\tilde{\mu}^{(d)})]_\alpha = \frac{1}{d} \sum_{\substack{r=1 \\ k_r \in \mathcal{K}}}^{w} \mathrm{Tr}\left[S_{k_1}^{(n)} \dots A_{k_r}^{(n)} \dots S_{k_w}^{(n)}\right] \tag{164}$$

where $A_k^{(n)} := -\frac{1}{4}\left(M_k^{(n)} S_k^{(n)} + S_k^{(n)} (M_k^{(n)})^T\right)$ and $M_k^{(n)} := H_k^{(n)} + \gamma I_d$ with $H_k^{(n)}$ defined in Lemma 2.

The following lemma establishes the integral equation for the empirical trajectory (7).

Lemma 4. *For every $t \in [0,T]$, the auxiliary process satisfies*

$$\check{\mu}^{(d)}(t) = \check{\mu}^{(d)}(0) + \mathbf{1}_{\mathcal{E}_d(T)} \int_0^t V_d(\tilde{\mu}^{(d)}(s))\, ds + \check{E}_{\mathrm{total}}^{(d)}(t). \tag{165}$$

For every $\rho > C_*^2$, all terms in (165) *belong to X_ρ^0 almost surely, and*

$$\sup_{t\in[0,T]} \left\| \check{E}_{\mathrm{total}}^{(d)}(t) \right\|_\rho \xrightarrow{\mathbb{P}} 0. \tag{166}$$

Proof. Define

$$J := \inf\left\{ n \ge 0 : \max_{1\le k\le K} \|W_k^{(n)}\|_{\mathrm{op}} > C_* \right\}. \tag{167}$$

By Lemma 3,

$$\mathbb{P}(J > N_T) \to 1. \tag{168}$$

All estimates below are uniform on $\{n < J\}$.

Step 1: One-step expansion of the attention matrices. For $k \in \mathcal{K}$, Lemma 2 gives

$$\bar{g}_k^{(n)} = \frac{1}{d} H_k^{(n)} W_k^{(n)} + \mathcal{E}_{\mathrm{Stein},k}^{(n)}, \tag{169}$$

where $\|H_k^{(n)}\|_{\mathrm{op}} \le C_H$ and $\|\mathcal{E}_{\mathrm{Stein},k}^{(n)}\|_{\mathrm{op}} \le C_E d^{-3/2}$, uniformly on $\{n < J\}$. Define

$$\xi_k^{(n)} := g_k^{(n)} - \bar{g}_k^{(n)}, \qquad M_k^{(n)} := H_k^{(n)} + \gamma I_d. \tag{170}$$

Then $\mathbb{E}[\xi_k^{(n)} \mid \mathcal{F}_n] = 0$, and the SGD (6) can be written as

$$W_k^{(n+1)} = W_k^{(n)} - \frac{\tau_d}{4} M_k^{(n)} W_k^{(n)} - \alpha_d \xi_k^{(n)} - \alpha_d \mathcal{E}_{\mathrm{Stein},k}^{(n)}. \tag{171}$$

For $k \notin \mathcal{K}$, $W_k^{(n+1)} = W_k^{(n)}$. Let

$$\delta W_k^{(n)} := W_k^{(n+1)} - W_k^{(n)}. \tag{172}$$

Expanding $S_k^{(n+1)} = \left(W_k^{(n)} + \delta W_k^{(n)}\right)\left(W_k^{(n)} + \delta W_k^{(n)}\right)^T$ gives

$$S_k^{(n+1)} = S_k^{(n)} + \tau_d A_k^{(n)} + \alpha_d B_k^{(n)} + C_k^{(n)}, \tag{173}$$

where, for $k \in \mathcal{K}$,

$$A_k^{(n)} := -\frac{1}{4}\left(M_k^{(n)} S_k^{(n)} + S_k^{(n)} (M_k^{(n)})^T\right), \tag{174}$$

$$B_k^{(n)} := -\left(\xi_k^{(n)} (W_k^{(n)})^T + W_k^{(n)} (\xi_k^{(n)})^T\right), \tag{175}$$

and

$$C_k^{(n)} := -\alpha_d \left(\mathcal{E}_{\mathrm{Stein},k}^{(n)} (W_k^{(n)})^T + W_k^{(n)} (\mathcal{E}_{\mathrm{Stein},k}^{(n)})^T\right) + \delta W_k^{(n)} (\delta W_k^{(n)})^T. \tag{176}$$

For $k \notin \mathcal{K}$, we set

$$A_k^{(n)} = B_k^{(n)} = C_k^{(n)} = 0. \tag{177}$$

On $\{n < J\}$,

$$\|S_k^{(n)}\|_{\mathrm{op}} \le C_*^2, \qquad \|M_k^{(n)}\|_{\mathrm{op}} \le C_M \tag{178}$$

for a constant C_M independent of d and n. In particular,

$$\|A_k^{(n)}\|_{\mathrm{op}} \le C_A. \tag{179}$$

We shall also use moment bounds for the stochastic gradient noise. Similarly to the estimates in Lemma 3 we can write

$$\mathbf{1}_{\{n<J\}} \mathbb{E}\left[\|\xi_k^{(n)}\|_F^p \,\middle|\, \mathcal{F}_n\right] \le C_p d^{p/2}. \tag{180}$$

Fix a multi-index $\alpha = (k_1, \dots, k_w)$. Using (173) in every factor of $\mu_\alpha^{(n+1)} = \frac{1}{d}\mathrm{Tr}\left[S_{k_1}^{(n+1)} \cdots S_{k_w}^{(n+1)}\right]$, we separate the expansion into four classes.

The zeroth-order term is $\mu_\alpha^{(n)}$. The terms containing exactly one factor $\tau_d A_{k_r}^{(n)}$ define the empirical drift

$$[V_d(\mu^{(n)})]_\alpha := \sum_{\substack{1 \le r \le w \\ k_r \in \mathcal{K}}} \frac{1}{d}\mathrm{Tr}\left[S_{k_1}^{(n)} \cdots A_{k_r}^{(n)} \cdots S_{k_w}^{(n)}\right]. \tag{181}$$

The terms containing exactly one factor $\alpha_d B_{k_r}^{(n)}$ define the martingale

$$\eta_\alpha^{(n)} := \sum_{\substack{1 \le r \le w \\ k_r \in \mathcal{K}}} \frac{1}{d}\mathrm{Tr}\left[S_{k_1}^{(n)} \cdots B_{k_r}^{(n)} \cdots S_{k_w}^{(n)}\right]. \tag{182}$$

Since all $S_k^{(n)}$ are $\mathcal{F}_n$-measurable and $B_k^{(n)}$ is linear in $\xi_k^{(n)}$, $\mathbb{E}[\eta_\alpha^{(n)} \mid \mathcal{F}_n] = 0$.

Finally, let $R_\alpha^{(n)}$ denote the sum of all remaining terms in the expansion. Thus $R_\alpha^{(n)}$ contains, in particular,

- every term involving at least one $C_{k_r}^{(n)}$;
- every term involving at least two first-order increments chosen from $\{\tau A_{k_r}^{(n)}, \alpha B_{k_r}^{(n)}\}$.

With this definition we obtain the expansion:

$$\mu_\alpha^{(n+1)} - \mu_\alpha^{(n)} = \tau_d [V_d(\mu^{(n)})]_\alpha + \alpha_d \eta_\alpha^{(n)} + R_\alpha^{(n)}. \tag{183}$$

Step 2: Vanishing of the remainder. We next bound the remainder. For every fixed multi-index α, there exists a constant $C_\alpha < \infty$ such that, uniformly on $\{n < J\}$,

$$\mathbb{E}\left[|R_\alpha^{(n)}| \,\middle|\, \mathcal{F}_n\right] \le C_\alpha \left(\alpha_d^2 + \alpha_d\tau_d + \tau_d^2 + \alpha_d d^{-3/2}\right). \tag{184}$$

To see this, first bound terms containing at least two B-factors, let $P_0, P_1, \cdots$ denote arbitrary products of the remaining S- and A-matrices. On $\{n < J\}$, their operator norms are bounded by a constant depending only on α. We obtain, for fixed $r_B \ge 2$,

$$\frac{1}{d}\mathbb{E}\left[|\mathrm{Tr}(P_0 B_1 P_1 \cdots B_{r_B} P_{r_B})| \,|\, \mathcal{F}_n\right] \le C_\alpha d^{(r_B-2)/2}. \tag{185}$$

Hence the corresponding contribution is bounded by

$$C_\alpha \alpha_d^{r_B} \tau_d^{r_A} d^{(r_B-2)/2} \le C_\alpha \alpha_d^2 \tag{186}$$

for all sufficiently large d, because $\alpha_d\sqrt{d} \to 0$. A term containing exactly one $\alpha_d B$ and at least one $\tau_d A$ is bounded by $C_\alpha\alpha_d\tau_d$, while a term containing at least two $\tau_d A$ factors is bounded by $C_\alpha\tau_d^2$.

It remains to consider terms containing at least one C-factor. By (176), each $C_k^{(n)}$ is a finite sum of terms of the following types:

$$O_{\mathrm{op}}(\alpha_d d^{-3/2}), \qquad \tau_d^2\, U_1 U_2^T, \qquad \alpha_d\tau_d\, U_1\xi_k^T, \qquad \alpha_d^2\, \xi_k\xi_k^T,$$

together with terms involving $\mathcal{E}_{\mathrm{Stein},k}^{(n)}$ that are of smaller order. Here the U_i have uniformly bounded operator norm and $\mathcal{O}(\sqrt{d})$ Frobenius norm. Applying the same trace inequalities, Hölder's inequality, and (180), every product containing at least one C-factor is bounded in conditional expectation by $C_\alpha\left(\alpha_d^2 + \alpha_d\tau_d + \tau_d^2 + \alpha_d d^{-3/2}\right)$, with any additional stochastic increment producing only an additional factor of order $\alpha_d\sqrt{d} = o(1)$. This proves (184).

Define

$$E_{\mathrm{rem},\alpha}^{(d)}(t) := \sum_{n=0}^{\lfloor t/\tau\rfloor - 1} R_\alpha^{(n)}. \tag{187}$$

Replace $R_\alpha^{(n)}$ temporarily by $R_\alpha^{(n)}\mathbf{1}_{\{n<J\}}$. By (184),

$$\mathbb{E}\left[\sum_{n=0}^{N_T-1} \mathbf{1}_{\{n<J\}}|R_\alpha^{(n)}|\right] \le N_T C_\alpha\left(\alpha_d^2 + \alpha_d\tau_d + \tau_d^2 + \alpha_d d^{-3/2}\right) \le C_{\alpha,T}\left(d\alpha_d + \alpha_d + \frac{\alpha_d}{d} + d^{-1/2}\right). \tag{188}$$

Since $\alpha_d \le \frac{c_0}{d\log d}$, the right-hand side tends to zero. Hence, by Markov's inequality,

$$\sup_{t\in[0,T]}\left|E_{\mathrm{rem},\alpha}^{(d)}(t)\right| \xrightarrow{\mathbb{P}} 0 \tag{189}$$

on $\{J > N_T\}$. Together with (168), this proves (189) for the original trajectory.

Step 3: Vanishing of the first-order martingale noise. Define

$$E_{\mathrm{noise},\alpha}^{(d)}(t) := \sum_{n=0}^{\lfloor t/\tau_d\rfloor - 1} \alpha_d \eta_\alpha^{(n)}. \tag{190}$$

We claim that, for every fixed α,

$$\mathbf{1}_{\{n<J\}}\mathbb{E}\left[(\eta_\alpha^{(n)})^2 \,\middle|\, \mathcal{F}_n\right] \le \frac{C_\alpha}{d^2}. \tag{191}$$

Indeed, every constituent of (182) is a linear functional of $\xi_k^{(n)}$. Since $\xi_k^{(n)} = g_k^{(n)} - \mathbb{E}[g_k^{(n)} \mid \mathcal{F}_n]$, it suffices to bound the second moment of the corresponding linear functional of $g_k^{(n)}$. After inserting $g_k^{(n)} = \sum_{i,j}\partial_{(ijk)}\mathcal{L}(G^{(n)})X_{ij}^{(n)}W_k^{(n)}$, each term takes the form

$$\partial_{(ijk)}\mathcal{L}(G^{(n)})Z_{ij}(A), \qquad Z_{ij}(A) := \frac{1}{d}\mathrm{Tr}\left[AX_{ij}^{(n)}\right], \tag{192}$$

where, for fixed α, $\|A\|_{\mathrm{op}} \le C_\alpha$. Since $X_{ij}^{(n)}$ is a centered Gaussian quadratic-form matrix, standard Gaussian moment estimates give

$$\mathbb{E}\left[|Z_{ij}(A)|^4 \,\middle|\, \mathcal{F}_n\right] \le \frac{C_\alpha}{d^4}. \tag{193}$$

On the other hand, Assumption 2.1 and the uniform moment bounds for $G^{(n)}$ imply

$$\mathbb{E}\left[\left|\partial_{(ijk)}\mathcal{L}(G^{(n)})\right|^4 \,\middle|\, \mathcal{F}_n\right] \le C_\alpha. \tag{194}$$

Hence, by Cauchy–Schwarz,

$$\mathbb{E}\left[\left|\partial_{(ijk)}\mathcal{L}(G^{(n)})Z_{ij}(A)\right|^2 \,\middle|\, \mathcal{F}_n\right] \le \frac{C_\alpha}{d^2}. \tag{195}$$

Since K, L, and the length of α are fixed, this proves (191).

Therefore the stopped process

$$M_{\alpha,m} := \sum_{n=0}^{m-1} \alpha_d \eta_\alpha^{(n)} \mathbf{1}_{\{n<J\}} \tag{196}$$

is a square-integrable martingale with predictable quadratic variation bounded by

$$\mathbb{E}[\langle M_\alpha\rangle_{N_T}] \le N_T \alpha_d^2 \frac{C_\alpha}{d^2} \le C_{\alpha,T}\frac{\alpha_d}{d}. \tag{197}$$

Doob's maximal inequality yields, for every $\varepsilon > 0$,

$$\mathbb{P}\left(\max_{m\le N_T} |M_{\alpha,m}| > \varepsilon\right) \le \frac{C_{\alpha,T}\alpha_d}{\varepsilon^2 d} \to 0 \tag{198}$$

Using again $\mathbb{P}(J > N_T) \to 1$, we conclude that

$$\sup_{t\in[0,T]} \left|E^{(d)}_{\mathrm{noise},\alpha}(t)\right| \xrightarrow{\mathbb{P}} 0. \tag{199}$$

Step 4: From the discrete recursion to the integral equation. For $m(t) := \lfloor \frac{t}{\tau} \rfloor$, summing (183) from $n = 0$ to $m(t) - 1$ gives

$$\tilde\mu_\alpha^{(d)}(t) = \tilde\mu_\alpha^{(d)}(0) + \tau_d \sum_{n=0}^{m(t)-1} [V_d(\mu^{(n)})]_\alpha + E^{(d)}_{\mathrm{noise},\alpha}(t) + E^{(d)}_{\mathrm{rem},\alpha}(t). \tag{200}$$

Since the interpolation (7) is piecewise constant,

$$\int_0^t [V_d(\tilde\mu^{(d)}(s))]_\alpha \, ds = \tau_d \sum_{n=0}^{m(t)-1} [V_d(\mu^{(n)})]_\alpha + (t - m(t)\tau_d)[V_d(\mu^{(m(t))})]_\alpha. \tag{201}$$

Define the discretization error

$$E^{(d)}_{\mathrm{disc},\alpha}(t) := -(t - m(t)\tau_d)[V_d(\mu^{(m(t))})]_\alpha. \tag{202}$$

Thus we define $\check E^{(d)}_{\mathrm{total},\alpha} := \mathbf{1}_{\mathcal{E}_d(T)} E^{(d)}_{\mathrm{total},\alpha}$ with

$$E^{(d)}_{\mathrm{total},\alpha} = E^{(d)}_{\mathrm{noise},\alpha} + E^{(d)}_{\mathrm{rem},\alpha} + E^{(d)}_{\mathrm{disc},\alpha}, \tag{203}$$

and (165) follows coordinatewise.

On $\{n < J\}$, (179) implies that for $|\alpha| = w$,

$$|[V_d(\mu^{(n)})]_\alpha| \le B w C_*^{2w} \tag{204}$$

for a constant B independent of d, n, and w. Consequently, for every fixed α,

$$\sup_{t\in[0,T]} |E^{(d)}_{\mathrm{disc},\alpha}(t)| \le \tau B|\alpha| C_*^{2|\alpha|} \to 0. \tag{205}$$

Combining this with (189) and (199), we have, for every fixed multi-index α,

$$\sup_{t\in[0,T]} |E^{(d)}_{\mathrm{total},\alpha}(t)| \xrightarrow{\mathbb{P}} 0. \tag{206}$$

Step 5: Convergence in X_ρ^0. We now upgrade the coordinatewise convergence to convergence in X_ρ^0 on $\mathcal{E}_d(T)$. On $\mathcal{E}_d(T)$,

$$\sup_{t\in[0,T]}\sup_{|\alpha|=w}|\tilde{\mu}_\alpha^{(d)}(t)|\le C_*^{2w}. \tag{207}$$

Together with (204) and

$$E_{\text{total}}^{(d)}(t)=\tilde{\mu}^{(d)}(t)-\tilde{\mu}^{(d)}(0)-\int_0^t V_d(\tilde{\mu}^{(d)}(s))\,ds, \tag{208}$$

this yields, on $\mathcal{E}_d(T)$,

$$\sup_{t\in[0,T]}\sup_{|\alpha|=w}|E_{\text{total},\alpha}^{(d)}(t)|\le(2+TBw)C_*^{2w}. \tag{209}$$

Since $\rho>C_*^2$,

$$(2+TBw)\left(\frac{C_*^2}{\rho}\right)^w\longrightarrow 0\qquad\text{as } w\to\infty. \tag{210}$$

Fix $\varepsilon>0$. Choose $w_\star$ sufficiently large that

$$\sup_{w>w_\star}(2+TBw)\left(\frac{C_*^2}{\rho}\right)^w<\frac{\varepsilon}{2}. \tag{211}$$

For degrees $w\le w_\star$, there are only finitely many multi-indices, since K is fixed. Hence (206) and a finite union bound imply

$$\max_{|\alpha|\le w_\star}\sup_{t\in[0,T]}\frac{|E_{\text{total},\alpha}^{(d)}(t)|}{\rho^{|\alpha|}}\xrightarrow{\mathbb{P}}0 \tag{212}$$

on $\mathcal{E}_d(T)$.

By $\check{E}_{\text{total}}^{(d)}(t):=\mathbf{1}_{\mathcal{E}_d(T)}E_{\text{total}}^{(d)}(t)$, $\check{E}_{\text{total}}^{(d)}(t)=0$ on $\mathcal{E}_d(T)^c$. On $\mathcal{E}_d(T)$, (209)–(211) show that $\check{E}_{\text{total}}^{(d)}(t)\in X_\rho^0$ and that

$$\sup_{t\in[0,T]}\left\|\check{E}_{\text{total}}^{(d)}(t)\right\|_\rho\xrightarrow{\mathbb{P}}0. \tag{213}$$

Finally, on $\mathcal{E}_d(T)$ we have $\check{\mu}^{(d)}=\tilde{\mu}^{(d)}$, and hence (208) gives

$$\check{\mu}^{(d)}(t)=\check{\mu}^{(d)}(0)+\mathbf{1}_{\mathcal{E}_d(T)}\int_0^t V_d(\tilde{\mu}^{(d)}(s))\,ds+\check{E}_{\text{total}}^{(d)}(t). \tag{214}$$

On $\mathcal{E}_d(T)^c$, both the integral term and $\check{E}_{\text{total}}^{(d)}$ vanish, while by definition $\check{\mu}^{(d)}(t)=\check{\mu}^{(d)}(0)$. Therefore the same identity holds almost surely on the whole probability space. This proves the lemma. □

D.1.4 Compactness of the empirical trajectories

Lemma 5. *Under the conditions of Theorem 2, for any finite $T>0$ and any $\rho>C_*^2$, the sequence of auxiliary moment processes $\{\check{\mu}^{(d)}(\cdot)\}_{d\ge1}$ is C-tight in $D([0,T];X_\rho^0)$*[5].

Proof. Define the continuous approximation

$$\hat{\mu}^{(d)}(t):=\check{\mu}^{(d)}(0)+\mathbf{1}_{\mathcal{E}_d(T)}\int_0^t V_d(\tilde{\mu}^{(d)}(s))\,ds. \tag{215}$$

By Lemma 4,

$$\check{\mu}^{(d)}(t)-\hat{\mu}^{(d)}(t)=\check{E}_{\text{total}}^{(d)}(t), \tag{216}$$

and therefore

$$\sup_{t\in[0,T]}\left\|\check{\mu}^{(d)}(t)-\hat{\mu}^{(d)}(t)\right\|_\rho\xrightarrow{\mathbb{P}}0. \tag{217}$$

[5]We say that a sequence of processes is C-tight in $D([0,T];X_\rho^0)$ if it is tight (relatively compact in distribution) and every subsequential weak limit is supported on $C([0,T];X_\rho^0)$.

We first establish compactness of the family $\{\hat{\mu}^{(d)}\}$. By (160) and (204), for every multi-index α of length w,

$$\sup_{t\in[0,T]} |\hat{\mu}_\alpha^{(d)}(t)| \le (1+TBw)C_*^{2w} =: a_w \qquad \text{almost surely.} \tag{218}$$

Indeed, on $\mathcal{E}_d(T)$ this follows from Lemma 3, while on $\mathcal{E}_d(T)^c$ the process $\hat{\mu}^{(d)}$ is constant.

Since

$$\frac{a_w}{\rho^w} = (1+TBw)\left(\frac{C_*^2}{\rho}\right)^w \longrightarrow 0, \tag{219}$$

the set

$$\mathcal{K}_{T,\rho} := \left\{ \mu \in X_\rho^0 : \sup_{|\alpha|=w} |\mu_\alpha| \le a_w \text{ for every } w \ge 1 \right\} \tag{220}$$

is compact in X_ρ^0. This is because for every $\varepsilon > 0$, the coordinates of degree larger than a sufficiently large M contribute at most ε uniformly over $\mathcal{K}_{T,\rho}$, whereas the coordinates of degree at most M form a bounded subset of a finite-dimensional space.

Moreover, for $0 \le s < t \le T$ and $|\alpha| = w$,

$$|\hat{\mu}_\alpha^{(d)}(t) - \hat{\mu}_\alpha^{(d)}(s)| \le B|t-s|\, wC_*^{2w}. \tag{221}$$

Consequently,

$$\left\|\hat{\mu}^{(d)}(t) - \hat{\mu}^{(d)}(s)\right\|_\rho \le B|t-s| \sup_{w\ge1} w\left(\frac{C_*^2}{\rho}\right)^w =: L_\rho|t-s|, \tag{222}$$

where $L_\rho < \infty$ is independent of d.

Thus the trajectories $\hat{\mu}^{(d)}$ take values almost surely in the compact set $\mathcal{K}_{T,\rho}$ and are uniformly Lipschitz in X_ρ^0. By the Arzelà–Ascoli theorem, the family $\{\hat{\mu}^{(d)}\}$ is tight in $C([0,T]; X_\rho^0)$.

Finally, (217) shows that $\check{\mu}^{(d)}$ is asymptotically equivalent to the tight family $\hat{\mu}^{(d)}$ in the uniform metric. It follows that $\{\check{\mu}^{(d)}\}_{d\ge1}$ is tight in $D([0,T]; X_\rho^0)$, and every subsequential weak limit is supported on $C([0,T]; X_\rho^0)$. Hence the sequence is C-tight. □

D.1.5 Global well-posedness

We show below that V_d converges uniformly to V under the spectral bounds in Lemma 3.

Lemma 6. *Fix $C_* > 0$ and $\rho > C_*^2$. Let*

$$\mathcal{A}_d(C_*) := \left\{ (S_1,\dots,S_K) : S_k = W_kW_k^T,\ \max_{1\le k\le K} \|W_k\|_{\mathrm{op}} \le C_* \right\}. \tag{223}$$

For $\mathbf{S} \in \mathcal{A}_d(C_*)$, let $\mu(\mathbf{S})$ denote its associated moment sequence. Then

$$\sup_{\mathbf{S}\in\mathcal{A}_d(C_*)} \|V_d(\mu(\mathbf{S})) - V(\mu(\mathbf{S}))\|_\rho \longrightarrow 0 \qquad \text{as } d\to\infty. \tag{224}$$

Consequently, on the spectral-bound event $\mathcal{E}_d(T)$,

$$\sup_{t\in[0,T]} \left\| V_d(\tilde{\mu}^{(d)}(t)) - V(\tilde{\mu}^{(d)}(t)) \right\|_\rho \xrightarrow{\mathbb{P}} 0. \tag{225}$$

Proof. Define

$$\Gamma_{kl}^{(d)} := \frac{1}{2} \sum_{i,j,i',j'} (\mathcal{C}_{ii'}\mathcal{C}_{jj'} + \mathcal{C}_{ij'}\mathcal{C}_{ji'})\, \mathbb{E}_x\left[\partial^2_{(ijk),(i'j'l)}\mathcal{L}(G^{(d)})\right]. \tag{226}$$

By the definition of H_k,

$$H_k = 4\sum_{l=1}^K \Gamma_{kl}^{(d)} S_l. \tag{227}$$

Hence the empirical vector field (164) can be written as

$$[V_d(\mu)]_\alpha = -\sum_{\substack{r=1\\ k_r\in\mathcal{K}}}^{w}\sum_{l=1}^{K}\Gamma^{(d)}_{k_r l}\left(\mu_{\alpha^{r,l,-}}+\mu_{\alpha^{r,l,+}}\right)-\frac{\gamma}{2}|\alpha|_{\mathcal{K}}\mu_\alpha, \tag{228}$$

where $\alpha^{r,l,-}$ and $\alpha^{r,l,+}$ denote the two multi-indices obtained by inserting l before and after k_r, respectively.

The limiting vector field has the same form with $\Gamma^{(d)}_{kl}$ replaced by $\Gamma_{kl}(q)$ defined in (55). Therefore, defining

$$\delta_d(C_*) := \sup_{\mathbf{S}\in\mathcal{A}_d(C_*)}\max_{k,l}\left|\Gamma^{(d)}_{kl}-\Gamma_{kl}(q(\mathbf{S}))\right|, \tag{229}$$

we have, for every $|\alpha|=w$,

$$|[V_d(\mu)-V(\mu)]_\alpha| \le 2Kw\,\delta_d(C_*)\,C_*^{2(w+1)}. \tag{230}$$

Thus,

$$\|V_d(\mu)-V(\mu)\|_\rho \le 2KC_*^2\delta_d(C_*)\sup_{w\ge1} w\left(\frac{C_*^2}{\rho}\right)^w. \tag{231}$$

It remains to prove $\delta_d(C_*)\to 0$.

Suppose otherwise. Then there exist $\epsilon>0$, a subsequence d_m, and $\mathbf{S}^{(d_m)}\in\mathcal{A}_{d_m}(C_*)$ such that

$$\max_{k,l}\left|\Gamma^{(d_m)}_{kl}-\Gamma_{kl}(q^{(d_m)})\right|\ge\epsilon. \tag{232}$$

Since K is fixed and $|q^{(d_m)}_{kl}|\le C_*^4$, after taking a further subsequence we may assume $q^{(d_m)}\to q^*\in\mathbb{S}^K_+$.

Define the active index set

$$I := \{k\in\{1,\dots,K\} : q^*_{kk}>0\}. \tag{233}$$

For every $k\in I$,

$$\frac{1}{d_m}\|S^{(d_m)}_k\|_F^2 = q^{(d_m)}_{kk}\longrightarrow q^*_{kk}>0. \tag{234}$$

Since $\|S^{(d_m)}_k\|_{\mathrm{op}}\le C_*^2$, it follows that

$$\frac{\|S^{(d_m)}_k\|_{\mathrm{op}}}{\|S^{(d_m)}_k\|_F}\longrightarrow 0. \tag{235}$$

Hence Corollary 5 applies to $k\in I$. For $k\notin I$, we instead have $q^{(d_m)}_{kk}\to 0$, which gives

$$G^{(d_m)}_{ijk}\longrightarrow 0 \qquad \text{in } L^2. \tag{236}$$

Moreover, all covariances involving an inactive index converge to zero. Combining them,

$$G^{(d_m)}\to G^*\sim\mathcal{N}(0,\Sigma(q^*)) \tag{237}$$

in distribution. Therefore, by Assumption 2.1 and uniform integrability,

$$\mathbb{E}\left[\partial^2_{(ijk),(i'j'l)}\mathcal{L}(G^{(d_m)})\right]\longrightarrow\mathbb{E}\left[\partial^2_{(ijk),(i'j'l)}\mathcal{L}(G^*)\right]. \tag{238}$$

Consequently, by the definitions of $\Gamma^{(d_m)}$ and $\Gamma(q^*)$,

$$\Gamma^{(d_m)}_{kl}\longrightarrow\Gamma_{kl}(q^*) \qquad \text{for every } k,l. \tag{239}$$

This contradicts (232), and hence

$$\delta_d(C_*)\to 0. \tag{240}$$

The claimed convergence in X_ρ follows. □

The next lemma establishes the global well-posedness of (8).

Lemma 7. *Under the conditions of Theorem 2, let C_* be the uniform spectral bound in Lemma 3. For every finite $T > 0$ and every pair $C_*^2 < \rho_- < \rho_+$, there exists a unique trajectory $\bar{\mu} \in C([0,T]; X^0_{\rho_-})$ such that*

$$\bar{\mu}(t) = \bar{\mu}(0) + \int_0^t V(\bar{\mu}(s))\, ds, \qquad t \in [0,T], \tag{241}$$

where the identity is understood in $X^0_{\rho_+}$. Moreover,

$$q(\bar{\mu}(t)) \succeq 0, \qquad t \in [0,T], \tag{242}$$

and $|\bar{\mu}_\alpha(t)| \le C_*^{2|\alpha|}$ for every multi-index α.

Proof. Fix $T > 0$ and $C_*^2 < \rho_- < \rho_+$. By Lemma 5, the sequence $\{\check{\mu}^{(d)}\}_{d\ge 1}$ is C-tight in $D([0,T]; X^0_{\rho_-})$. Hence, from any subsequence we may extract a further subsequence, still indexed by d, such that

$$\check{\mu}^{(d)} \to \mu^* \qquad \text{in } D([0,T]; X^0_{\rho_-}), \tag{243}$$

where $\mu^* \in C([0,T]; X^0_{\rho_-})$ almost surely.

We next rewrite the auxiliary equation in terms of the limiting vector field. By Lemma 4,

$$\check{\mu}^{(d)}(t) = \check{\mu}^{(d)}(0) + \mathbf{1}_{\mathcal{E}_d(T)} \int_0^t V_d(\tilde{\mu}^{(d)}(s))\, ds + \check{E}^{(d)}_{\mathrm{total}}(t), \tag{244}$$

with

$$\sup_{t\in[0,T]} \left\| \check{E}^{(d)}_{\mathrm{total}}(t) \right\|_{\rho_+} \xrightarrow{\mathbb{P}} 0. \tag{245}$$

Define

$$R_d(t) := \check{E}^{(d)}_{\mathrm{total}}(t) + \int_0^t \left[\mathbf{1}_{\mathcal{E}_d(T)} V_d(\tilde{\mu}^{(d)}(s)) - V(\check{\mu}^{(d)}(s)) \right] ds. \tag{246}$$

Then

$$\check{\mu}^{(d)}(t) = \check{\mu}^{(d)}(0) + \int_0^t V(\check{\mu}^{(d)}(s))\, ds + R_d(t). \tag{247}$$

We claim that

$$\sup_{t\in[0,T]} \|R_d(t)\|_{\rho_+} \xrightarrow{\mathbb{P}} 0. \tag{248}$$

Indeed, on $\mathcal{E}_d(T)$ we have $\check{\mu}^{(d)} = \tilde{\mu}^{(d)}$, and therefore

$$\mathbf{1}_{\mathcal{E}_d(T)} \left\| V_d(\tilde{\mu}^{(d)}(s)) - V(\check{\mu}^{(d)}(s)) \right\|_{\rho_+} = \mathbf{1}_{\mathcal{E}_d(T)} \left\| V_d(\tilde{\mu}^{(d)}(s)) - V(\tilde{\mu}^{(d)}(s)) \right\|_{\rho_+}. \tag{249}$$

Lemma 6 implies

$$\mathbf{1}_{\mathcal{E}_d(T)} \sup_{s\in[0,T]} \left\| V_d(\tilde{\mu}^{(d)}(s)) - V(\tilde{\mu}^{(d)}(s)) \right\|_{\rho_+} \xrightarrow{\mathbb{P}} 0. \tag{250}$$

On the other hand, by (160), the auxiliary process always satisfies

$$|\check{\mu}^{(d)}_\alpha(s)| \le C_*^{2|\alpha|}. \tag{251}$$

Since $\widetilde{\Gamma}$ is bounded on the corresponding bounded set of second moments, there exists a constant $C_V = C_V(C_*, \rho_+) < \infty$ such that

$$\sup_{d\ge 1} \sup_{s\in[0,T]} \|V(\check{\mu}^{(d)}(s))\|_{\rho_+} \le C_V \qquad \text{almost surely.} \tag{252}$$

Consequently,

$$\mathbf{1}_{\mathcal{E}_d(T)^c} \sup_{s\in[0,T]} \|V(\check{\mu}^{(d)}(s))\|_{\rho_+} \xrightarrow{\mathbb{P}} 0, \tag{253}$$

because $\mathbb{P}(\mathcal{E}_d(T)^c) \to 0$. Together with (245) and (250), this proves (248).

We also have $\check{\mu}^{(d)}(0) = \tilde{\mu}^{(d)}(0)$. Assumption 2.3 gives convergence in probability of every coordinate, while Assumption 2.2 gives $|\check{\mu}_\alpha^{(d)}(0)| \le C_0^{2|\alpha|}$. Since $\rho_- > C_*^2 > C_0^2$, the contribution of sufficiently high degrees is uniformly exponentially small, which yields

$$\check{\mu}^{(d)}(0) \xrightarrow{\mathbb{P}} \bar{\mu}(0) \qquad \text{in } X^0_{\rho_-}. \tag{254}$$

Because $X^0_{\rho_-}$ is separable, we may apply the Skorokhod representation theorem to the joint laws of $\big(\check{\mu}^{(d)}, \check{\mu}^{(d)}(0), R_d\big)$. After passing to a further subsequence and realizing these random elements on a common probability space, we may assume that, almost surely,

$$\check{\mu}^{(d)} \to \mu^* \qquad \text{in } D([0,T]; X^0_{\rho_-}), \tag{255}$$

$$\check{\mu}^{(d)}(0) \longrightarrow \bar{\mu}(0) \qquad \text{in } X^0_{\rho_-}, \tag{256}$$

and

$$\sup_{t\in[0,T]} \|R_d(t)\|_{\rho_+} \longrightarrow 0. \tag{257}$$

Since μ^* is continuous, convergence in the Skorokhod topology implies

$$\sup_{t\in[0,T]} \left\|\check{\mu}^{(d)}(t) - \mu^*(t)\right\|_{\rho_-} \longrightarrow 0 \qquad \text{almost surely.} \tag{258}$$

The vector field V satisfies the Cauchy estimate established in Lemma 1. Since the convergent trajectories remain in a bounded subset of $X^0_{\rho_-}$, there exists $L < \infty$ such that

$$\|V(\mu) - V(\nu)\|_{\rho_+} \le \frac{L}{\rho_+ - \rho_-}\|\mu - \nu\|_{\rho_-} \tag{259}$$

for all states μ, ν under consideration. Hence

$$\sup_{t\in[0,T]} \left\|\int_0^t \Big[V(\check{\mu}^{(d)}(s)) - V(\mu^*(s))\Big]\, ds\right\|_{\rho_+} \le \frac{LT}{\rho_+ - \rho_-} \sup_{s\in[0,T]} \left\|\check{\mu}^{(d)}(s) - \mu^*(s)\right\|_{\rho_-} \longrightarrow 0. \tag{260}$$

Passing to the limit in (247) yields

$$\mu^*(t) = \bar{\mu}(0) + \int_0^t V(\mu^*(s))\, ds, \qquad t \in [0,T], \tag{261}$$

as an identity in $X^0_{\rho_+}$.

We now show that the limiting trajectory satisfies the original ODE and inherits the uniform moment bounds. For every finite d and every $t \in [0,T]$,

$$q_{kl}(\check{\mu}^{(d)}(t)) = \frac{1}{d}\mathrm{Tr}\left[\check{S}_k^{(d)}(t)\check{S}_l^{(d)}(t)\right]. \tag{262}$$

Therefore

$$q(\check{\mu}^{(d)}(t)) \succeq 0 \qquad \text{almost surely for every } d, t. \tag{263}$$

Since the PSD cone is closed and convergence in $X^0_{\rho_-}$ implies convergence of second-order coordinates,

$$q(\mu^*(t)) \succeq 0, \qquad t \in [0,T]. \tag{264}$$

Thus

$$\mathscr{P}(q(\mu^*(t))) = q(\mu^*(t)), \tag{265}$$

and consequently

$$\widetilde{\Gamma}(q(\mu^*(t))) = \Gamma(q(\mu^*(t))). \tag{266}$$

Therefore μ^* satisfies the original equation (8). Moreover, by (161),

$$|\check{\mu}_\alpha^{(d)}(t)| \le C_*^{2|\alpha|} \qquad \text{almost surely for every } d, t. \tag{267}$$

Passing to the limit gives

$$|\mu_\alpha^*(t)| \le C_*^{2|\alpha|}, \qquad \alpha,\ t \in [0,T]. \tag{268}$$

We finally establish uniqueness. Suppose that μ_1 and μ_2 are two solutions of (8). Both are also solutions of (58). By Lemma 1, the solution of (58) is locally unique in the Banach scale. Hence, if $\mu_1(t_0) = \mu_2(t_0)$ at some $t_0 < T$, the two solutions coincide on a finite interval to the right of t_0. Starting from $t_0 = 0$ and applying this local uniqueness with a contradiction argument shows that

$$\mu_1(t) = \mu_2(t), \qquad t \in [0,T]. \tag{269}$$

Thus the solution is unique.

Since the initial condition $\bar{\mu}(0)$ is deterministic, uniqueness implies that the random limit μ^* is almost surely equal to a deterministic trajectory, which we denote by $\bar{\mu}$ as a unique trajectory. □

D.1.6 Final proof of Theorem 2

Fix an arbitrary finite horizon $T > 0$, and choose

$$C_*^2 < \rho_- < \rho_+. \tag{270}$$

By Lemma 5, the auxiliary trajectories $\{\check{\mu}^{(d)}\}_{d\ge1}$ are C-tight in $D([0,T]; X^0_{\rho_-})$.

By Lemma 7, every subsequential weak limit is the same deterministic path $\bar{\mu}$, and therefore

$$\check{\mu}^{(d)} \to \bar{\mu} \qquad \text{in } D([0,T]; X^0_{\rho_-}). \tag{271}$$

Since $\bar{\mu}$ is deterministic, convergence in distribution to $\bar{\mu}$ is equivalent to convergence in probability:

$$d_{J_1}\left(\check{\mu}^{(d)}, \bar{\mu}\right) \xrightarrow{\mathbb{P}} 0, \tag{272}$$

where d_{J_1} denotes the Skorokhod J_1 metric. Moreover, $\bar{\mu}$ is continuous. Convergence in the Skorokhod topology to a continuous limit implies convergence in the uniform topology. Hence

$$\sup_{t\in[0,T]} \left\|\check{\mu}^{(d)}(t) - \bar{\mu}(t)\right\|_{\rho_-} \xrightarrow{\mathbb{P}} 0. \tag{273}$$

For every fixed multi-index α,

$$\sup_{t\in[0,T]} \left|\check{\mu}_\alpha^{(d)}(t) - \bar{\mu}_\alpha(t)\right| \le \rho_-^{|\alpha|} \sup_{t\in[0,T]} \left\|\check{\mu}^{(d)}(t) - \bar{\mu}(t)\right\|_{\rho_-}, \tag{274}$$

and therefore

$$\sup_{t\in[0,T]} \left|\check{\mu}_\alpha^{(d)}(t) - \bar{\mu}_\alpha(t)\right| \xrightarrow{\mathbb{P}} 0. \tag{275}$$

It remains only to transfer this convergence from the auxiliary trajectory to the original trajectory. By (163), for every fixed multi-index α and every $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{t\in[0,T]} \left|\tilde{\mu}_\alpha^{(d)}(t) - \bar{\mu}_\alpha(t)\right| > \varepsilon\right) \le \mathbb{P}(\mathcal{E}_d(T)^c) + \mathbb{P}\left(\sup_{t\in[0,T]} \left|\check{\mu}_\alpha^{(d)}(t) - \bar{\mu}_\alpha(t)\right| > \varepsilon\right). \tag{276}$$

The first term tends to zero by Lemma 3, while the second tends to zero by (275). Consequently,

$$\lim_{d\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \left|\tilde{\mu}_\alpha^{(d)}(t) - \bar{\mu}_\alpha(t)\right| > \varepsilon\right) = 0. \tag{277}$$

This proves Theorem 2.

D.2 Proof of Corollary 2

We define a projection operator $\Pi_M : X_s \to X_s$ for a $M \geq 2$, such that for any multi-index α:

$$[\Pi_M \mu]_\alpha = \begin{cases} \mu_\alpha, & \text{if } |\alpha| \leq M \\ 0, & \text{if } |\alpha| > M \end{cases}. \tag{278}$$

Let $R := C_*^2$. By Lemma 7,

$$|\bar{\mu}_\alpha(t)| \leq R^{|\alpha|}, \qquad t \geq 0. \tag{279}$$

In particular, $\|\bar{q}(t)\|_F \leq KR^2$.

We first show that the finite-dimensional truncation is equivalently the restriction of the extended infinite-dimensional ODE (58) to an invariant subspace. Indeed, define

$$Y_M := \{\mu : \mu_\alpha = 0 \text{ for every } |\alpha| > M\}. \tag{280}$$

If $\mu \in Y_M$ and $|\alpha| > M$, then every moment appearing in $[V(\mu)]_\alpha$ has degree either $|\alpha|$ or $|\alpha| + 1$, and hence vanishes. Therefore, $[V(\mu)]_\alpha = 0$ for all $|\alpha| > M$. Thus Y_M is invariant under (58), and $\mu^{(M)}$ is precisely the solution of

$$\frac{d}{dt}\mu^{(M)}(t) = V(\mu^{(M)}(t)), \qquad \mu^{(M)}(0) = \Pi_M \bar{\mu}(0). \tag{281}$$

We next establish stability with respect to this truncation. Let

$$\mathcal{U} := \left\{q \in \mathbb{R}^{K\times K} : \|q\|_F \leq KR^2 + 1\right\}. \tag{282}$$

Since $\widetilde{\Gamma}$ is locally bounded and locally Lipschitz (see Lemma 1), there exist finite constants C_Γ, L_Γ such that

$$\max_{k,l} |\widetilde{\Gamma}_{kl}(q)| \leq C_\Gamma, \qquad q \in \mathcal{U}, \tag{283}$$

and

$$\max_{k,l} \left|\widetilde{\Gamma}_{kl}(q) - \widetilde{\Gamma}_{kl}(q')\right| \leq L_\Gamma \|q - q'\|_F, \qquad q, q' \in \mathcal{U}. \tag{284}$$

Choose $s_0 > R$ and γ sufficiently large (Assumption 2.4) such that

$$\kappa := \frac{4KC_\Gamma s_0}{\gamma} + \frac{4K^2 L_\Gamma s_0^2 R}{\gamma} < 1. \tag{285}$$

Define the error

$$e^{(M)}(t) := \mu^{(M)}(t) - \bar{\mu}(t) \tag{286}$$

and

$$E_M(t) := \|e^{(M)}(t)\|_{s_0} = \sup_\alpha \frac{|e_\alpha^{(M)}(t)|}{s_0^{|\alpha|}}. \tag{287}$$

Let $q^{(M)}(t) := q(\mu^{(M)}(t))$. As long as $q^{(M)}(t) \in \mathcal{U}$, we have

$$\|q^{(M)}(t) - \bar{q}(t)\|_F \leq Ks_0^2 E_M(t). \tag{288}$$

Let $\alpha^{i,l,-}$ and $\alpha^{i,l,+}$ denote the multi-indices obtained by inserting l before and after k_i, respectively. Subtracting the equations for $\mu^{(M)}$ and $\bar{\mu}$, and using $\widetilde{\Gamma}(\bar{q}) = \Gamma(\bar{q})$, gives

$$\begin{aligned} \frac{d}{dt}e_\alpha^{(M)} = &- \sum_{\substack{1\leq i\leq w \\ k_i \in \mathcal{K}}} \sum_{l=1}^{K} \widetilde{\Gamma}_{k_i l}(q^{(M)}) \left(e^{(M)}_{\alpha^{i,l,-}} + e^{(M)}_{\alpha^{i,l,+}}\right) \\ &- \sum_{\substack{1\leq i\leq w \\ k_i \in \mathcal{K}}} \sum_{l=1}^{K} \left[\widetilde{\Gamma}_{k_i l}(q^{(M)}) - \Gamma_{k_i l}(\bar{q})\right] \left(\bar{\mu}_{\alpha^{i,l,-}} + \bar{\mu}_{\alpha^{i,l,+}}\right) - \frac{\gamma}{2}|\alpha|_{\mathcal{K}} e_\alpha^{(M)}. \end{aligned} \tag{289}$$

If $|\alpha|_{\mathcal{K}} = 0$, the right-hand side vanishes and hence $e_\alpha^{(M)}(t) = e_\alpha^{(M)}(0)$. Suppose now that $|\alpha|_{\mathcal{K}} \geq 1$. Applying the variation-of-constants formula to (289) and using (283) gives

$$\frac{|e_\alpha^{(M)}(t)|}{s_0^w} \leq \frac{|e_\alpha^{(M)}(0)|}{s_0^w} + 2KC_\Gamma |\alpha|_{\mathcal{K}} s_0 \int_0^t e^{-\frac{\gamma}{2}|\alpha|_{\mathcal{K}}(t-u)} E_M(u)\, du + \frac{1}{s_0^w} \int_0^t e^{-\frac{\gamma}{2}|\alpha|_{\mathcal{K}}(t-u)} \sum_{i:k_i \in \mathcal{K}} \sum_{l=1}^{K} \left| \widetilde{\Gamma}_{k_i l}(q^{(M)}(u)) - \Gamma_{k_i l}(\bar{q}(u)) \right| \times \left(|\bar{\mu}_{\alpha^{i,l,-}}(u)| + |\bar{\mu}_{\alpha^{i,l,+}}(u)| \right) du. \tag{290}$$

By (284), (288), and (279), the second integral is bounded by

$$2K^2 L_\Gamma s_0^2 R\, |\alpha|_{\mathcal{K}} \left(\frac{R}{s_0}\right)^w \int_0^t e^{-\frac{\gamma}{2}|\alpha|_{\mathcal{K}}(t-u)} E_M(u)\, du. \tag{291}$$

Since

$$|\alpha|_{\mathcal{K}} \int_0^t e^{-\frac{\gamma}{2}|\alpha|_{\mathcal{K}}(t-u)}\, du \leq \frac{2}{\gamma}, \tag{292}$$

and $R/s_0 < 1$, we obtain

$$\frac{|e_\alpha^{(M)}(t)|}{s_0^w} \leq \frac{|e_\alpha^{(M)}(0)|}{s_0^w} + \kappa \sup_{0 \leq u \leq t} E_M(u), \tag{293}$$

with κ defined in (285). The same inequality trivially holds when $|\alpha|_{\mathcal{K}} = 0$.

Taking the supremum over α and over $0 \leq u \leq t$, we therefore obtain

$$\sup_{0 \leq u \leq t} E_M(u) \leq E_M(0) + \kappa \sup_{0 \leq u \leq t} E_M(u). \tag{294}$$

Since $\kappa < 1$,

$$\sup_{0 \leq u \leq t} E_M(u) \leq \frac{1}{1-\kappa} E_M(0). \tag{295}$$

It remains to estimate the initial truncation error. By definition,

$$e_\alpha^{(M)}(0) = \begin{cases} 0, & |\alpha| \leq M, \\ -\bar{\mu}_\alpha(0), & |\alpha| > M. \end{cases} \tag{296}$$

Using (279),

$$E_M(0) = \sup_{w > M} \sup_{|\alpha| = w} \frac{|\bar{\mu}_\alpha(0)|}{s_0^w} \leq \sup_{w > M} \left(\frac{R}{s_0}\right)^w = \left(\frac{R}{s_0}\right)^{M+1}. \tag{297}$$

Combining (295) and (297) gives

$$\sup_{0 \leq t \leq T} \|\mu^{(M)}(t) - \bar{\mu}(t)\|_{s_0} \leq \frac{1}{1-\kappa} \left(\frac{R}{s_0}\right)^{M+1}, \tag{298}$$

provided $q^{(M)}(t) \in \mathcal{U}$.

We finally close this condition by a standard bootstrap argument. Since

$$\|q^{(M)}(t) - \bar{q}(t)\|_F \leq \frac{K s_0^2}{1-\kappa} \left(\frac{R}{s_0}\right)^{M+1}, \tag{299}$$

the right-hand side is smaller than 1 for all sufficiently large M. Because $\|\bar{q}(t)\|_F \leq KR^2$, it follows that

$$\|q^{(M)}(t)\|_F < KR^2 + 1, \tag{300}$$

so $q^{(M)}(t)$ cannot leave $\mathcal{U}$. Moreover, (298) bounds every coordinate of (10), so its solution cannot undergo finite-time blow-up. Hence, for all sufficiently large M, the solution of (10) exists on $[0, T]$, and (298) holds throughout the interval. This finishes the proof.

D.3 An exactly solvable example

Consider the squared loss $\mathcal{L}(G_1, G_2) = (G_1 - G_2)^2$ without weight decay ($\gamma = 0$). It is a simplified setting where $L = 1$ and $K = 2$. We assume only the first index is learnable ($\mathcal{K} = \{1\}$). The potential evaluates to:

$$\Phi(q) = \mathbb{E}_{z_1,z_2}||z_1 - z_2||^2 = 2q_{11} + 2q_{22} - 4q_{12}. \tag{301}$$

Because the partial derivatives of $\Phi(q)$ are constant, the dynamics in (8) form an infinite-dimensional linear ODE hierarchy. Truncating this system at a finite degree N yields a closed linear ODE with polynomial solutions. At truncation degree $N = 1$, the solution is static: $\bar{\mu}_1(t) = \bar{\mu}_1(0)$. At $N = 2$, the dynamics activate:

$$\begin{aligned} &\bar{q}_{11}(t) = \bar{q}_{11}(0),\ \bar{q}_{12}(t) = \bar{q}_{12}(0), \\ &\bar{\mu}_1(t) = \bar{\mu}_1(0) - 4t[\bar{q}_{11}(0) - \bar{q}_{12}(0)]. \end{aligned} \tag{302}$$

Extending to $N = 3$, the solution captures higher-order curvature:

$$\begin{aligned} &\bar{q}_{11}(t) = \bar{q}_{11}(0) - 8t[\bar{\mu}_{111}(0) - \bar{\mu}_{112}(0)],\ \bar{q}_{12}(t) = \bar{q}_{12}(0) - 4t[\bar{\mu}_{112}(0) - \bar{\mu}_{122}(0)], \\ &\bar{\mu}_1(t) = \bar{\mu}_1(0) - 4t[\bar{q}_{11}(0) - q_{12}(0)] + 8t^2[2\bar{\mu}_{111}(0) - 3\bar{\mu}_{112}(0) + \bar{\mu}_{122}(0)]. \end{aligned} \tag{303}$$

Remarkably, these truncated polynomial solutions match the Taylor expansion of the original matrix flow evaluated around $t = 0$. To see this, note that the gradient flow for the student forms a matrix Riccati differential equation:

$$\frac{d}{dt}S_1 = 2(S_1S_2 + S_2S_1) - 4S_1^2. \tag{304}$$

Assuming that $S_1(0)$ and S_2 are invertible, (304) admits a closed-form solution:

$$S_1(t) = \left[e^{-2S_2t}S_1(0)^{-1}e^{-2S_2t} + S_2^{-1}\left(I - e^{-4S_2t}\right)\right]^{-1}, \tag{305}$$

Taking the first-order Taylor expansion of the inner inverse matrix around $t = 0$ yields:

$$\begin{aligned} S_1(t)^{-1} &\approx (I - 2S_2t)S_1(0)^{-1}(I - 2S_2t) + S_2^{-1}(4S_2t) \\ &\approx S_1(0)^{-1} + \left(4I - 2S_2S_1(0)^{-1} - 2S_1(0)^{-1}S_2\right)t. \end{aligned} \tag{306}$$

Using the matrix identity $(A + tB)^{-1} \approx A^{-1} - A^{-1}BA^{-1}t$, we recover the first-order expansion for $S_1(t)$:

$$S_1(t) \approx S_1(0) + (2S_1(0)S_2 + 2S_2S_1(0) - 4S_1(0)^2)t. \tag{307}$$

Taking the normalized traces of this matrix expansion reproduces the linear terms of the truncated moments:

$$\begin{aligned} &\bar{q}_{11}(t) = \bar{q}_{11}(0) - 8t[\bar{\mu}_{111}(0) - \bar{\mu}_{112}(0)],\ \bar{q}_{12}(t) = \bar{q}_{12}(0) - 4t[\bar{\mu}_{112}(0) - \bar{\mu}_{122}(0)], \\ &\bar{\mu}_1(t) = \bar{\mu}_1(0) - 4t[\bar{q}_{11}(0) - \bar{q}_{12}(0)]. \end{aligned} \tag{308}$$

Numerical results are shown in Figure 4.

The closed-form solution also provides a direct interpretation of why the moment dynamics is intrinsically infinite-dimensional. To see this, consider the special case $S_2 = cI_d$ for some $c > 0$. Then each eigenvalue $\lambda(t)$ of $S_1(t)$ evolves independently according to $\dot{\lambda}(t) = 4c\lambda(t) - 4\lambda(t)^2$, and hence

$$\lambda(t) = \frac{c\lambda(0)}{\lambda(0) + (c - \lambda(0))e^{-4ct}}. \tag{309}$$

Therefore, if ρ_0 denotes the limiting empirical spectral distribution of $S_1(0)$, the first moment satisfies

$$\mu_1(t) = \int \frac{c\lambda}{\lambda + (c - \lambda)e^{-4ct}}\, d\rho_0(\lambda). \tag{310}$$

Restricting ρ_0 to finitely supported measures $\rho_0 = \sum_{j=0}^{M} p_j\delta_{\lambda_j}$ with distinct $\lambda_j \neq c$, the functions $t \longmapsto \frac{c\lambda_j}{\lambda_j + (c-\lambda_j)e^{-4ct}}$ are linearly independent. Consequently, the trajectory $\{\mu_1(t) : t \geq 0\}$ uniquely determines the weights $(p_0, \ldots, p_M)$.

Suppose, for contradiction, that these trajectories admitted a fixed r-dimensional ODE representation whose initial state depends continuously on the initial spectrum. For each M, the map from the interior

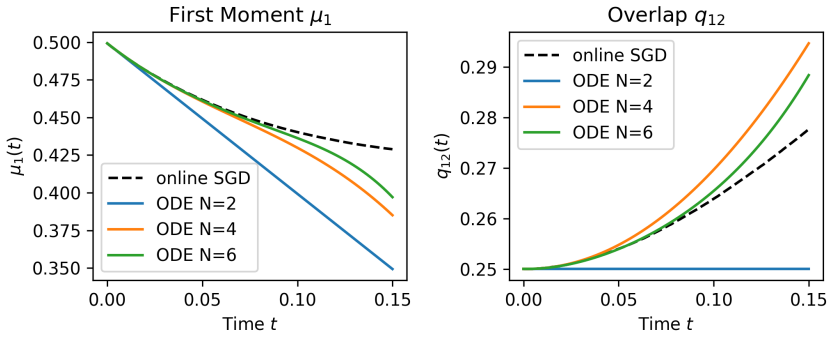


Figure 4: Comparison between online SGD and the theory at different truncation orders N for the linear activation, $L = 1$ and the MSE loss. As N grows larger, the theory is better aligned with the experiments. We choose $d = 400$, learning rate 0.01, batch size 512 and no weight decay ($\gamma = 0$).

of the M-dimensional simplex $(p_0, \ldots, p_M)$ to the ODE initial state must then be injective, since identical ODE initial states generate identical trajectories, while the trajectory uniquely determines the weights. For $M > r$, however, no continuous injective map from an M-dimensional open set into $\mathbb{R}^r$ exists (as a consequence of Brouwer's invariance of domain theorem). Since M is arbitrary, no fixed finite-dimensional ODE with a continuous initialization map can reproduce this family of trajectories. This shows that the infinite-dimensionality of (8) is intrinsic and cannot in general be removed by replacing the moment hierarchy with a fixed finite set of order parameters.

D.4 Proof of Corollary 3

Let

$$\bar{p}_{12}(t) = \bar{q}_{12}(t) - \bar{t}_1(t)\bar{t}_2, \tag{311}$$

where $\bar{t}_2$ is constant because the teacher is fixed. We first derive its limiting dynamics from (8).

For the order parameter $\bar{q}_{12} = \bar{\mu}_{(1,2)}$, only the student index 1 is trainable. Therefore, from (8) we have

$$\frac{d}{dt}\bar{q}_{12} = -2\Gamma_{11}(\bar{q})\bar{\mu}_{(1,1,2)} - 2\Gamma_{12}(\bar{q})\bar{\mu}_{(1,2,2)} - \frac{\gamma}{2}\bar{q}_{12}. \tag{312}$$

Similarly, for $\bar{t}_1 = \bar{\mu}_{(1)}$,

$$\frac{d}{dt}\bar{t}_1 = -2\Gamma_{11}(\bar{q})\bar{q}_{11} - 2\Gamma_{12}(\bar{q})\bar{q}_{12} - \frac{\gamma}{2}\bar{t}_1. \tag{313}$$

Subtracting $\bar{t}_2$ times (313) from (312) gives

$$\frac{d}{dt}\bar{p}_{12} = -2\Gamma_{11}(\bar{q})\left(\bar{\mu}_{(1,1,2)} - \bar{t}_2\bar{q}_{11}\right) - 2\Gamma_{12}(\bar{q})\left(\bar{\mu}_{(1,2,2)} - \bar{t}_2\bar{q}_{12}\right) - \frac{\gamma}{2}\bar{p}_{12}. \tag{314}$$

Define the centered third-order moments

$$\mathring{\bar{\mu}}_{(1,1,2)} := \lim_{d\to\infty}\frac{1}{d}\mathrm{Tr}[\mathring{S}_1^2\mathring{S}_2], \qquad \mathring{\bar{\mu}}_{(1,2,2)} := \lim_{d\to\infty}\frac{1}{d}\mathrm{Tr}[\mathring{S}_1\mathring{S}_2^2]. \tag{315}$$

Expanding $S_k = \mathring{S}_k + \bar{t}_k I$ yields

$$\bar{\mu}_{(1,1,2)} - \bar{t}_2\bar{q}_{11} = \mathring{\bar{\mu}}_{(1,1,2)} + 2\bar{t}_1\bar{p}_{12}, \tag{316}$$

and

$$\bar{\mu}_{(1,2,2)} - \bar{t}_2\bar{q}_{12} = \mathring{\bar{\mu}}_{(1,2,2)} + \bar{t}_2\bar{p}_{12} + \bar{t}_1\bar{p}_{22}. \tag{317}$$

Consequently,

$$\frac{d}{dt}\bar{p}_{12} = -2\Gamma_{11}(\bar{q})\left(\mathring{\bar{\mu}}_{(1,1,2)} + 2\bar{t}_1\bar{p}_{12}\right) - 2\Gamma_{12}(\bar{q})\left(\mathring{\bar{\mu}}_{(1,2,2)} + \bar{t}_2\bar{p}_{12} + \bar{t}_1\bar{p}_{22}\right) - \frac{\gamma}{2}\bar{p}_{12}. \tag{318}$$

At $t = 0$, Assumption 3 gives

$$\bar{p}_{12}(0) = \bar{\tilde{\mu}}_{(1,1,2)}(0) = \bar{\tilde{\mu}}_{(1,2,2)}(0) = 0. \tag{319}$$

Hence (318) reduces to

$$\left.\frac{d}{dt}\bar{p}_{12}(t)\right|_{t=0} = -2\Gamma_{12}(\bar{q}(0))\bar{t}_1(0)\bar{p}_{22}. \tag{320}$$

By Assumption 3, this quantity is nonzero. Set

$$v_0 := \left|\left.\frac{d}{dt}\bar{p}_{12}(t)\right|_{t=0}\right| = 2|\Gamma_{12}(\bar{q}(0))|\bar{t}_1(0)\bar{p}_{22} > 0. \tag{321}$$

Since the limiting trajectory satisfies (8) and the relevant finite-order components of $V(\bar{\mu}(t))$ are continuous in t, $\bar{p}_{12}$ is continuously differentiable. By continuity there exists $T_0 > 0$, independent of d, such that

$$\left|\frac{d}{dt}\bar{p}_{12}(t) - \frac{d}{dt}\bar{p}_{12}(0)\right| \le \frac{v_0}{2}, \qquad 0 \le t \le T_0. \tag{322}$$

In particular, by continuity $\frac{d}{dt}\bar{p}_{12}(t)$ has the same sign as $\frac{d}{dt}\bar{p}_{12}(0)$ throughout $[0, T_0]$. Since $\bar{p}_{12}(0) = 0$,

$$|\bar{p}_{12}(T_0)| \ge \frac{v_0 T_0}{2}. \tag{323}$$

Choose $c := \frac{v_0 T_0}{4} > 0$ and $T_+ := T_0$. Then

$$|\bar{p}_{12}(T_+)| \ge 2c. \tag{324}$$

On the other hand, because $\bar{p}_{12}(0) = 0$ and $\bar{p}_{12}$ is continuous, there exists $T_- \in (0, T_+)$ such that

$$\sup_{0 \le t \le T_-} |\bar{p}_{12}(t)| < \frac{c}{2}. \tag{325}$$

Let $\tilde{p}_{12}^{(d)}(t)$ denote the piecewise-constant interpolation of the empirical structural overlap. Since

$$\tilde{p}_{12}^{(d)} = \tilde{q}_{12}^{(d)} - \tilde{t}_1^{(d)}\tilde{t}_2^{(d)}, \tag{326}$$

Theorem 2 implies that, for every fixed finite T,

$$\sup_{t \in [0,T]} \left|\tilde{p}_{12}^{(d)}(t) - \bar{p}_{12}(t)\right| \xrightarrow{\mathbb{P}} 0. \tag{327}$$

Applying (327) on $[0, T_+]$ and using (324), we obtain

$$\mathbb{P}\left(|\tilde{p}_{12}^{(d)}(T_+)| \ge c\right) \longrightarrow 1. \tag{328}$$

Hence, with probability tending to one,

$$N_{\mathrm{wr}}^{(d)}(c) \le \left\lceil \frac{T_+}{\tau_d} \right\rceil = \left\lceil \frac{T_+ d}{4\alpha_d} \right\rceil. \tag{329}$$

Likewise, by (325) and (327),

$$\mathbb{P}\left(\sup_{0 \le t \le T_-} |\tilde{p}_{12}^{(d)}(t)| < c\right) \longrightarrow 1. \tag{330}$$

Therefore, with probability tending to one,

$$N_{\mathrm{wr}}^{(d)}(c) > \left\lfloor \frac{T_-}{\tau_d} \right\rfloor = \left\lfloor \frac{T_- d}{4\alpha_d} \right\rfloor. \tag{331}$$

Combining (329) and (331) proves

$$N_{\mathrm{wr}}^{(d)}(c) = \Theta_{\mathbb{P}}\left(\frac{d}{\alpha_d}\right). \tag{332}$$

Finally, choosing

$$\alpha_d = \Theta\left(\frac{1}{d \log d}\right), \tag{333}$$

which is the largest learning rate covered by Theorem 2, gives

$$N_{\mathrm{wr}}^{(d)}(c) = \Theta_{\mathbb{P}}(d^2 \log d). \tag{334}$$

This completes the proof.

D.5 Convergence rate of strong recovery

In this section we study the convergence rate after weak recovery. Because the weight decay depends on the first moments of the matrices, it is convenient to introduce

$$t_k := \mu_{(k)} = \frac{1}{d}\mathrm{Tr}[S_k], \qquad q_{kl} := \mu_{(k,l)} = \frac{1}{d}\mathrm{Tr}[S_k S_l]. \tag{335}$$

The limiting regularized population loss is defined as

$$\Psi(t,q) := \Phi(q) + \frac{\gamma}{2}\sum_{k\in\mathcal{K}} t_k. \tag{336}$$

Let $\mathcal{Q}_+^{(1,2)}$ denote the asymptotically realizable domain of the pair (t,q), namely the closure of all limits generated by sequences of PSD matrices $S_k = W_k W_k^T \succeq 0$, subject to the fixed teacher components when $k \notin \mathcal{K}$. We define

$$\Psi^* := \inf_{(t,q)\in\mathcal{Q}_+^{(1,2)}} \Psi(t,q). \tag{337}$$

We refer to the convergence to the global minimizers of Ψ as "strong recovery".

Let $\bar{t}_k(t) := \bar{\mu}_{(k)}(t)$ and $\bar{q}_{kl}(t) := \bar{\mu}_{(k,l)}(t)$ denote the corresponding components of the solution of (8). We impose an effective Polyak–Łojasiewicz condition along the trajectory.

Assumption 6. *There exist constants $T_0 < \infty$ and $c > 0$ such that, for every $t \geq T_0$,*

$$-\frac{d}{dt}\Psi(\bar{t}(t),\bar{q}(t)) \geq 2c\left(\Psi(\bar{t}(t),\bar{q}(t)) - \Psi^*\right). \tag{338}$$

For a finite d, define the regularized population loss

$$\Psi_d(W) := \mathbb{E}_x\left[\mathcal{L}\left(\left\{\frac{1}{\sqrt{d}}\mathrm{Tr}\left[W_k W_k^T (x_j x_i^T - \mathcal{C}_{ij} I_d)\right]\right\}_{i,j,k=1}^{L,L,K}\right)\right] + \frac{\gamma}{2d}\sum_{k\in\mathcal{K}}\|W_k\|_F^2. \tag{339}$$

Corollary 6. *Suppose that the conditions of Theorem 2 and Assumption 6 hold. Let $n_0 := \left\lceil \frac{T_0 d}{4\alpha_d} \right\rceil$. Then, for every fixed target accuracy $\epsilon > 0$, there exists*

$$n_\epsilon = n_0 + O\left(\frac{d}{\alpha_d}\log_+\frac{1}{\epsilon}\right), \tag{340}$$

such that

$$\lim_{d\to\infty}\mathbb{P}\left(\Psi_d(W^{(n_\epsilon)}) - \Psi^* > \epsilon\right) = 0. \tag{341}$$

where $\log_+(x) := \max\{0, \log x\}$. In particular, if $T_0 = \Theta(1)$, reaching an ϵ-suboptimal population loss requires at most $O\left(\frac{d}{\alpha_d}\log\frac{1}{\epsilon}\right)$ additional samples.

Proof. We first analyze the limiting trajectory in Theorem 2. Define

$$\Delta(t) := \Psi(\bar{t}(t),\bar{q}(t)) - \Psi^* \geq 0. \tag{342}$$

By Assumption 6, for every $t \geq T_0$,

$$\frac{d}{dt}\Delta(t) \leq -2c\Delta(t). \tag{343}$$

Hence Grönwall's inequality gives

$$\Delta(t) \leq \Delta(T_0)e^{-2c(t-T_0)}, \qquad t \geq T_0. \tag{344}$$

Let

$$\Delta_0 := \Psi(\bar{t}(T_0),\bar{q}(T_0)) - \Psi^*. \tag{345}$$

If $\Delta_0 = 0$, there is nothing to prove. Otherwise, for any fixed $\epsilon > 0$, define

$$T_\epsilon := T_0 + \frac{1}{2c}\log_+\left(\frac{2\Delta_0}{\epsilon}\right). \tag{346}$$

Then

$$\Psi(\bar{t}(T_\epsilon), \bar{q}(T_\epsilon)) - \Psi^* \leq \frac{\epsilon}{2}. \tag{347}$$

We next transfer this bound to the finite-dimensional population loss. Recall that the discrete steps and continuous times are related by $t = n\tau_d$ with $\tau_d = \frac{4\alpha_d}{d}$. For $t \geq 0$, define $W^{(d)}(t) := W^{(\lfloor t/\tau_d \rfloor)}$. For every fixed finite T, Theorem 2, together with the uniform spectral bound of Lemma 3, implies

$$\sup_{t\in[0,T]} \left| \Psi_d(W^{(d)}(t)) - \Psi(\bar{t}(t), \bar{q}(t)) \right| \xrightarrow{\mathbb{P}} 0. \tag{348}$$

This is because on the spectral-bound event, the same argument used in Lemma 6, together with Corollary 5 and uniform integrability, gives uniformly over the spectrally bounded class

$$\mathbb{E}_x[\mathcal{L}(G(x; W))] - \Phi(q(W)) \longrightarrow 0. \tag{349}$$

The regularization term is exactly

$$\frac{\gamma}{2d} \sum_{k\in\mathcal{K}} \|W_k\|_F^2 = \frac{\gamma}{2} \sum_{k\in\mathcal{K}} t_k. \tag{350}$$

The convergence of the first- and second-order moments supplied by Theorem 2, together with the continuity of Φ, therefore yields (348).

For the fixed accuracy $\epsilon > 0$, the time T_ϵ in (346) is finite and independent of d. Thus (348), evaluated on $[0, T_\epsilon]$, implies

$$\mathbb{P}\left(\left| \Psi_d(W^{(d)}(T_\epsilon)) - \Psi(\bar{t}(T_\epsilon), \bar{q}(T_\epsilon)) \right| > \frac{\epsilon}{2} \right) \longrightarrow 0. \tag{351}$$

Set $n_\epsilon := \left\lceil \frac{T_\epsilon d}{4\alpha_d} \right\rceil$. Therefore, using (347),

$$\mathbb{P}\left(\Psi_d(W^{(n_\epsilon)}) - \Psi^* > \epsilon \right) \longrightarrow 0. \tag{352}$$

Finally, by (346) and the definition of n_ϵ,

$$n_\epsilon - n_0 = O\left(\frac{d}{\alpha_d} \log_+ \frac{1}{\epsilon} \right), \tag{353}$$

which proves the claimed sample complexity. □

Example. *Consider the setting $\gamma = 0$ and $\mathcal{L}(G_1, G_2) = (G_1 - G_2)^2$ with one learnable student $S_1 = W_1 W_1^T$ and one fixed teacher S_2. Then*

$$\Phi(q) = 2q_{11} + 2q_{22} - 4q_{12} = 2 \lim_{d\to\infty} \frac{1}{d} \mathrm{Tr}[(S_1 - S_2)^2]. \tag{354}$$

If the teacher is normalized by $q_{22} = 1$, the global minimum $\Phi^* = 0$ is attained at perfect recovery, $q_{11} = q_{12} = 1$.

At finite d, up to the normalization induced by the time scaling, the corresponding population loss is proportional to $\frac{1}{d}\mathrm{Tr}[(S_1 - S_2)^2]$. Its gradient with respect to W_1 satisfies

$$\|\nabla_{W_1} \mathcal{L}\|_F^2 \propto \frac{1}{d^2} \mathrm{Tr}[S_1(S_1 - S_2)^2]. \tag{355}$$

Hence the effective PL condition reduces, up to fixed normalization constants, to a bound of the form

$$\mathrm{Tr}[S_1(S_1 - S_2)^2] \geq c_0 \mathrm{Tr}[(S_1 - S_2)^2] \tag{356}$$

for some $c_0 > 0$. A sufficient condition for (356) *is $S_1 \succeq c_0 I_d$. More generally, it requires non-degeneracy only along the directions relevant to $S_1 - S_2$.*

E Online SGD over S

In this section, we study online SGD directly over the attention matrices $\{S_k\}_{k=1}^K$. At each step n, a fresh sample $x^{(n)} \sim \mathcal{N}(0, \mathcal{C} \otimes I_d)$ is drawn, and for $k \in \{1, \dots, K\}$ we consider the update

$$S_k^{(n+1)} = S_k^{(n)} - \alpha_d \chi_k \left(\nabla_{S_k} \mathcal{L}(G^{(n)}(x^{(n)})) + \frac{\gamma}{d} S_k^{(n)} \right), \tag{357}$$

where $\chi_k := \mathbf{1}_{\{k \in \mathcal{K}\}}$ so that S_k remains fixed whenever $k \notin \mathcal{K}$. The tensor G in (357) is defined as

$$G_{ijk}^{(n)} := \frac{1}{\sqrt{d}} \mathrm{Tr}\left[S_k^{(n)} \left(x_j^{(n)} (x_i^{(n)})^T - \mathcal{C}_{ij} I_d \right) \right]. \tag{358}$$

We introduce the time as $\tau_d := \frac{\alpha_d}{d}$ and $t := n\tau_d$.

The order parameters are defined as:

$$t_k^{(d)} := \frac{1}{d} \mathrm{Tr}[S_k], \qquad \Psi_{kl}^{(d)} := \frac{1}{d} \mathrm{Tr}[S_k S_l], \qquad \Omega_{kl}^{(d)} := \frac{1}{d} \mathrm{Tr}[S_k S_l^T]. \tag{359}$$

Moreover, writing $S_k^+ := \frac{S_k + S_k^T}{2}$ and $S_k^- := \frac{S_k - S_k^T}{2}$, we have $\frac{\Omega_{kl} + \Psi_{kl}}{2} = \frac{1}{d} \langle S_k^+, S_l^+ \rangle_F$ and $\frac{\Omega_{kl} - \Psi_{kl}}{2} = \frac{1}{d} \langle S_k^-, S_l^- \rangle_F$. Consequently, the feasible domain of the order parameters is

$$\mathcal{D} := \left\{ (\Psi, \Omega) \in \mathbb{S}^K \times \mathbb{S}^K : \frac{\Omega + \Psi}{2} \succeq 0, \quad \frac{\Omega - \Psi}{2} \succeq 0 \right\}. \tag{360}$$

For $(\Psi, \Omega) \in \mathcal{D}$, define the potential

$$\Phi(\Psi, \Omega) := \mathbb{E}_{G_{\Psi,\Omega}} \left[\mathcal{L}(G_{\Psi,\Omega}) \right], \tag{361}$$

where $G_{\Psi,\Omega} \in \mathbb{R}^{L \times L \times K}$ is a centered Gaussian tensor with covariance $\mathrm{Cov}\left((G_{\Psi,\Omega})_{ijk}, (G_{\Psi,\Omega})_{i'j'l} \right) = \mathcal{C}_{ii'} \mathcal{C}_{jj'} \Omega_{kl} + \mathcal{C}_{ij'} \mathcal{C}_{ji'} \Psi_{kl}$. We denote by $\nabla_\Psi \Phi$ and $\nabla_\Omega \Phi$ the corresponding matrix derivatives w.r.t. Ψ and Ω.

Assumption 7. *The following conditions hold.*

1. *The loss $\mathcal{L} : \mathbb{R}^{L \times L \times K} \to \mathbb{R}$ is four times continuously differentiable, and $\mathcal{L}$ together with its derivatives up to fourth order have at most polynomial growth.*
2. *There exists a constant $C_0 > 0$, independent of d, such that $\max_{1 \le k \le K} \|S_k^{(0)}\|_{\mathrm{op}} \le C_0$ almost surely.*
3. *The initial order parameters converge in probability to deterministic limits: $t_k^{(d)}(0) \to t_k(0)$, $\Psi_{kl}^{(d)}(0) \to \bar{\Psi}_{kl}(0)$ and $\Omega_{kl}^{(d)}(0) \to \bar{\Omega}_{kl}(0)$ for every k, l.*
4. *The regularization is sufficiently large, i.e., $\gamma > \gamma_*(C_0, \mathcal{C}, \mathcal{L}, K, L)$ for some γ_* independent of d.*

Let $\tilde{S}_k^{(d)}(t)$ denote the piecewise-constant interpolation

$$\tilde{S}_k^{(d)}(t) = S_k^{(n)}, \qquad t \in [n\tau_d, (n+1)\tau_d), \tag{362}$$

and define $\tilde{t}^{(d)}$, $\tilde{\Psi}^{(d)}$, and $\tilde{\Omega}^{(d)}$ in the same way according to (359).

The following theorem shows that, unlike Theorems 2 and 3, direct optimization over S closes at the level of finitely many first- and second-order parameters.

Theorem 4. *Under Assumption 7, suppose that the learning rate satisfies $\alpha_d \le \frac{c_0}{d \log d}$ and $\alpha_d = \Omega(d^{-\iota})$ for some constants $\iota > 0$ and $c_0 > 0$ sufficiently small. Then, for every finite $T > 0$, the empirical order parameters converge uniformly in probability to a deterministic trajectory $(\bar{t}(t), \bar{\Psi}(t), \bar{\Omega}(t))$. More precisely, for every k, l and every $\epsilon > 0$,*

$$\lim_{d \to \infty} \mathbb{P}\left(\sup_{t \in [0,T]} |\tilde{t}_k^{(d)}(t) - \bar{t}_k(t)| + |\tilde{\Psi}_{kl}^{(d)}(t) - \bar{\Psi}_{kl}(t)| + |\tilde{\Omega}_{kl}^{(d)}(t) - \bar{\Omega}_{kl}(t)| > \epsilon \right) = 0. \tag{363}$$

The limiting trajectory is the unique solution of

$$\frac{d}{dt}\bar{t}_k = -2\chi_k \sum_{m=1}^{K} \left([\nabla_\Omega \Phi]_{km} + [\nabla_\Psi \Phi]_{km}\right) \bar{t}_m - \gamma \chi_k \bar{t}_k, \tag{364}$$

$$\frac{d}{dt}\bar{\Psi}_{kl} = -2\chi_k \sum_{m=1}^{K} \left([\nabla_\Omega \Phi]_{km} \bar{\Psi}_{ml} + [\nabla_\Psi \Phi]_{km} \bar{\Omega}_{ml}\right) - 2\chi_l \sum_{m=1}^{K} \left([\nabla_\Omega \Phi]_{lm} \bar{\Psi}_{km} + [\nabla_\Psi \Phi]_{lm} \bar{\Omega}_{km}\right) - \gamma(\chi_k + \chi_l)\bar{\Psi}_{kl}, \tag{365}$$

$$\frac{d}{dt}\bar{\Omega}_{kl} = -2\chi_k \sum_{m=1}^{K} \left([\nabla_\Omega \Phi]_{km} \bar{\Omega}_{ml} + [\nabla_\Psi \Phi]_{km} \bar{\Psi}_{ml}\right) - 2\chi_l \sum_{m=1}^{K} \left([\nabla_\Omega \Phi]_{lm} \bar{\Omega}_{km} + [\nabla_\Psi \Phi]_{lm} \bar{\Psi}_{km}\right) - \gamma(\chi_k + \chi_l)\bar{\Omega}_{kl}, \tag{366}$$

where the gradients of Φ are evaluated at $(\bar{\Psi}(t), \bar{\Omega}(t)) \in \mathcal{D}$ for $t \in [0, T]$.

Building upon this uniform convergence, now we consider the weak recovery. First define the traceless parts

$$\mathring{S}_k := S_k - t_k I_d, \tag{367}$$

and the structural overlaps

$$p^{(d)}_{\Psi,kl} := \frac{1}{d}\mathrm{Tr}[\mathring{S}_k \mathring{S}_l] = \Psi^{(d)}_{kl} - t^{(d)}_k t^{(d)}_l, \tag{368}$$

$$p^{(d)}_{\Omega,kl} := \frac{1}{d}\mathrm{Tr}[\mathring{S}_k \mathring{S}_l^T] = \Omega^{(d)}_{kl} - t^{(d)}_k t^{(d)}_l. \tag{369}$$

It is useful to resolve these overlaps into their symmetric and skew-symmetric channels:

$$p_{+,12} := \frac{p_{\Omega,12} + p_{\Psi,12}}{2}, \qquad p_{-,12} := \frac{p_{\Omega,12} - p_{\Psi,12}}{2}. \tag{370}$$

For the teacher, define the corresponding signal strengths

$$a_+ := \frac{p_{\Omega,22} + p_{\Psi,22}}{2} = \lim_{d\to\infty} \frac{1}{d} \left\| \frac{\mathring{S}_2 + \mathring{S}_2^T}{2} \right\|_F^2, \tag{371}$$

and

$$a_- := \frac{p_{\Omega,22} - p_{\Psi,22}}{2} = \lim_{d\to\infty} \frac{1}{d} \left\| \frac{\mathring{S}_2 - \mathring{S}_2^T}{2} \right\|_F^2. \tag{372}$$

Thus $a_+, a_- \geq 0$. The following is the uninformative initialization and informative teacher assumption.

Assumption 8. *As $d \to \infty$, the initial student $S_1(0)$ and the fixed teacher S_2 satisfy*

$$\bar{t}_1(0) = 0, \qquad \bar{p}_{\Psi,12}(0) = 0, \qquad \bar{p}_{\Omega,12}(0) = 0. \tag{373}$$

Moreover, the teacher is informative in the sense that $a_+ + a_- > 0$.

For example, (373) is satisfied by the standard Gaussian initialization with centered i.i.d. entries of variance $\Theta(d^{-1})$.

Define the effective gradients in the symmetric and skew-symmetric parts by

$$g_+ := [\nabla_\Omega \Phi]_{12} + [\nabla_\Psi \Phi]_{12}, \qquad g_- := [\nabla_\Omega \Phi]_{12} - [\nabla_\Psi \Phi]_{12}. \tag{374}$$

For a fixed $c > 0$, define the weak-recovery sample complexity

$$N^{(d)}_{\mathrm{wr},S}(c) := \inf\left\{ n \geq 0 : \max\left(|p^{(d,n)}_{\Psi,12}|, |p^{(d,n)}_{\Omega,12}| \right) \geq c \right\}. \tag{375}$$

The following corollary gives the weak recovery conditions.

Corollary 7. *Consider* (357) *with $\mathcal{K} = \{1\}$, where S_1 is the trainable student and S_2 is a fixed teacher. Suppose the conditions of Theorem 4 and Assumption 8 hold.*

1. *If, at initialization,*

$$a_+ g_+(0) \neq 0 \qquad \text{or} \qquad a_- g_-(0) \neq 0, \tag{376}$$

then there exist constants $c > 0$ and $0 < T_- < T_+ < \infty$, independent of d, such that

$$\lim_{d\to\infty} \mathbb{P}\left(\left\lfloor \frac{T_- d}{\alpha_d} \right\rfloor < N^{(d)}_{\mathrm{wr},S}(c) \leq \left\lceil \frac{T_+ d}{\alpha_d} \right\rceil \right) = 1. \tag{377}$$

In particular, choosing the largest learning-rate scaling $\alpha_d = \Theta\left(\frac{1}{d \log d}\right)$ covered by Theorem 4 gives the weak recovery sample complexity

$$N^{(d)}_{\mathrm{wr},S}(c) = \Theta_{\mathbb{P}}(d^2 \log d). \tag{378}$$

2. *Let the uninformative manifold be $\mathcal{U} := \{(t, \Psi, \Omega) : t_1 = 0, \Psi_{12} = 0, \Omega_{12} = 0\}$. Suppose that*

$$[\nabla_\Omega \Phi(\Psi, \Omega)]_{12} = [\nabla_\Psi \Phi(\Psi, \Omega)]_{12} = 0 \qquad \text{for every state in } \mathcal{U}. \tag{379}$$

Then for every fixed $T < \infty$ and every $\epsilon > 0$,

$$\lim_{d\to\infty} \mathbb{P}\left(\sup_{t\in[0,T]} \max\left(|p^{(d)}_{\Psi,12}(t)|, |p^{(d)}_{\Omega,12}(t)| \right) > \epsilon \right) = 0. \tag{380}$$

In particular, for any learning-rates covered by Theorem 4, S-SGD cannot achieve weak recovery within $\mathcal{O}(d^2 \log d)$ samples.

E.1 Proof of Theorem 4

The proof combines uniform spectral bounds based on the Matrix Freedman inequality with a martingale expansion of the empirical order parameters, and concludes by applying Grönwall's inequality to the resulting approximate ODE.

E.1.1 Uniform spectral bounds

Lemma 8. *Under the conditions of Theorem 4, there exists a constant $C_* > C_0$, independent of d and T, such that for every fixed $T > 0$,*

$$\lim_{d\to\infty} \mathbb{P}\left(\sup_{0\leq n\leq N_T} \max_{1\leq k\leq K} \|S^{(n)}_k\|_{\mathrm{op}} \leq C_* \right) = 1, \tag{381}$$

where $N_T := \left\lfloor \frac{T}{\tau_d} \right\rfloor = \left\lfloor \frac{Td}{\alpha_d} \right\rfloor$.

The proof follows essentially the same stopping-time and Matrix Freedman argument as Lemma 3. The main difference is that direct optimization over S_k yields an expected gradient of the form $d^{-1}H_k$ rather than $d^{-1}H_k W_k$ (see Lemma 2); moreover, since S_k need not be symmetric, H_k contains both S_l and S_l^T.

Proof. The matrices with $k \notin \mathcal{K}$ remain fixed, and hence satisfy $\|S^{(n)}_k\|_{\mathrm{op}} \leq C_0$. It therefore suffices to control the trainable indices $k \in \mathcal{K}$.

Fix $C_* > C_0$, to be chosen below, and define the stopping time

$$J := \inf\left\{ n \geq 0 : \max_{1\leq k\leq K} \|S^{(n)}_k\|_{\mathrm{op}} > C_* \right\}. \tag{382}$$

On $\{n < J\}$,

$$\max_k \|S^{(n)}_k\|_{\mathrm{op}} \leq C_*. \tag{383}$$

Step 1: Truncation of the fresh sample. As in the proof of Lemma 3, define

$$\mathcal{A}_n := \left\{ \|x^{(n)}\|^2 \le c_x d \right\} \cap \left\{ \max_{i,j,k} |G^{(n)}_{ijk}| \le c_G \log d \right\}, \tag{384}$$

where $c_x, c_G > 0$ are sufficiently large constants. Similarly to Lemma 3, we have

$$\mathbf{1}_{\{n<J\}} \mathbb{P}(\mathcal{A}_n^c \mid \mathcal{F}_n) \le p_d, \tag{385}$$

where $p_d \le C(e^{-c_1 d} + e^{-c_2 (\log d)^2})$. Consequently, defining

$$\mathcal{A}_{\text{data}} := \bigcap_{n=0}^{N_T - 1} \left(\{n \ge J\} \cup \mathcal{A}_n \right), \tag{386}$$

we have

$$\mathbb{P}(\mathcal{A}_{\text{data}}^c) \le N_T p_d \longrightarrow 0, \tag{387}$$

because $\alpha_d = \Omega(d^{-\iota})$.

Step 2: Conditional drift and stopped martingale noise. For $k \in \mathcal{K}$, define the gradient

$$g_k^{(n)} := \nabla_{S_k} \mathcal{L}(G^{(n)}) = \sum_{i,j=1}^{L} \partial_{(ijk)} \mathcal{L}(G^{(n)}) X_{ij}^{(n)}, \tag{388}$$

where $X_{ij}^{(n)} := \frac{1}{\sqrt{d}} \left(x_i^{(n)} (x_j^{(n)})^T - \mathcal{C}_{ij} I_d \right)$. Let $\bar{g}_k^{(n)} := \mathbb{E}[g_k^{(n)} \mid \mathcal{F}_n]$. The same calculation as in Lemma 2 gives

$$\bar{g}_k^{(n)} = \frac{1}{d} H_k^{(n)} + \mathcal{E}_{\text{Stein},k}^{(n)}, \tag{389}$$

where

$$H_k^{(n)} := \sum_{i,j,i',j'=1}^{L} \sum_{l=1}^{K} \mathbb{E}_x \left[\partial^2_{(ijk),(i'j'l)} \mathcal{L}(G^{(n)}) \right] \left(\mathcal{C}_{ii'} \mathcal{C}_{jj'} S_l^{(n)} + \mathcal{C}_{ij'} \mathcal{C}_{ji'} (S_l^{(n)})^T \right). \tag{390}$$

Using Assumption 7.1 and the uniform moment bounds for $G^{(n)}$ under (383), the same estimates as in Lemma 2 yield

$$\|H_k^{(n)}\|_{\text{op}} \le C_H(C_*), \qquad \|\mathcal{E}_{\text{Stein},k}^{(n)}\|_{\text{op}} \le C_E(C_*) d^{-3/2}, \tag{391}$$

uniformly on $\{n < J\}$.

Define the stopped and truncated gradient

$$\widehat{g}_k^{(n)} := \mathbf{1}_{\{n<J\}} g_k^{(n)} \mathbf{1}_{\mathcal{A}_n}, \tag{392}$$

its conditional mean

$$\widehat{\bar{g}}_k^{(n)} := \mathbb{E}[\widehat{g}_k^{(n)} \mid \mathcal{F}_n], \tag{393}$$

and the martingale difference

$$\widehat{\xi}_k^{(n)} := \widehat{g}_k^{(n)} - \widehat{\bar{g}}_k^{(n)}. \tag{394}$$

Then

$$\widehat{\bar{g}}_k^{(n)} = \mathbf{1}_{\{n<J\}} \left(\bar{g}_k^{(n)} - \mathcal{T}_k^{(n)} \right), \qquad \mathcal{T}_k^{(n)} := \mathbb{E}\left[g_k^{(n)} \mathbf{1}_{\mathcal{A}_n^c} \mid \mathcal{F}_n \right]. \tag{395}$$

As in the proof of Lemma 3, Hölder's inequality and moment bounds give

$$\mathbf{1}_{\{n<J\}} \max \left\{ \left\| \mathbb{E}[\widehat{\xi}_k^{(n)} (\widehat{\xi}_k^{(n)})^T \mid \mathcal{F}_n] \right\|_{\text{op}}, \left\| \mathbb{E}[(\widehat{\xi}_k^{(n)})^T \widehat{\xi}_k^{(n)} \mid \mathcal{F}_n] \right\|_{\text{op}} \right\} \le C_g(C_*). \tag{396}$$

On $\mathcal{A}_n \cap \{n < J\}$, we also have

$$\|\widehat{g}_k^{(n)}\|_{\text{op}} \le C\sqrt{d}\,(1 + (\log d)^r) =: B_d \tag{397}$$

for some fixed integer r, and consequently

$$\|\widehat{\xi}_k^{(n)}\|_{\text{op}} \le 2B_d. \tag{398}$$

Finally, by Cauchy–Schwarz and (385), the truncation error satisfies

$$\mathbf{1}_{\{n<J\}} \|\mathcal{T}_k^{(n)}\|_{\text{op}} \le r_d, \qquad r_d = o(d^{-M}) \tag{399}$$

for every fixed $M > 0$.

Step 3: Uniform control of the martingale noise. Set $a_d := 1 - \frac{\alpha_d \gamma}{d} = 1 - \tau_d \gamma$. For sufficiently large d, $a_d \in (0,1)$. Define $D_{k,n} := -\alpha_d \widehat{\xi}_k^{(n)}$ and, for each terminal time $m \geq 1$,

$$Y_{k,m} := \sum_{n=0}^{m-1} a_d^{m-1-n} D_{k,n}. \tag{400}$$

For each fixed m, the summands form a square matrix martingale difference sequence. By (398),

$$\|a_d^{m-1-n} D_{k,n}\|_{\mathrm{op}} \leq 2\alpha_d B_d =: R_d, \tag{401}$$

where $R_d = o((\log d)^{-1})$ under $\alpha_d \leq c_0 (d \log d)^{-1}$. Moreover, by (396),

$$\sigma_d^2 \leq C_g \alpha_d^2 \sum_{r=0}^{\infty} a_d^{2r} = \frac{C_g \alpha_d^2}{1 - a_d^2} \leq C \frac{d\alpha_d}{\gamma} \leq \frac{C c_0}{\gamma \log d}. \tag{402}$$

Hence the Matrix Freedman inequality gives, for every fixed $\delta > 0$,

$$\mathbb{P}(\|Y_{k,m}\|_{\mathrm{op}} \geq \delta) \leq 2d \exp\left[-\frac{\delta^2/2}{\sigma_d^2 + R_d \delta/3} \right] \leq 2d^{1 - c\delta^2/c_0} \tag{403}$$

for all sufficiently large d. Since $N_T \leq C_T d^{1+\iota}$ and K is fixed, a union bound gives

$$\mathbb{P}\left(\max_{k \in \mathcal{K}} \max_{1 \leq m \leq N_T} \|Y_{k,m}\|_{\mathrm{op}} > \delta \right) \leq C_T d^{2+\iota - c\delta^2/c_0}. \tag{404}$$

Choosing $c_0 > 0$ sufficiently small yields

$$\mathbb{P}(\mathcal{A}_{\mathrm{noise}}^c) \longrightarrow 0, \tag{405}$$

where

$$\mathcal{A}_{\mathrm{noise}} := \left\{ \max_{k \in \mathcal{K}} \max_{1 \leq m \leq N_T} \|Y_{k,m}\|_{\mathrm{op}} \leq \delta \right\}. \tag{406}$$

Step 4: Closing the stopping-time argument. On $\mathcal{A}_{\mathrm{data}}$, for every $m \leq J$ and every $n < m$,

$$g_k^{(n)} = \bar{g}_k^{(n)} - \mathcal{T}_k^{(n)} + \widehat{\xi}_k^{(n)}. \tag{407}$$

For $k \in \mathcal{K}$, unrolling (357) therefore gives

$$S_k^{(m)} = a_d^m S_k^{(0)} - \alpha_d \sum_{n=0}^{m-1} a_d^{m-1-n} \bar{g}_k^{(n)} + \alpha_d \sum_{n=0}^{m-1} a_d^{m-1-n} \mathcal{T}_k^{(n)} + Y_{k,m}. \tag{408}$$

Since $\alpha_d \sum_{r=0}^{\infty} a_d^r = \frac{d}{\gamma}$, (391) and (399) imply that, on $\mathcal{A}_{\mathrm{data}} \cap \mathcal{A}_{\mathrm{noise}}$ and for $m \leq J \wedge N_T$,

$$\|S_k^{(m)}\|_{\mathrm{op}} \leq C_0 + \frac{C_H(C_*)}{\gamma} + \frac{C_E(C_*)}{\gamma\sqrt{d}} + \frac{d r_d}{\gamma} + \delta = C_0 + \frac{C_H(C_*)}{\gamma} + \delta + o(1). \tag{409}$$

Fix, for instance, $C_* := 2C_0 + 1$ and $\delta := \frac{C_* - C_0}{4}$. By Assumption 7.4, γ can be chosen sufficiently large so that

$$\frac{C_H(C_*)}{\gamma} \leq \frac{C_* - C_0}{4}. \tag{410}$$

Then, for all sufficiently large d,

$$\max_k \|S_k^{(m)}\|_{\mathrm{op}} < C_* \qquad \text{for every } m \leq J \wedge N_T \tag{411}$$

on $\mathcal{A}_{\mathrm{data}} \cap \mathcal{A}_{\mathrm{noise}}$.

If $J \leq N_T$, taking $m = J$ gives $\max_k \|S_k^{(J)}\|_{\mathrm{op}} < C_*$, contradicting the definition of J. Hence $J > N_T$ on $\mathcal{A}_{\mathrm{data}} \cap \mathcal{A}_{\mathrm{noise}}$ for all sufficiently large d.

Finally,

$$\mathbb{P}(J \leq N_T) \leq \mathbb{P}(\mathcal{A}_{\mathrm{data}}^c) + \mathbb{P}(\mathcal{A}_{\mathrm{noise}}^c) \longrightarrow 0 \tag{412}$$

by (387) and (405). This proves the lemma. □

E.1.2 Moment expansion

Let

$$\mathbf{Q}^{(d)}(t) := \Big(t^{(d)}(t), \Psi^{(d)}(t), \Omega^{(d)}(t)\Big) \tag{413}$$

denote the piecewise-constant interpolation of the empirical order parameters, and let $F = (F_t, F_\Psi, F_\Omega)$ denote the finite-dimensional vector field on the right-hand sides of (364)–(366).

Lemma 9. *Under the conditions of Theorem 4, for every fixed $T > 0$,*

$$\mathbf{Q}^{(d)}(t) = \mathbf{Q}^{(d)}(0) + \int_0^t F(\mathbf{Q}^{(d)}(s))\, ds + E^{(d)}_{\text{total}}(t), \qquad t \in [0,T], \tag{414}$$

where

$$\sup_{t\in[0,T]} \Big\| E^{(d)}_{\text{total}}(t) \Big\|_\infty \xrightarrow{\mathbb{P}} 0. \tag{415}$$

The proof follows the same martingale-decomposition and drift-convergence strategy as Lemma 4. The main simplification is that direct optimization over S closes at the finite-dimensional level of (t, Ψ, Ω), so no higher-order moments or weighted moment-space estimates are required; the only additional feature is the simultaneous appearance of S_l and S_l^T due to the possible asymmetry of S_l.

Proof. Let C_* be the constant in Lemma 8, and define

$$J := \inf\left\{ n \geq 0 : \max_{1\leq k\leq K} \|S_k^{(n)}\|_{\text{op}} > C_* \right\}. \tag{416}$$

Then, for $N_T := \left\lfloor \frac{T}{\tau_d} \right\rfloor = \left\lfloor \frac{Td}{\alpha_d} \right\rfloor$, Lemma 8 gives

$$\mathbb{P}(J > N_T) \longrightarrow 1. \tag{417}$$

All estimates below are uniform on $\{n < J\}$.

Step 1: Conditional drift. Recall from the proof of Lemma 8 that

$$\bar{g}_k^{(n)} = \frac{1}{d} H_k^{(n)} + \mathcal{E}^{(n)}_{\text{Stein},k}, \qquad \|\mathcal{E}^{(n)}_{\text{Stein},k}\|_{\text{op}} \leq C_E d^{-3/2}, \tag{418}$$

where

$$H_k^{(n)} = \sum_{i,j,i',j',m} \mathbb{E}_x\left[\partial^2_{(ijk),(i'j'm)} \mathcal{L}(G^{(n)})\right] \Big(\mathcal{C}_{ii'}\mathcal{C}_{jj'} S_m^{(n)} + \mathcal{C}_{ij'}\mathcal{C}_{ji'} (S_m^{(n)})^T\Big). \tag{419}$$

Define the finite-dimensional coefficients

$$\Gamma^{(d)}_{\Omega,km} := \frac{1}{2} \sum_{i,j,i',j'} \mathcal{C}_{ii'}\mathcal{C}_{jj'}\, \mathbb{E}_x\left[\partial^2_{(ijk),(i'j'm)} \mathcal{L}(G^{(n)})\right], \tag{420}$$

$$\Gamma^{(d)}_{\Psi,km} := \frac{1}{2} \sum_{i,j,i',j'} \mathcal{C}_{ij'}\mathcal{C}_{ji'}\, \mathbb{E}_x\left[\partial^2_{(ijk),(i'j'm)} \mathcal{L}(G^{(n)})\right]. \tag{421}$$

Then

$$H_k^{(n)} = 2 \sum_{m=1}^K \Big(\Gamma^{(d)}_{\Omega,km} S_m^{(n)} + \Gamma^{(d)}_{\Psi,km} (S_m^{(n)})^T\Big). \tag{422}$$

We claim that, uniformly over the spectrally bounded class,

$$\delta_d(C_*) := \sup_{\max_m \|S_m\|_{\text{op}} \leq C_*} \max_{k,m} \Big(\Big|\Gamma^{(d)}_{\Omega,km} - [\nabla_\Omega \Phi(\Psi,\Omega)]_{km}\Big| + \Big|\Gamma^{(d)}_{\Psi,km} - [\nabla_\Psi \Phi(\Psi,\Omega)]_{km}\Big| \Big) \longrightarrow 0. \tag{423}$$

The proof is the same subsequence argument as in Lemma 6.

Step 2: One-step expansion of the order parameters. Define

$$\xi_k^{(n)} := g_k^{(n)} - \bar{g}_k^{(n)}. \tag{424}$$

Using $\tau_d = \alpha_d/d$, the SGD update can be written as

$$\Delta S_k^{(n)} := S_k^{(n+1)} - S_k^{(n)} = -\tau_d \chi_k \left(H_k^{(n)} + \gamma S_k^{(n)} \right) - \alpha_d \chi_k \xi_k^{(n)} - \alpha_d \chi_k \mathcal{E}_{\mathrm{Stein},k}^{(n)}. \tag{425}$$

For the normalized trace,

$$t_k^{(n+1)} - t_k^{(n)} = \frac{1}{d} \mathrm{Tr}[\Delta S_k^{(n)}]. \tag{426}$$

For the two second-order overlaps,

$$\Psi_{kl}^{(n+1)} - \Psi_{kl}^{(n)} = \frac{1}{d} \mathrm{Tr}[\Delta S_k^{(n)} S_l^{(n)} + S_k^{(n)} \Delta S_l^{(n)}] + \frac{1}{d} \mathrm{Tr}[\Delta S_k^{(n)} \Delta S_l^{(n)}], \tag{427}$$

$$\Omega_{kl}^{(n+1)} - \Omega_{kl}^{(n)} = \frac{1}{d} \mathrm{Tr}[\Delta S_k^{(n)} (S_l^{(n)})^T + S_k^{(n)} (\Delta S_l^{(n)})^T] + \frac{1}{d} \mathrm{Tr}[\Delta S_k^{(n)} (\Delta S_l^{(n)})^T]. \tag{428}$$

Substituting (422) gives, for example,

$$\frac{1}{d} \mathrm{Tr}[H_k^{(n)}] = 2 \sum_m \left(\Gamma_{\Omega,km}^{(d)} + \Gamma_{\Psi,km}^{(d)} \right) t_m^{(n)}, \tag{429}$$

$$\frac{1}{d} \mathrm{Tr}[H_k^{(n)} S_l^{(n)}] = 2 \sum_m \left(\Gamma_{\Omega,km}^{(d)} \Psi_{ml}^{(n)} + \Gamma_{\Psi,km}^{(d)} \Omega_{ml}^{(n)} \right), \tag{430}$$

and

$$\frac{1}{d} \mathrm{Tr}[H_k^{(n)} (S_l^{(n)})^T] = 2 \sum_m \left(\Gamma_{\Omega,km}^{(d)} \Omega_{ml}^{(n)} + \Gamma_{\Psi,km}^{(d)} \Psi_{ml}^{(n)} \right). \tag{431}$$

Therefore the first-order terms coincide with $\tau_d F(\mathbf{Q}^{(n)})$, up to an error $\tau_d \varepsilon_{\mathrm{drift}}^{(n)}$ satisfying

$$\mathbf{1}_{\{n<J\}} \|\varepsilon_{\mathrm{drift}}^{(n)}\|_\infty \le C(C_*) \delta_d(C_*). \tag{432}$$

The first-order martingale terms are

$$\eta_{t,k}^{(n)} := -\frac{\chi_k}{d} \mathrm{Tr}[\xi_k^{(n)}], \tag{433}$$

$$\eta_{\Psi,kl}^{(n)} := -\frac{1}{d} \mathrm{Tr} \left[\chi_k \xi_k^{(n)} S_l^{(n)} + \chi_l S_k^{(n)} \xi_l^{(n)} \right], \tag{434}$$

$$\eta_{\Omega,kl}^{(n)} := -\frac{1}{d} \mathrm{Tr} \left[\chi_k \xi_k^{(n)} (S_l^{(n)})^T + \chi_l S_k^{(n)} (\xi_l^{(n)})^T \right]. \tag{435}$$

Collect them into $\eta^{(n)}$. The remaining terms in (427)–(428), are denoted by $R^{(n)}$. Hence

$$\mathbf{Q}^{(n+1)} - \mathbf{Q}^{(n)} = \tau_d F(\mathbf{Q}^{(n)}) + \alpha_d \eta^{(n)} + R^{(n)} + \tau_d \varepsilon_{\mathrm{drift}}^{(n)}. \tag{436}$$

Step 3: Vanishing of the martingale noise. As in Step 3 of the proof of Lemma 4, every coordinate of $\eta^{(n)}$ is a finite sum of linear functionals of $\xi_k^{(n)}$. The same moment estimate therefore gives

$$\mathbf{1}_{\{n<J\}} \mathbb{E} \left[|\eta_a^{(n)}|^2 \,\middle|\, \mathcal{F}_n \right] \le \frac{C}{d^2} \tag{437}$$

for every coordinate a of (t, Ψ, Ω).

Thus the stopped process

$$M_a^{(m)} := \sum_{n=0}^{m-1} \alpha_d \eta_a^{(n)} \mathbf{1}_{\{n<J\}} \tag{438}$$

is a square-integrable martingale with

$$\mathbb{E}[\langle M_a \rangle_{N_T}] \le N_T \alpha_d^2 \frac{C}{d^2} \le C_T \frac{\alpha_d}{d}. \tag{439}$$

By Doob's maximal inequality,

$$\max_a \sup_{m \le N_T} |M_a^{(m)}| \xrightarrow{\mathbb{P}} 0. \tag{440}$$

Here the maximum is over only finitely many coordinates, since K is fixed.

Step 4: Vanishing of the higher-order residuals. On $\{n < J\}$, the same moment estimates as in Lemma 4 give

$$\mathbb{E}\left[\|R^{(n)}\|_\infty \,\middle|\, \mathcal{F}_n\right] \le C\left(\alpha_d^2 + \alpha_d\tau_d + \tau_d^2 + \alpha_d d^{-3/2}\right). \tag{441}$$

The first three terms arise from the products of the drift and stochastic increments in (427)–(428), while the last term comes from the residual in (425).

Since $N_T = O\left(\frac{d}{\alpha_d}\right)$ and $\tau_d = \frac{\alpha_d}{d}$, we obtain

$$\mathbb{E}\left[\sum_{n=0}^{N_T-1} \mathbf{1}_{\{n<J\}}\|R^{(n)}\|_\infty\right] \le C_T\left(d\alpha_d + \alpha_d + \frac{\alpha_d}{d} + d^{-1/2}\right) \longrightarrow 0, \tag{442}$$

because $\alpha_d \le c_0(d\log d)^{-1}$. Hence, by Markov's inequality,

$$\sup_{m\le N_T}\left\|\sum_{n=0}^{m-1} R^{(n)}\right\|_\infty \xrightarrow{\mathbb{P}} 0. \tag{443}$$

Likewise, by (432),

$$\sup_{m\le N_T}\left\|\tau_d\sum_{n=0}^{m-1}\varepsilon_{\text{drift}}^{(n)}\right\|_\infty \le TC(C_*)\delta_d(C_*) \longrightarrow 0. \tag{444}$$

Step 5: Integral equation. For $m(t) := \left\lfloor \frac{t}{\tau_d}\right\rfloor$, summing (436) gives

$$\mathbf{Q}^{(d)}(t) = \mathbf{Q}^{(d)}(0) + \tau_d\sum_{n=0}^{m(t)-1} F(\mathbf{Q}^{(n)}) + E_{\text{noise}}^{(d)}(t) + E_{\text{rem}}^{(d)}(t) + E_{\text{drift}}^{(d)}(t), \tag{445}$$

where $E_{\text{noise}}^{(d)}(t) := \alpha_d\sum_{n=0}^{m(t)-1}\eta^{(n)}$, $E_{\text{rem}}^{(d)}(t) := \sum_{n=0}^{m(t)-1}R^{(n)}$ and $E_{\text{drift}}^{(d)}(t) := \tau_d\sum_{n=0}^{m(t)-1}\varepsilon_{\text{drift}}^{(n)}$.

Because $\mathbf{Q}^{(d)}$ is piecewise constant,

$$\int_0^t F(\mathbf{Q}^{(d)}(s))\,ds = \tau_d\sum_{n=0}^{m(t)-1} F(\mathbf{Q}^{(n)}) + (t - m(t)\tau_d)F(\mathbf{Q}^{(m(t))}). \tag{446}$$

On $\{J > N_T\}$, the spectral bound implies that $\mathbf{Q}^{(d)}(t)$ remains in a bounded subset, on which F is bounded. Therefore the discretization error $E_{\text{disc}}^{(d)}(t) := (t - m(t)\tau_d)F(\mathbf{Q}^{(m(t))})$ satisfies

$$\sup_{t\in[0,T]}\|E_{\text{disc}}^{(d)}(t)\|_\infty \le C\tau_d \longrightarrow 0. \tag{447}$$

Combining (440), (443), (444), and (447), we conclude that $\sup_{t\in[0,T]}\left\|E_{\text{total}}^{(d)}(t)\right\|_\infty \xrightarrow{\mathbb{P}} 0$, where

$$E_{\text{total}}^{(d)} = E_{\text{noise}}^{(d)} + E_{\text{rem}}^{(d)} + E_{\text{drift}}^{(d)} + E_{\text{disc}}^{(d)}. \tag{448}$$

This proves the lemma. □

E.1.3 Final proof of Theorem 4

Fix an arbitrary finite time horizon $T > 0$, and let C_* be the constant in Lemma 8. Define

$$\mathcal{E}_d(T) := \left\{\sup_{0\le n\le N_T}\max_{1\le k\le K}\|S_k^{(n)}\|_{\text{op}} \le C_*\right\}, \qquad N_T = \left\lfloor\frac{Td}{\alpha_d}\right\rfloor. \tag{449}$$

By Lemma 8,

$$\mathbb{P}(\mathcal{E}_d(T)) \longrightarrow 1. \tag{450}$$

On $\mathcal{E}_d(T)$, the empirical order parameters satisfy

$$|t_k^{(d)}(t)| \le C_*, \qquad |\Psi_{kl}^{(d)}(t)| \le C_*^2, \qquad |\Omega_{kl}^{(d)}(t)| \le C_*^2, \tag{451}$$

uniformly over $t \in [0,T]$.

We first address the fact that the vector field F is naturally defined only on the domain $\mathbb{R}^K \times \mathcal{D}$. For $(\Psi,\Omega) \in \mathbb{S}^K \times \mathbb{S}^K$, write

$$Q^+ := \frac{\Omega+\Psi}{2}, \qquad Q^- := \frac{\Omega-\Psi}{2}, \tag{452}$$

and define

$$\widehat{Q}^{\pm} := \Pi_{\mathbb{S}_+^K}(Q^{\pm}), \qquad \widehat{\Omega} := \widehat{Q}^+ + \widehat{Q}^-, \qquad \widehat{\Psi} := \widehat{Q}^+ - \widehat{Q}^-. \tag{453}$$

Using $\nabla_\Psi \Phi(\widehat{\Psi},\widehat{\Omega})$ and $\nabla_\Omega \Phi(\widehat{\Psi},\widehat{\Omega})$ in (364)–(366) defines an extension $\widetilde{F}$ of F to the finite-dimensional space $\mathcal{X} := \mathbb{R}^K \times \mathbb{S}^K \times \mathbb{S}^K$. By Assumption 7.1, Price's theorem, and the non-expansiveness of the PSD projection, $\widetilde{F}$ is locally Lipschitz on $\mathcal{X}$.

Choose a fixed radius R larger than the bound in (451). By composing $\widetilde{F}$ with the projection onto the closed ball $B_R \subset \mathcal{X}$, we obtain a globally Lipschitz vector field $\widehat{F}$ which agrees with F on all states satisfying (451). Let $L_R < \infty$ denote a global Lipschitz constant of $\widehat{F}$.

Consider the globally well-posed ODE

$$\frac{d}{dt}\bar{\mathbf{Q}}(t) = \widehat{F}(\bar{\mathbf{Q}}(t)), \qquad \bar{\mathbf{Q}}(0) = \left(\bar{t}(0), \bar{\Psi}(0), \bar{\Omega}(0)\right), \tag{454}$$

where $\bar{\mathbf{Q}} = (\bar{t}, \bar{\Psi}, \bar{\Omega})$.

By Assumption 7.3 and the fact that K is fixed,

$$\left\|\mathbf{Q}^{(d)}(0) - \bar{\mathbf{Q}}(0)\right\|_\infty \xrightarrow{\mathbb{P}} 0. \tag{455}$$

On the other hand, Lemma 9 gives

$$\mathbf{Q}^{(d)}(t) = \mathbf{Q}^{(d)}(0) + \int_0^t F(\mathbf{Q}^{(d)}(s))\, ds + E_{\text{total}}^{(d)}(t), \tag{456}$$

with

$$\sup_{t\in[0,T]} \left\|E_{\text{total}}^{(d)}(t)\right\|_\infty \xrightarrow{\mathbb{P}} 0. \tag{457}$$

On $\mathcal{E}_d(T)$, every empirical state satisfies (451). Hence

$$F(\mathbf{Q}^{(d)}(t)) = \widehat{F}(\mathbf{Q}^{(d)}(t)), \qquad t \in [0,T]. \tag{458}$$

Subtracting the integral form of (454) from (456), we obtain on $\mathcal{E}_d(T)$

$$\left\|\mathbf{Q}^{(d)}(t) - \bar{\mathbf{Q}}(t)\right\|_\infty \le \left\|\mathbf{Q}^{(d)}(0) - \bar{\mathbf{Q}}(0)\right\|_\infty + \sup_{s\le T}\left\|E_{\text{total}}^{(d)}(s)\right\|_\infty + L_R \int_0^t \left\|\mathbf{Q}^{(d)}(s) - \bar{\mathbf{Q}}(s)\right\|_\infty ds. \tag{459}$$

Therefore, by Grönwall's inequality,

$$\sup_{t\in[0,T]} \left\|\mathbf{Q}^{(d)}(t) - \bar{\mathbf{Q}}(t)\right\|_\infty \le e^{L_R T}\left(\left\|\mathbf{Q}^{(d)}(0) - \bar{\mathbf{Q}}(0)\right\|_\infty + \sup_{t\in[0,T]}\left\|E_{\text{total}}^{(d)}(t)\right\|_\infty\right) \tag{460}$$

on $\mathcal{E}_d(T)$. Combining (450), (455), and (457) yields

$$\sup_{t\in[0,T]} \left\|\mathbf{Q}^{(d)}(t) - \bar{\mathbf{Q}}(t)\right\|_\infty \xrightarrow{\mathbb{P}} 0. \tag{461}$$

It remains to show that $\bar{\mathbf{Q}}$ is a solution of the original ODE rather than merely a solution of its extension. For every finite d and every t, $\left(\Psi^{(d)}(t), \Omega^{(d)}(t)\right) \in \mathcal{D}$. Since $\mathcal{D}$ is closed, (461) implies

$\big(\bar{\Psi}(t), \bar{\Omega}(t)\big) \in \mathcal{D}$ for $t \in [0, T]$. Likewise, (451) and (461) imply that $\bar{\mathbf{Q}}(t)$ remains inside the region where $\widehat{F} = F$. Consequently,

$$\frac{d}{dt}\bar{\mathbf{Q}}(t) = F(\bar{\mathbf{Q}}(t)), \qquad t \in [0, T], \tag{462}$$

and hence $\bar{\mathbf{Q}}$ satisfies (364)–(366).

Finally, the locally Lipschitz extension $\widetilde{F}$ implies local uniqueness of the ODE. Thus the solution $\bar{\mathbf{Q}}$ is unique. This proves Theorem 4.

E.2 Proof of Corollary 7

Recall that S_1 is trainable and S_2 is fixed. For brevity, write

$$A_{kl}(t) := [\nabla_\Omega \Phi(\bar{\Psi}(t), \bar{\Omega}(t))]_{kl}, \qquad B_{kl}(t) := [\nabla_\Psi \Phi(\bar{\Psi}(t), \bar{\Omega}(t))]_{kl}. \tag{463}$$

Since S_2 is fixed, $\bar{t}_2, p_{\Psi,22}, p_{\Omega,22}, a_+, a_-$ are constant in time.

We first derive the limiting dynamics of the structural overlaps. From (364), with $\chi_1 = 1$ and $\chi_2 = 0$,

$$\frac{d}{dt}\bar{t}_1 = -2(A_{11} + B_{11})\bar{t}_1 - 2(A_{12} + B_{12})\bar{t}_2 - \gamma \bar{t}_1. \tag{464}$$

Similarly, (365) and (366) give

$$\frac{d}{dt}\bar{\Psi}_{12} = -2A_{11}\bar{\Psi}_{12} - 2B_{11}\bar{\Omega}_{12} - 2A_{12}\bar{\Psi}_{22} - 2B_{12}\bar{\Omega}_{22} - \gamma\bar{\Psi}_{12}, \tag{465}$$

$$\frac{d}{dt}\bar{\Omega}_{12} = -2A_{11}\bar{\Omega}_{12} - 2B_{11}\bar{\Psi}_{12} - 2A_{12}\bar{\Omega}_{22} - 2B_{12}\bar{\Psi}_{22} - \gamma\bar{\Omega}_{12}. \tag{466}$$

Recall

$$\bar{p}_{\Psi,12} = \bar{\Psi}_{12} - \bar{t}_1\bar{t}_2, \qquad \bar{p}_{\Omega,12} = \bar{\Omega}_{12} - \bar{t}_1\bar{t}_2. \tag{467}$$

Subtracting $\bar{t}_2$ times (464) from (465) and (466), respectively, we obtain

$$\frac{d}{dt}\bar{p}_{\Psi,12} = -2A_{11}\bar{p}_{\Psi,12} - 2B_{11}\bar{p}_{\Omega,12} - 2A_{12}p_{\Psi,22} - 2B_{12}p_{\Omega,22} - \gamma\bar{p}_{\Psi,12}, \tag{468}$$

$$\frac{d}{dt}\bar{p}_{\Omega,12} = -2A_{11}\bar{p}_{\Omega,12} - 2B_{11}\bar{p}_{\Psi,12} - 2A_{12}p_{\Omega,22} - 2B_{12}p_{\Psi,22} - \gamma\bar{p}_{\Omega,12}. \tag{469}$$

It is convenient to use

$$\bar{p}_+ := \frac{\bar{p}_{\Omega,12} + \bar{p}_{\Psi,12}}{2}, \qquad \bar{p}_- := \frac{\bar{p}_{\Omega,12} - \bar{p}_{\Psi,12}}{2}. \tag{470}$$

Define

$$h_+(t) := A_{11}(t) + B_{11}(t), \qquad h_-(t) := A_{11}(t) - B_{11}(t), \tag{471}$$

and recall

$$g_+(t) := A_{12}(t) + B_{12}(t), \qquad g_-(t) := A_{12}(t) - B_{12}(t). \tag{472}$$

Adding and subtracting (468)–(469) yields the decoupled equations

$$\frac{d}{dt}\bar{p}_+ = -\left(2h_+(t) + \gamma\right)\bar{p}_+ - 2a_+ g_+(t), \tag{473}$$

$$\frac{d}{dt}\bar{p}_- = -\left(2h_-(t) + \gamma\right)\bar{p}_- - 2a_- g_-(t). \tag{474}$$

By Assumption 8, $\bar{p}_+(0) = \bar{p}_-(0) = 0$. Consequently,

$$\dot{\bar{p}}_+(0) = -2a_+ g_+(0), \qquad \dot{\bar{p}}_-(0) = -2a_- g_-(0). \tag{475}$$

We now prove the two claims separately.

Case 1. Suppose (376) holds. Then there exists $\sigma \in \{+, -\}$ such that

$$v_\sigma := \dot{\bar{p}}_\sigma(0) \neq 0. \tag{476}$$

By Theorem 4, the limiting trajectory is continuously differentiable, and hence $\dot{\bar{p}}_\sigma(t)$ is continuous. Therefore there exists $T_0 > 0$, independent of d, such that

$$|\dot{\bar{p}}_\sigma(t) - v_\sigma| \leq \frac{|v_\sigma|}{2}, \qquad 0 \leq t \leq T_0. \tag{477}$$

In particular, $\dot{\bar{p}}_\sigma(t)$ has the same sign as v_σ throughout $[0, T_0]$ and

$$|\dot{\bar{p}}_\sigma(t)| \geq \frac{|v_\sigma|}{2}. \tag{478}$$

Since $\bar{p}_\sigma(0) = 0$,

$$|\bar{p}_\sigma(T_0)| \geq \frac{|v_\sigma| T_0}{2}. \tag{479}$$

Moreover,

$$\max\left(|\bar{p}_{\Psi,12}|, |\bar{p}_{\Omega,12}|\right) \geq \max(|\bar{p}_+|, |\bar{p}_-|). \tag{480}$$

Set $c := \frac{|v_\sigma| T_0}{4}$ and $T_+ := T_0$. Then

$$\max\left(|\bar{p}_{\Psi,12}(T_+)|, |\bar{p}_{\Omega,12}(T_+)|\right) \geq 2c. \tag{481}$$

On the other hand, because both structural overlaps vanish at initialization and are continuous, there exists $T_- \in (0, T_+)$ such that

$$\sup_{0 \leq t \leq T_-} \max\left(|\bar{p}_{\Psi,12}(t)|, |\bar{p}_{\Omega,12}(t)|\right) < \frac{c}{2}. \tag{482}$$

Let $\tilde{p}^{(d)}_{\Psi,12}(t)$ and $\tilde{p}^{(d)}_{\Omega,12}(t)$ denote the corresponding empirical piecewise-constant interpolations. Since

$$\tilde{p}^{(d)}_{\Psi,12} = \tilde{\Psi}^{(d)}_{12} - \tilde{t}^{(d)}_1 \tilde{t}^{(d)}_2, \qquad \tilde{p}^{(d)}_{\Omega,12} = \tilde{\Omega}^{(d)}_{12} - \tilde{t}^{(d)}_1 \tilde{t}^{(d)}_2, \tag{483}$$

Theorem 4 implies

$$\sup_{t \in [0,T]} \max\left(\left|\tilde{p}^{(d)}_{\Psi,12}(t) - \bar{p}_{\Psi,12}(t)\right|, \left|\tilde{p}^{(d)}_{\Omega,12}(t) - \bar{p}_{\Omega,12}(t)\right|\right) \xrightarrow{\mathbb{P}} 0 \tag{484}$$

for every fixed $T < \infty$.

Applying (484) on $[0, T_+]$ and using (481), we obtain

$$\mathbb{P}\left(\max\left(|\tilde{p}^{(d)}_{\Psi,12}(T_+)|, |\tilde{p}^{(d)}_{\Omega,12}(T_+)|\right) \geq c\right) \longrightarrow 1. \tag{485}$$

Hence, with probability tending to one,

$$N^{(d)}_{\mathrm{wr},S}(c) \leq \left\lceil \frac{T_+}{\tau_d} \right\rceil = \left\lceil \frac{T_+ d}{\alpha_d} \right\rceil. \tag{486}$$

Likewise, (482) and (484) imply

$$\mathbb{P}\left(\sup_{0 \leq t \leq T_-} \max\left(|\tilde{p}^{(d)}_{\Psi,12}(t)|, |\tilde{p}^{(d)}_{\Omega,12}(t)|\right) < c\right) \longrightarrow 1. \tag{487}$$

Therefore, with probability tending to one,

$$N^{(d)}_{\mathrm{wr},S}(c) > \left\lfloor \frac{T_- d}{\alpha_d} \right\rfloor. \tag{488}$$

Combining (486) and (488) proves (377). In particular, $N^{(d)}_{\mathrm{wr},S}(c) = \Theta_{\mathbb{P}}\left(\frac{d}{\alpha_d}\right)$. Choosing $\alpha_d = \Theta\left(\frac{1}{d \log d}\right)$ gives

$$N^{(d)}_{\mathrm{wr},S}(c) = \Theta_{\mathbb{P}}(d^2 \log d). \tag{489}$$

Case 2. Suppose now that (379) holds. Consider the uninformative manifold

$$\mathcal{U} = \{(t, \Psi, \Omega) : t_1 = 0, \quad \Psi_{12} = 0, \quad \Omega_{12} = 0\}. \tag{490}$$

On $\mathcal{U}$, we have

$$p_{\Psi,12} = p_{\Omega,12} = p_+ = p_- = 0. \tag{491}$$

Moreover, by (379), $A_{12} = B_{12} = 0$. Therefore (464) gives $\dot{\bar{t}}_1 = 0$ while (473)–(474) give $\dot{\bar{p}}_+ = \dot{\bar{p}}_- = 0$. Thus the limiting vector field is tangent to $\mathcal{U}$.

By uniqueness of the solution established in Theorem 4, and since Assumption 8 places the initial state on $\mathcal{U}$, the limiting trajectory remains on this manifold:

$$\bar{t}_1(t) = 0, \qquad \bar{\Psi}_{12}(t) = \bar{\Omega}_{12}(t) = 0, \qquad t \geq 0. \tag{492}$$

Equivalently,

$$\bar{p}_{\Psi,12}(t) = \bar{p}_{\Omega,12}(t) = 0, \qquad t \geq 0. \tag{493}$$

Combining (493) with (484), for every fixed $T < \infty$ and every $\epsilon > 0$,

$$\mathbb{P}\left(\sup_{t\in[0,T]} \max\left(|\tilde{p}^{(d)}_{\Psi,12}(t)|, |\tilde{p}^{(d)}_{\Omega,12}(t)| \right) > \epsilon \right) \longrightarrow 0, \tag{494}$$

which proves (380).

Finally, let $C > 0$ be arbitrary and consider $M_d := \lfloor Cd^2 \log d \rfloor$ SGD steps, satisfying

$$M_d \tau_d = M_d \frac{\alpha_d}{d} \leq Cd \log d\, \alpha_d \leq Cc_0 \tag{495}$$

under the learning-rate condition of Theorem 4. Hence, for every fixed recovery threshold $c > 0$,

$$\mathbb{P}\left(N^{(d)}_{\mathrm{wr},S}(c) \leq Cd^2 \log d \right) \longrightarrow 0. \tag{496}$$

Since $C > 0$ is arbitrary, S-SGD cannot achieve weak recovery within $\mathcal{O}(d^2 \log d)$ samples.

This completes the proof.

F Theory in Section 4

F.1 Assumptions for Theorem 3

In this section we formally state the assumptions for Theorem 3.

Assumption 9. *The following conditions hold.*

1. *The loss $\mathcal{L} : \mathbb{R}^{L\times L\times K} \to \mathbb{R}$ is six times continuously differentiable, and $\mathcal{L}$ together with its derivatives up to sixth order has at most polynomial growth.*
2. *For every k, $\|U_k^{(0)}\|_F + \|V_k^{(0)}\|_F = \Theta(\sqrt{d})$. There exists a constant $C_0 > 0$, independent of d, such that $\max_k \left\{ \|U_k^{(0)}\|_{\mathrm{op}}, \|V_k^{(0)}\|_{\mathrm{op}} \right\} \leq C_0$ almost surely.*
3. *There exist deterministic limits $\tilde{m}_k(0)$ and $\bar{\mu}_\alpha(0)$ such that $m_k^{(d)}(0) \to \tilde{m}_k(0)$ and $\mu_\alpha^{(d)}(0) \to \bar{\mu}_\alpha(0)$ in probability for every k and every α.*
4. *For every trainable index $k \in \mathcal{K}$, $\bar{t}_k^A(0) > 0$, $\bar{t}_k^B(0) > 0$ and $\left|\bar{t}_k^A(0) - \bar{t}_k^B(0)\right| \geq c_{\mathrm{bal}}$ for some $c_{\mathrm{bal}} > 0$.*
5. *The regularization is sufficiently large, i.e., $\gamma > \gamma_*(C_0, \mathcal{C}, \mathcal{L}, K, L, T)$ for some γ_* independent of d.*

Assumption 9.4 is a technical non-degeneracy condition used to control the denominator $t_k^A + t_k^B$ of (17) and obtain uniform bounds along the slow trajectory. It is satisfied, for example, by independent Gaussian initializations with slightly different variances for U_k and V_k. Assumption 9.5 plays the same spectral-control role as in the tied case.

Assumption 10 (Local strong convexity)**.** *Fix a time horizon $T > 0$. Let $m_{\mathcal{K}}$ denote the trainable mean components, while $m_{\setminus\mathcal{K}}$ is fixed at $m_{\setminus\mathcal{K}}(0)$. Assume that there exist an open convex set $\mathcal{M} \subset \mathbb{R}^{|\mathcal{K}|}$, an open set $\mathcal{Q}$ of covariance parameters $q = (q^{(1)}, q^{(2)})$, and a constant $\lambda_0 > 0$ such that:*

1. *For every* $(m_{\mathcal{K}}, q) \in \mathcal{M} \times \mathcal{Q}$, $\nabla^2_{m_{\mathcal{K}}} \Phi(m, q) \succeq \lambda_0 I$.
2. *Writing* $q_0 = (\bar{q}^{(1)}(0), \bar{q}^{(2)}(0))$, *we have* $\tilde{m}_{\mathcal{K}}(0) \in \mathcal{M}$, $q_0 \in \mathcal{Q}$, *and*
$$\left\{m_{\mathcal{K}} \in \mathcal{M} : \Phi(m_{\mathcal{K}}, m_{\setminus\mathcal{K}}(0), q_0) \leq \Phi(\tilde{m}(0), q_0)\right\} \Subset \mathcal{M}. \tag{497}$$
3. *For every* $q \in \mathcal{Q}$, *the equation* $\nabla_{m_{\mathcal{K}}} \Phi(m, q) = 0$ *has a solution* $m^\star_{\mathcal{K}}(q) \in \mathcal{M}$. *The solution* $\big(m^\star_{\mathcal{K}}(\bar{q}(\tau)), \bar{q}(\tau)\big)$ *of* (16)-(19) *remains in a compact subset of* $\mathcal{M} \times \mathcal{Q}$ *for* $0 \leq \tau \leq T$[6] *and sufficiently large* γ.

Assumption 10 ensures that, for each nearby covariance state, the fast dynamics has a locally unique attracting critical point $m^\star(q)$. This property allows the fast mean variables to be eliminated on the slow timescale.

F.2 Proof of Theorem 3

F.2.1 A stronger version of Theorem 1

We first estimate the error scaling in Theorem 1. It is closely related to existing work on the multivariate Gaussian approximation [18, 19], and we present its independent proof in Appendix F.6 for completeness.

Lemma 10. *Fix constants* $C_*, R_* < \infty$. *Consider deterministic matrices* $M_k \in \mathbb{R}^{d\times d}$ *satisfying* $\max_k \|M_k\|_{\mathrm{op}} \leq C_*$ *and* $\max_k \frac{1}{\sqrt{d}}|\mathrm{Tr}(M_k)| \leq R_*$. *Let* $G^{(d)}_{ijk} = \frac{1}{\sqrt{d}} x_i^T M_k x_j$, *where* $x \sim \mathcal{N}(0, \mathcal{C} \otimes I_d)$, *and let* $G^{(d)}_{\mathrm{G}}$ *be a Gaussian tensor having the same mean and covariance as* $G^{(d)}$.

Then for every $f \in C^3(\mathbb{R}^{L\times L\times K})$ *whose derivatives up to third order have at most polynomial growth, there exists a constant* $C = C(C_*, R_*, \mathcal{C}, K, L, f)$ *independent of* d *and of the matrices* M_k, *such that*
$$\left|\mathbb{E} f(G^{(d)}) - \mathbb{E} f(G^{(d)}_{\mathrm{G}})\right| \leq \frac{C}{\sqrt{d}}. \tag{498}$$

F.2.2 Local well-posedness

We first introduce an extension of the vector field in (19) outside the set of covariance parameters. Define
$$\mathfrak{Q} := \left\{(q^{(1)}, q^{(2)}) \in \mathbb{S}^K \times \mathbb{S}^K : q^{(1)} + q^{(2)} \succeq 0, \quad q^{(1)} - q^{(2)} \succeq 0\right\}. \tag{499}$$
Let $\mathscr{P} : \mathbb{R}^{K\times K} \times \mathbb{R}^{K\times K} \longrightarrow \mathfrak{Q}$ denote the orthogonal projection onto the closed convex cone $\mathfrak{Q}$. In particular,
$$\|\mathscr{P}(q) - \mathscr{P}(q')\|_F \leq \|q - q'\|_F. \tag{500}$$
We next define the extended ODE for (19). Let $q(\mu) := \big(q^{(1)}(\mu), q^{(2)}(\mu)\big)$ and $\widehat{q}(\mu) := \mathscr{P}(q(\mu))$. We set
$$\widetilde{m}^\star_{\mathcal{K}}(\mu) := m^\star_{\mathcal{K}}(\widehat{q}(\mu)), \tag{501}$$
with the non-trainable components fixed at $m_{\setminus\mathcal{K}}(0)$, and define
$$a^{(r)}_{kl}(\mu) := \left[\nabla_{q^{(r)}} \Phi\left(\widetilde{m}^\star(\mu), \widehat{q}(\mu)\right)\right]_{kl}, \qquad r = 1, 2. \tag{502}$$
For $k \in \mathcal{K}$, let
$$\psi^{\mathrm{ext}}_k(\mu) := 2\sum_{l=1}^{K}\sum_{r=1}^{2} a^{(r)}_{kl}(\mu)\left[\mu_{(M_l^{(r)}, B_k)} + \mu_{(A_k, M_l^{(r)})}\right], \tag{503}$$
and
$$\beta^{\mathrm{ext}}_k(\mu) := -\frac{\psi^{\mathrm{ext}}_k(\mu)}{t^A_k(\mu) + t^B_k(\mu)}. \tag{504}$$
The extended vector field V_{ext} is defined by
$$[V_{\mathrm{ext}}(\mu)]_\alpha = -\sum_{\substack{i=1\\ k_i\in\mathcal{K}}}^{w} \beta^{\mathrm{ext}}_{k_i}(\mu)\mathcal{T}^{\mathrm{bias}}_i(\mu) - 2\sum_{\substack{i=1\\ k_i\in\mathcal{K}}}^{w}\sum_{l=1}^{K}\sum_{r=1}^{2} a^{(r)}_{k_i l}(\mu)\mathcal{T}^{\mathrm{diff}}_{i,l,r}(\mu) - 2\gamma|\alpha|_{\mathcal{K}}\mu_\alpha. \tag{505}$$

[6] More precisely, on every interval of local existence contained in $[0, T]$.

We refer to

$$\frac{d\mu}{d\tau} = V_{\mathrm{ext}}(\mu) \tag{506}$$

as an extension of the original ODE (19).

Whenever $q(\mu) \in \mathfrak{Q}$, the projection acts trivially, $\widehat{q}(\mu) = q(\mu)$. Hence, on the covariance set, $\widetilde{m}^\star_{\mathcal{K}}(\mu) = m^\star_{\mathcal{K}}(q(\mu))$, and (506) agrees with the original slow system (16)–(19).

For $s > 0$, define the open subset

$$\mathcal{U}_s := \left\{ \mu \in X_s : \mathscr{P}(q(\mu)) \in \mathcal{Q}, \quad \min_{k\in\mathcal{K}} \big(t_k^A(\mu) + t_k^B(\mu)\big) > 0 \right\}. \tag{507}$$

Lemma 11. *Let $s_0 > 0$ and $\mu_0 \in \mathcal{U}_{s_0}$. Under Assumptions 9 and 10, there exist $T_{\mathrm{local}} > 0$ and a strictly increasing continuous function $s : [0, T_{\mathrm{local}}) \to (s_0, \infty)$ with $s(0) = s_0$ such that* (506) *admits a unique solution $\mu(\tau) \in X_{s(\tau)}$ for $0 \le \tau < T_{\mathrm{local}}$ satisfying $\mu(\tau) \in \mathcal{U}_{s(\tau)}$.*

Proof. Fix $s_0 > 0$ and $\mu_0 \in \mathcal{U}_{s_0}$, and write

$$q_0 := \mathscr{P}(q(\mu_0)), \qquad d_0 := \min_{k\in\mathcal{K}} \big(t_k^A(\mu_0) + t_k^B(\mu_0)\big) > 0. \tag{508}$$

Since $q_0 \in \mathcal{Q}$, we may choose a neighborhood $\mathcal{V}_q$ of q_0 in $\mathfrak{Q}$ whose closure is contained in $\mathcal{Q}$.

We first prove the local regularity of the slaved mean $m^\star_{\mathcal{K}}(q)$. By Assumption 10, for every $q \in \mathcal{V}_q$ the map $m_{\mathcal{K}} \mapsto \Phi(m, q)$ is uniformly strongly convex on $\mathcal{M}$ and admits a critical point $m^\star_{\mathcal{K}}(q) \in \mathcal{M}$, which is therefore unique. Moreover, on compact subsets of $\mathcal{M} \times \mathcal{V}_q$ there exists $L_q < \infty$ such that

$$\|\nabla_{m_{\mathcal{K}}} \Phi(m, q) - \nabla_{m_{\mathcal{K}}} \Phi(m, q')\| \le L_q \|q - q'\|_F. \tag{509}$$

Uniform strong convexity gives

$$\langle \nabla_{m_{\mathcal{K}}} \Phi(m, q) - \nabla_{m_{\mathcal{K}}} \Phi(m', q), m - m' \rangle \ge \lambda_0 \|m - m'\|^2. \tag{510}$$

Applying (509) and (510) to $m = m^\star(q)$ and $m' = m^\star(q')$ yields

$$\|m^\star_{\mathcal{K}}(q) - m^\star_{\mathcal{K}}(q')\| \le \frac{L_q}{\lambda_0} \|q - q'\|_F. \tag{511}$$

Hence $m^\star_{\mathcal{K}}$ is locally Lipschitz on $\mathcal{V}_q$.

Since $\mathscr{P}$ is non-expansive and the finite-order coordinates of μ are continuous on X_s, there exists a neighborhood $\mathcal{V}_\mu$ of μ_0 in X_{s_0} such that for every $\mu \in \mathcal{V}_\mu$,

$$\mathscr{P}(q(\mu)) \in \mathcal{V}_q, \qquad t_k^A(\mu) + t_k^B(\mu) \ge \frac{d_0}{2}, \quad k \in \mathcal{K}. \tag{512}$$

It follows from (511), Assumption 9.1, and the definition that $a_{kl}^{(r)}(\mu)$ is locally bounded and locally Lipschitz on $\mathcal{V}_\mu$. The same is true for $\psi_k^{\mathrm{ext}}(\mu)$, and by the lower bound (512), for $\beta_k^{\mathrm{ext}}(\mu)$.

Then following Lemma 1, we can verify that V_{ext} is locally bounded and locally Lipschitz from X_s to $X_{s'}$. The Ovsyannikov theorem therefore yields $T_{\mathrm{local}} > 0$ and a strictly increasing continuous scale $s(\tau)$ for which (506) has a unique solution $\mu(\tau) \in X_{s(\tau)}$ on $[0, T_{\mathrm{local}})$. □

F.2.3 Uniform spectral bounds

Now we prove that the SGD trajectory does not blow up, i.e. the spectral norm of the weights remains uniformly bounded across the training steps.

Step 1: A lower bound for the traces.

Lemma 12. *Fix $T > 0$, and let $N_T := \left\lfloor \frac{Td}{\alpha_d} \right\rfloor$. Let J be any stopping time such that, on $\{n < J\}$,*

$$\max_{1\le k\le K} \left\{ \|U_k^{(n)}\|_{\mathrm{op}}, \|V_k^{(n)}\|_{\mathrm{op}} \right\} \le C_*, \qquad \|m^{(n)}\|_\infty \le R_* \tag{513}$$

for some constants $C_*, R_* < \infty$ independent of d, T, γ. Under the conditions of Theorem 3, there exists $c_D > 0$, independent of d, such that

$$\lim_{d\to\infty} \mathbb{P}\left(\inf_{\substack{0\le n< N_T\wedge J \\ k\in\mathcal{K}}} \left(t_k^{A,(n)} + t_k^{B,(n)} \right) \ge c_D \right) = 1. \tag{514}$$

Proof. Fix $k \in \mathcal{K}$, and suppress the index k when there is no ambiguity. Let

$$g_U^{(n)} := \nabla_U \mathcal{L}(G^{(n)}), \qquad g_V^{(n)} := \nabla_V \mathcal{L}(G^{(n)}) \tag{515}$$

denote the stochastic gradients. The SGD updates can be written as

$$U^{(n+1)} = a_d U^{(n)} - \alpha_d g_U^{(n)}, \qquad V^{(n+1)} = a_d V^{(n)} - \alpha_d g_V^{(n)}, \tag{516}$$

where $a_d := 1 - \frac{\alpha_d \gamma}{d}$. The key observation is that

$$\left\langle g_U^{(n)}, U^{(n)} \right\rangle_F = \sum_{i,j=1}^{L} \partial_{ijk} \mathcal{L}(G^{(n)}) G_{ijk}^{(n)} = \left\langle g_V^{(n)}, V^{(n)} \right\rangle_F. \tag{517}$$

Therefore, defining $\delta_k^{(n)} := t_k^{A,(n)} - t_k^{B,(n)}$, (516) gives the recursion

$$\delta^{(n+1)} = a_d^2 \delta^{(n)} + r^{(n)}, \qquad r^{(n)} := \frac{\alpha_d^2}{d} \left(\|g_U^{(n)}\|_F^2 - \|g_V^{(n)}\|_F^2 \right). \tag{518}$$

On $\{n < J\}$, the bounds in (513) imply uniform bounds on all moments of $G^{(n)}$. By Assumption 9.1, the same holds for the polynomially growing first derivatives of $\mathcal{L}$. Since $g_U^{(n)} = \frac{1}{\sqrt{d}} \sum_{i,j=1}^{L} \partial_{ijk} \mathcal{L}(G^{(n)}) x_i^{(n)} (x_j^{(n)})^T V^{(n)}$, and similarly for $g_V^{(n)}$, the estimates in Lemma 2 give

$$\mathbf{1}_{\{n<J\}} \mathbb{E} \left[\|g_U^{(n)}\|_F^2 + \|g_V^{(n)}\|_F^2 \,\middle|\, \mathcal{F}_n \right] \le Cd, \tag{519}$$

where C depends on $C_*, R_*, \mathcal{C}, \mathcal{L}, K, L$, but not on d or n.

Unrolling (518), for every $m < N_T \wedge J$,

$$\delta^{(m)} - a_d^{2m} \delta^{(0)} = \sum_{n=0}^{m-1} a_d^{2(m-1-n)} r^{(n)}. \tag{520}$$

Hence

$$\sup_{m < N_T \wedge J} \left| \delta^{(m)} - a_d^{2m} \delta^{(0)} \right| \le \sum_{n=0}^{N_T - 1} \mathbf{1}_{\{n<J\}} |r^{(n)}|. \tag{521}$$

Using (519),

$$\mathbb{E} \left[\sum_{n=0}^{N_T-1} \mathbf{1}_{\{n<J\}} |r^{(n)}| \right] \le \frac{\alpha_d^2}{d} \sum_{n=0}^{N_T-1} Cd \le C_T d \alpha_d \longrightarrow 0, \tag{522}$$

because $\alpha_d \le c_0 (d \log d)^{-1}$. Markov's inequality proves

$$\max_{k \in \mathcal{K}} \sup_{0 \le n < N_T \wedge J} \left| \delta_k^{(n)} - a_d^{2n} \delta_k^{(0)} \right| \xrightarrow{\mathbb{P}} 0. \tag{523}$$

Since $K = \Theta(1)$, the convergence holds simultaneously for all $k \in \mathcal{K}$.

Finally, Assumptions 9.3–9.4 imply

$$\mathbb{P} \left(\min_{k \in \mathcal{K}} |\delta_k^{(0)}| \ge \frac{c_{\mathrm{bal}}}{2} \right) \longrightarrow 1. \tag{524}$$

Moreover, uniformly for $n \le N_T$,

$$a_d^{2n} = \left(1 - \frac{\alpha_d \gamma}{d}\right)^{2n} \ge e^{-3\gamma T} \tag{525}$$

for all sufficiently large d. Combining this with (523), with probability tending to one,

$$\inf_{\substack{0 \le n < N_T \wedge J \\ k \in \mathcal{K}}} |\delta_k^{(n)}| \ge \frac{c_{\mathrm{bal}}}{4} e^{-3\gamma T}. \tag{526}$$

Since

$$t_k^{A,(n)} + t_k^{B,(n)} \ge |t_k^{A,(n)} - t_k^{B,(n)}| = |\delta_k^{(n)}|, \tag{527}$$

we obtain (514), for example with $c_D := \frac{c_{\mathrm{bal}}}{4} e^{-3\gamma T}$. □

Step 2: Finite-dimensional Stein expansion.

Lemma 13. *Fix a compact set $\mathfrak{D} \Subset \mathcal{M} \times \mathcal{Q}$ and a constant $C_* < \infty$. Let J be a stopping time such that, on $\{n < J\}$,*

$$\left(m_{\mathcal{K}}^{(n)}, q^{(1,n)}, q^{(2,n)}\right) \in \mathfrak{D}, \quad \max_{1\le l\le K} \left\{\|U_l^{(n)}\|_{\mathrm{op}}, \|V_l^{(n)}\|_{\mathrm{op}}\right\} \le C_*. \tag{528}$$

Write $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_n]$, and define $b_k^{(n)} := \sum_{i,j=1}^L \mathcal{C}_{ij}\, \mathbb{E}_n\left[\partial_{ijk}\mathcal{L}(G^{(n)})\right]$. For $r = 1, 2$, set $\mathcal{C}^{(1)}_{ij,i'j'} := \mathcal{C}_{ii'}\mathcal{C}_{jj'}$ and $\mathcal{C}^{(2)}_{ij,i'j'} := \mathcal{C}_{ij'}\mathcal{C}_{ji'}$ and define $a_{kl}^{(r,d,n)} := \frac{1}{2}\sum_{i,j,i',j'=1}^L \mathcal{C}^{(r)}_{ij,i'j'}\, \mathbb{E}_n\left[\partial^2_{(ijk),(i'j'l)}\mathcal{L}(G^{(n)})\right]$.

Then, uniformly for $n < J$ and $k \in \mathcal{K}$,

$$\bar{g}_{U,k}^{(n)} = \frac{b_k^{(n)}}{\sqrt{d}} V_k^{(n)} + \frac{2}{d}\sum_{l=1}^K \sum_{r=1}^2 a_{kl}^{(r,d,n)} M_l^{(r,n)} V_k^{(n)} + \mathcal{E}_{U,k}^{(n)}, \tag{529}$$

$$\bar{g}_{V,k}^{(n)} = \frac{b_k^{(n)}}{\sqrt{d}} U_k^{(n)} + \frac{2}{d}\sum_{l=1}^K \sum_{r=1}^2 a_{kl}^{(r,d,n)} M_l^{(\bar{r},n)} U_k^{(n)} + \mathcal{E}_{V,k}^{(n)}. \tag{530}$$

Moreover, there exists $C = C(C_*, \mathfrak{D}, \mathcal{C}, \mathcal{L}, K, L) < \infty$ such that $\max_{k,l,r} |a_{kl}^{(r,d,n)}| + \max_k |b_k^{(n)}| \le C$,
$\max_{k\in\mathcal{K}} \left\{\|\mathcal{E}_{U,k}^{(n)}\|_{\mathrm{op}}, \|\mathcal{E}_{V,k}^{(n)}\|_{\mathrm{op}}\right\} \le C d^{-3/2}$ and $\max_{1\le k\le K} \left|b_k^{(n)} - \frac{\partial\Phi}{\partial m_k}\left(m^{(n)}, q^{(1,n)}, q^{(2,n)}\right)\right| \le \frac{C}{\sqrt{d}}$.

Proof. Throughout the proof, we condition on $\mathcal{F}_n$ and suppress the superscript (n). Conditional on $\mathcal{F}_n$, all weights are deterministic and $x = (x_1, \ldots, x_L) \sim \mathcal{N}(0, \mathcal{C} \otimes I_d)$. Write $f_{ijk}(x) := \partial_{ijk}\mathcal{L}(G(x))$.

The stochastic gradients are

$$g_{U,k} = \frac{1}{\sqrt{d}} \sum_{i,j=1}^L f_{ijk}(x) x_i x_j^T V_k, \tag{531}$$

$$g_{V,k} = \frac{1}{\sqrt{d}} \sum_{i,j=1}^L f_{ijk}(x) x_j x_i^T U_k. \tag{532}$$

Similarly to the expansion used in Lemma 2, we can obtain

$$\bar{g}_{U,k} = \frac{1}{\sqrt{d}} \sum_{i,j=1}^L \mathcal{C}_{ij}\mathbb{E}[f_{ijk}] V_k + \frac{1}{d} \sum_{\substack{i,j,i',j'=1\\ l=1}}^{L,L,L,L,K} \mathbb{E}\left[\partial^2_{(ijk),(i'j'l)}\mathcal{L}(G)\right] \left(\mathcal{C}_{ii'}\mathcal{C}_{jj'} M_l + \mathcal{C}_{ij'}\mathcal{C}_{ji'} M_l^T\right) V_k + \mathcal{E}_{U,k}.$$

and similarly for $g_{V,k}$. This proves (529) and (530).

The remainders are controlled following Lemma 2, which gives $\|\mathcal{E}_{U,k}\|_{\mathrm{op}} + \|\mathcal{E}_{V,k}\|_{\mathrm{op}} \le C d^{-3/2}$.

We now identify the coefficients. Let G_{G} be the Gaussian tensor having the same mean and covariance as G. Then we have

$$\frac{\partial\Phi}{\partial m_k} = \sum_{i,j=1}^L \mathcal{C}_{ij}\mathbb{E}\left[\partial_{ijk}\mathcal{L}(G_{\mathrm{G}})\right]. \tag{533}$$

and

$$\left[\nabla_{q^{(1)}}\Phi\right]_{kl} = \frac{1}{2} \sum_{i,j,i',j'=1}^L \mathcal{C}_{ii'}\mathcal{C}_{jj'}\mathbb{E}\left[\partial^2_{(ijk),(i'j'l)}\mathcal{L}(G_{\mathrm{G}})\right], \tag{534}$$

$$\left[\nabla_{q^{(2)}}\Phi\right]_{kl} = \frac{1}{2} \sum_{i,j,i',j'=1}^L \mathcal{C}_{ij'}\mathcal{C}_{ji'}\mathbb{E}\left[\partial^2_{(ijk),(i'j'l)}\mathcal{L}(G_{\mathrm{G}})\right]. \tag{535}$$

by Price's theorem. Applying Lemma 10 finishes the proof. □

Step 3: One-step increment. For any process $Y^{(n)}$, we define $\Delta Y^{(n)} := Y^{(n+1)} - Y^{(n)}$.

Lemma 14. *Under the conditions of Theorem 3, let J be a stopping time satisfying* (528)*. Set $g^{(n)} := \nabla_{m_\mathcal{K}} \Phi\left(m^{(n)}, q^{(1,n)}, q^{(2,n)}\right)$ and $D^{(n)} := \operatorname{diag}\left(t_k^{A,(n)} + t_k^{B,(n)}\right)_{k\in\mathcal{K}}$, and $\vartheta^{(n)} := \left(t^{A,(n)}, t^{B,(n)}, q^{(1,n)}, q^{(2,n)}\right)$. Then, uniformly on $\{n < J\}$,*

$$\mathbb{E}_n\left[\Delta m_\mathcal{K}^{(n)}\right] = -\alpha_d D^{(n)}\left(g^{(n)} + \frac{e^{(n)}}{\sqrt{d}}\right) + r_m^{(n)}, \tag{536}$$

where $\|e^{(n)}\| \le C$ and $\|r_m^{(n)}\| \le C\left(\frac{(1+\gamma)\alpha_d}{d} + \alpha_d^2\right)$. Moreover,

$$\mathbb{E}_n\left[\left\|\Delta m_\mathcal{K}^{(n)} - \mathbb{E}_n[\Delta m_\mathcal{K}^{(n)}]\right\|^2\right] \le C\alpha_d^2, \tag{537}$$

and

$$\left\|\mathbb{E}_n[\Delta\vartheta^{(n)}]\right\| \le C\left(\frac{\alpha_d}{\sqrt{d}}\|g^{(n)}\| + \frac{(1+\gamma)\alpha_d}{d}\right), \tag{538}$$

$$\mathbb{E}_n\left[\|\Delta\vartheta^{(n)}\|^2\right] \le C(1+\gamma)^2\frac{\alpha_d^2}{d}. \tag{539}$$

Here $C = C(C_*, \mathfrak{D}, \mathcal{C}, \mathcal{L}, K, L) < \infty$ is independent of d, n, T, and γ.

Proof. Throughout the proof, we condition on $\mathcal{F}_n$ and suppress the superscript (n). Thus, for any state variable $Y^{(n)}$, we write $Y := Y^{(n)}$, $Y^+ := Y^{(n+1)}$ and $\Delta Y := Y^+ - Y$. We write $a_d := 1 - \frac{\alpha_d\gamma}{d}$. All constants below are uniform over the stopped class (528).

The spectral bounds (528) imply $\max_{1\le l\le K}\{\|M_l\|_{\mathrm{op}}, \|A_l\|_{\mathrm{op}}, \|B_l\|_{\mathrm{op}}\} \le C_*^2$. Together with the compactness assumption on $(m_\mathcal{K}, q)$, this gives uniform bounds on all fixed moments of G. Assumption 9.1 therefore gives uniform bounds on all fixed moments of the derivatives of $\mathcal{L}(G)$.

We shall repeatedly use the following elementary conditional Gaussian estimates. If P is $\mathcal{F}_n$-measurable and $\|P\|_{\mathrm{op}} \le C$, then, for every fixed polynomially growing function $F(G)$ arising below,

$$\mathbb{E}_n\left[\left|\frac{1}{d}F(G)x_i^T P x_j\right|^2\right] \le C, \qquad \mathbb{E}_n\left[\left|\frac{1}{d^{3/2}}F(G)x_i^T P x_j\right|^2\right] \le \frac{C}{d}, \tag{540}$$

and

$$\mathbb{E}_n\left|x_i^T M_l^{(r)} x_j\right|^p \le C_p d^{p/2}, \qquad r = 1, 2, \tag{541}$$

which follows from $\|M_l\|_{\mathrm{op}} = \mathcal{O}(1)$, $\|M_l\|_F = \mathcal{O}(\sqrt{d})$, and $\mathrm{Tr}M_l = \mathcal{O}(\sqrt{d})$.

For $k \in \mathcal{K}$, by definition we have

$$M_k^+ = a_d^2 M_k - \alpha_d a_d\left(g_{U,k}V_k^T + U_k g_{V,k}^T\right) + \alpha_d^2 g_{U,k}g_{V,k}^T. \tag{542}$$

Consequently,

$$\mathbb{E}_n[\Delta m_k] = (a_d^2 - 1)m_k - \frac{\alpha_d a_d}{\sqrt{d}}\left[\mathrm{Tr}(\bar{g}_{U,k}V_k^T) + \mathrm{Tr}(U_k\bar{g}_{V,k}^T)\right] + \frac{\alpha_d^2}{\sqrt{d}}\mathbb{E}_n\left[\mathrm{Tr}(g_{U,k}g_{V,k}^T)\right], \tag{543}$$

where $\bar{g}_{U,k} := \mathbb{E}_n[g_{U,k}]$, $\bar{g}_{V,k} := \mathbb{E}_n[g_{V,k}]$. Let b_k and $a_{kl}^{(r,d)}$ be the coefficients from Lemma 13, and define

$$\psi_k^{(d)} := 2\sum_{l=1}^K\sum_{r=1}^2 a_{kl}^{(r,d)}\left[\mu_{(M_l^{(r)}, B_k)} + \mu_{(A_k, M_l^{(r)})}\right]. \tag{544}$$

Lemma 13 gives

$$\frac{1}{\sqrt{d}}\left[\mathrm{Tr}(\bar{g}_{U,k}V_k^T) + \mathrm{Tr}(U_k\bar{g}_{V,k}^T)\right] = D_k b_k + \frac{\psi_k^{(d)}}{\sqrt{d}} + \rho_k, \tag{545}$$

where $D_k := t_k^A + t_k^B$ and $|\rho_k| \le \frac{C}{d}$. Set

$$\zeta_k := \sqrt{d}\left[b_k - \frac{\partial \Phi}{\partial m_k}\left(m, q^{(1)}, q^{(2)}\right)\right]. \tag{546}$$

Lemma 13 gives $|\zeta_k| \le C$. Moreover, since $A_k, B_k \succeq 0$,

$$\left|\mu_{(M_l^{(r)}, B_k)}\right| \le \|M_l\|_{\mathrm{op}}\, t_k^B, \quad \left|\mu_{(A_k, M_l^{(r)})}\right| \le \|M_l\|_{\mathrm{op}}\, t_k^A. \tag{547}$$

The coefficients $a_{kl}^{(r,d)}$ are uniformly bounded (Lemma 13), and hence $|\psi_k^{(d)}| \le C D_k$. Define

$$e_k := \zeta_k + \frac{\psi_k^{(d)}}{D_k}, \tag{548}$$

where the ratio is set equal to zero when $D_k = 0$. Thus we have $|e_k| \le C$. Finally, define $\chi_k := \frac{1}{\sqrt{d}}\mathbb{E}_n\left[\mathrm{Tr}(g_{U,k} g_{V,k}^T)\right]$. Expanding the two gradients and using (541) gives $|\chi_k| \le C$.

Substituting $b_k = \frac{\partial \Phi}{\partial m_k} + \frac{\zeta_k}{\sqrt{d}}$ and (545) into (543), we obtain

$$\mathbb{E}_n[\Delta m_k] = -\alpha_d D_k\left(\frac{\partial \Phi}{\partial m_k} + \frac{e_k}{\sqrt{d}}\right) + r_{m,k}, \tag{549}$$

where

$$r_{m,k} = (a_d^2 - 1) m_k + \alpha_d (1 - a_d) D_k\left(\frac{\partial \Phi}{\partial m_k} + \frac{e_k}{\sqrt{d}}\right) - \alpha_d a_d \rho_k + \alpha_d^2 \chi_k. \tag{550}$$

By $1 - a_d = \frac{\alpha_d \gamma}{d}$, $|a_d^2 - 1| \le \frac{2\alpha_d\gamma}{d} + \frac{\alpha_d^2\gamma^2}{d^2}$ and (528),

$$|r_{m,k}| \le C\left(\frac{(1+\gamma)\alpha_d}{d} + \alpha_d^2\right) \tag{551}$$

for all sufficiently large d. Summing over the finitely many trainable indices proves (536).

Now we define

$$L_k := \frac{1}{\sqrt{d}}\left[\mathrm{Tr}(g_{U,k} V_k^T) + \mathrm{Tr}(U_k g_{V,k}^T)\right], \qquad Q_k := \frac{1}{\sqrt{d}}\mathrm{Tr}(g_{U,k} g_{V,k}^T). \tag{552}$$

Thus

$$\Delta m_k = (a_d^2 - 1) m_k - \alpha_d a_d L_k + \alpha_d^2 Q_k. \tag{553}$$

By (540), $\mathbb{E}_n |L_k|^2 \le C$. The argument used for χ also gives $\mathbb{E}_n |Q_k|^2 \le C$. Therefore, using $a_d \le 1$ for all sufficiently large d,

$$\mathbb{E}_n\left[\left|\Delta m_k - \mathbb{E}_n[\Delta m_k]\right|^2\right] \le C\alpha_d^2 + C\alpha_d^4 \le C\alpha_d^2. \tag{554}$$

Since $|\mathcal{K}| = \mathcal{O}(1)$, this proves (537).

Recall that for a trainable index k,

$$M_k^+ = a_d^2 M_k - \alpha_d a_d\left(g_{U,k} V_k^T + U_k g_{V,k}^T\right) + \alpha_d^2 g_{U,k} g_{V,k}^T, \tag{555}$$

$$A_k^+ = a_d^2 A_k - \alpha_d a_d\left(g_{U,k} U_k^T + U_k g_{U,k}^T\right) + \alpha_d^2 g_{U,k} g_{U,k}^T, \tag{556}$$

$$B_k^+ = a_d^2 B_k - \alpha_d a_d\left(g_{V,k} V_k^T + V_k g_{V,k}^T\right) + \alpha_d^2 g_{V,k} g_{V,k}^T. \tag{557}$$

The preceding expansions imply, for every $Z_k \in \{M_k, M_k^T, A_k, B_k\}$,

$$\|\mathbb{E}_n[\Delta Z_k]\|_{\mathrm{op}} \le C\left(\frac{\alpha_d}{\sqrt{d}}\|g\| + \frac{(1+\gamma)\alpha_d}{d} + \alpha_d^2\right). \tag{558}$$

They also give

$$\mathbb{E}_n\left[\|\Delta Z_k\|_F^2\right] \le C(1+\gamma)^2 \alpha_d^2 d. \tag{559}$$

Using (558), (559), and $\frac{1}{d}\left|\mathrm{Tr}(PQ)\right| \le \frac{1}{d}\|P\|_F\|Q\|_F$, we obtain

$$\|\mathbb{E}_n[\Delta\vartheta]\| \le C\left(\frac{\alpha_d}{\sqrt{d}}\|g\| + \frac{(1+\gamma)\alpha_d}{d} + \alpha_d^2\right). \tag{560}$$

The learning-rate condition in Theorem 3 implies $\alpha_d^2 \le C\alpha_d/d$ for all sufficiently large d, and hence proves (538).

It remains to estimate the second moments. Expanding (555)–(557) shows that

$$\Delta\vartheta = \alpha_d\Lambda + R_\vartheta, \tag{561}$$

where each component of Λ is a finite sum of terms of the form $\frac{1}{d^{3/2}}\partial_{ijk}\mathcal{L}(G)\, x_i^T P x_j$ for an $\mathcal{F}_n$-measurable matrix P satisfying $\|P\|_{\mathrm{op}} \le C$. Thus (540) gives

$$\mathbb{E}_n\|\Lambda\|^2 \le \frac{C}{d}. \tag{562}$$

The remainder R_ϑ consists of the weight-decay terms and terms containing at least two weight increments. Similarly we obtain

$$\mathbb{E}_n\left[\|R_\vartheta\|^2\right] \le C(1+\gamma)^2\left(\frac{\alpha_d^2}{d^2} + \alpha_d^4\right). \tag{563}$$

Since $\alpha_d^2 \le C/d$ for all sufficiently large d, (561)–(563) give

$$\mathbb{E}_n\left[\|\Delta\vartheta\|^2\right] \le C(1+\gamma)^2\frac{\alpha_d^2}{d}. \tag{564}$$

This proves (539). □

Step 4: A bound for the mean variable.

Lemma 15. *Under the conditions of Theorem 3, let J be a stopping time satisfying* (528)*. Write $q^{(n)} := \left(q^{(1,n)}, q^{(2,n)}\right)$ and $N_T := \left\lfloor \frac{Td}{\alpha_d} \right\rfloor$. Then there exist constants $C_{\mathrm{loc}} = C_{\mathrm{loc}}(C_*, \mathfrak{D}, \mathcal{C}, \mathcal{L}, K, L) < \infty$ and $c_{\mathrm{loc}} = c_{\mathrm{loc}}(T, \gamma, c_{\mathrm{bal}}, \lambda_0) > 0$, independent of d, such that for any $\epsilon > 0$*

$$\lim_{d\to\infty} \mathbb{P}\left(\sup_{0\le n<N_T\wedge J}\left[\left\|m_{\mathcal{K}}^{(n)} - m_{\mathcal{K}}^\star(q^{(n)})\right\| - C_{\mathrm{loc}}e^{-c_{\mathrm{loc}}\alpha_d n}\right]_+ > \epsilon\right) = 0. \tag{565}$$

Proof. Define the excess population loss

$$\mathscr{E}(m_{\mathcal{K}}, q) := \Phi(m_{\mathcal{K}}, q) - \Phi(m_{\mathcal{K}}^\star(q), q). \tag{566}$$

Let $\mathfrak{D}_q \Subset \mathcal{Q}$ denote the projection of $\mathfrak{D}$ onto the q-coordinates. By the implicit function theorem and the positive definiteness of $\nabla^2_{m_{\mathcal{K}}}\Phi$ (Assumption 10), the map $q \longmapsto m_{\mathcal{K}}^\star(q)$ is continuously differentiable on a neighborhood of $\overline{\mathfrak{D}}_q$. The envelope identity gives

$$\nabla_{m_{\mathcal{K}}}\mathscr{E}(m_{\mathcal{K}}, q) = \nabla_{m_{\mathcal{K}}}\Phi(m, q), \tag{567}$$

$$\nabla_q\mathscr{E}(m_{\mathcal{K}}, q) = \nabla_q\Phi(m, q) - \nabla_q\Phi(m^\star(q), q). \tag{568}$$

Assumption 9.1 implies $\mathscr{E}$ has bounded second derivatives on a compact neighborhood of (528).

Strong convexity and compactness give constants $\Lambda < \infty$ and $C_{\mathscr{E}} < \infty$ such that, uniformly over (528),

$$\frac{\lambda_0}{2}\left\|m_{\mathcal{K}} - m_{\mathcal{K}}^\star(q)\right\|^2 \le \mathscr{E}(m_{\mathcal{K}}, q) \le \frac{\Lambda}{2}\left\|m_{\mathcal{K}} - m_{\mathcal{K}}^\star(q)\right\|^2, \tag{569}$$

$$\left\|\nabla_{m_{\mathcal{K}}}\Phi(m, q)\right\|^2 \ge 2\lambda_0\mathscr{E}(m_{\mathcal{K}}, q), \tag{570}$$

$$\left\|\nabla_q\mathscr{E}(m_{\mathcal{K}}, q)\right\| \le C_{\mathscr{E}}\left\|m_{\mathcal{K}} - m_{\mathcal{K}}^\star(q)\right\| \le \frac{C_{\mathscr{E}}}{\lambda_0}\left\|\nabla_{m_{\mathcal{K}}}\Phi(m, q)\right\|. \tag{571}$$

The last inequality follows from $\|\nabla_{m_{\mathcal{K}}}\Phi(m, q)\| \ge \lambda_0\left\|m_{\mathcal{K}} - m_{\mathcal{K}}^\star(q)\right\|$.

For $k \in \mathcal{K}$, let $D_k^{(n)} := t_k^{A,(n)} + t_k^{B,(n)}$ and $D^{(n)} := \operatorname{diag}\left(D_k^{(n)}\right)_{k\in\mathcal{K}}$. The spectral bound in (528) implies $0 \le D_k^{(n)} \le 2C_*^2$. By Lemma 12, there exists $c_D = c_D(T, \gamma, c_{\mathrm{bal}}) > 0$ such that the event

$$\mathcal{B}_d := \left\{ \inf_{\substack{0\le n<N_T\wedge J \\ k\in\mathcal{K}}} \left(t_k^{A,(n)} + t_k^{B,(n)}\right) \ge c_D \right\} \tag{572}$$

satisfies $\mathbb{P}(\mathcal{B}_d) \longrightarrow 1$. We work on $\mathcal{B}_d$. Set $\mathscr{E}_n := \mathscr{E}\left(m_{\mathcal{K}}^{(n)}, q^{(n)}\right)$. For $n < J$, Taylor's expansion and the boundedness of the second derivatives of $\mathscr{E}$ (according to Assumption 9.1) give

$$\mathbb{E}_n[\mathscr{E}_{n+1} - \mathscr{E}_n] = \left\langle g^{(n)}, \mathbb{E}_n[\Delta m_{\mathcal{K}}^{(n)}] \right\rangle + \left\langle \nabla_q \mathscr{E}\left(m_{\mathcal{K}}^{(n)}, q^{(n)}\right), \mathbb{E}_n[\Delta q^{(n)}] \right\rangle + R_{\mathscr{E}}^{(n)}, \tag{573}$$

where

$$|R_{\mathscr{E}}^{(n)}| \le C\,\mathbb{E}_n\left[\|\Delta m_{\mathcal{K}}^{(n)}\|^2 + \|\Delta q^{(n)}\|^2\right] \le C(1+\gamma)^2\alpha_d^2. \tag{574}$$

Here and below $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_n]$. Using Lemma 14, $D_k^{(n)} \le 2C_*^2$, and (571), we obtain

$$\mathbb{E}_n[\mathscr{E}_{n+1} - \mathscr{E}_n] \le -\alpha_d (g^{(n)})^T D^{(n)} g^{(n)} + C\frac{\alpha_d}{\sqrt{d}}\left(\|g^{(n)}\| + \|g^{(n)}\|^2\right) + C(1+\gamma)\frac{\alpha_d}{d} + C(1+\gamma)^2\alpha_d^2. \tag{575}$$

All quantities appearing here are uniformly bounded over $\mathfrak{D}$. Since $D^{(n)} \succeq c_D I$ for $n < J$, the $d^{-1/2}$-terms to be absorbed into the leading term for all sufficiently large d. Moreover, $\alpha_d^2 \le \alpha_d/d$ under the learning-rate assumption. Consequently, there exist $c_1 = c_1(T, \gamma, c_{\mathrm{bal}}, \lambda_0) > 0$ and $C_1 = C_1(T, \gamma, C_*, \mathfrak{D}) < \infty$ such that

$$\mathbb{E}_n[\mathscr{E}_{n+1} - \mathscr{E}_n] \le -\frac{c_D\alpha_d}{2}\|g^{(n)}\|^2 + C_1\frac{\alpha_d}{d} \le -c_1\alpha_d\mathscr{E}_n + C_1\frac{\alpha_d}{d}, \tag{576}$$

for $n < J$, where the second inequality uses (570).

Define the stopped martingale differences

$$\xi^{(n+1)} := \mathbf{1}_{\{n<J\}}\left[\mathscr{E}_{n+1} - \mathscr{E}_n - \mathbb{E}_n[\mathscr{E}_{n+1} - \mathscr{E}_n]\right]. \tag{577}$$

Because $\mathscr{E}$ has bounded gradient, Lemma 14 implies

$$\mathbb{E}_n\left[|\xi^{(n+1)}|^2\right] \le C\alpha_d^2. \tag{578}$$

Thus, with $\rho_d := 1 - c_1\alpha_d$, (576) gives, for $n < J$,

$$\mathscr{E}_{n+1} \le \rho_d\mathscr{E}_n + C_1\frac{\alpha_d}{d} + \xi^{(n+1)}. \tag{579}$$

To control this recursion uniformly over $N_T = \Theta(d/\alpha_d)$ steps, divide the time interval into blocks of length $\ell_d := \left\lceil \frac{1}{\alpha_d} \right\rceil$ and define $B_d := \left\lceil \frac{N_T}{\ell_d} \right\rceil \le C_T d$. For a block starting at $s = b\ell_d$, set $S_{b,r} := \sum_{j=s}^{s+r-1} \xi^{(j+1)}$ for $0 \le r \le \ell_d$, with the convention that increments after J are zero. Doob's maximal inequality and (578) give

$$\mathbb{P}\left(\max_{0\le b<B_d}\max_{0\le r\le\ell_d}|S_{b,r}| > \varepsilon_d\right) \le \frac{CB_d\ell_d\alpha_d^2}{\varepsilon_d^2}, \tag{580}$$

where we choose $\varepsilon_d := (d\alpha_d)^{1/4}$. Since $B_d = \mathcal{O}(d)$, $\ell_d = \mathcal{O}(\alpha_d^{-1})$, and $d\alpha_d \to 0$,

$$\frac{B_d\ell_d\alpha_d^2}{\varepsilon_d^2} \le C_T(d\alpha_d)^{1/2} \longrightarrow 0. \tag{581}$$

On the complementary high-probability event, Abel summation gives, for any $s < m$ in the same block,

$$\left|\sum_{j=s}^{m-1}\rho_d^{m-1-j}\xi^{(j+1)}\right| \le 2\max_{1\le r\le m-s}|S_{b,r}| \le 2\varepsilon_d. \tag{582}$$

Moreover, $\rho_d^{\ell_d} \le e^{-c_1\alpha_d\ell_d} \le e^{-c_1} < 1$. Iterating (579) block by block, using (582) and $\frac{\alpha_d}{d}\sum_{r=0}^{\infty}\rho_d^r = \frac{1}{c_1 d}$, yields

$$\mathscr{E}_n \le C_0\rho_d^n + C\left(\varepsilon_d + \frac{1}{d}\right), \qquad 0 \le n < N_T \wedge J, \tag{583}$$

where $C_0 := \sup_{(m_{\mathcal{K}},q)\in\mathfrak{D}} \mathscr{E}(m_{\mathcal{K}}, q) < \infty$. Using (569) and $\rho_d^n \le e^{-c_1\alpha_d n}$, we conclude that, on an event whose probability tends to one,

$$\left\| m_{\mathcal{K}}^{(n)} - m_{\mathcal{K}}^{\star}(q^{(n)}) \right\| \le C_{\mathrm{loc}} e^{-\frac{c_1}{2}\alpha_d n} + C\left[(d\alpha_d)^{1/8} + d^{-1/2}\right] \tag{584}$$

uniformly for $0 \le n < N_T \wedge J$. The term in square brackets tends to zero, which proves (565), with $c_{\mathrm{loc}} = c_1/2$. Removing $\mathcal{B}_d$ is legitimate because $\mathbb{P}(\mathcal{B}_d^c) \to 0$. □

Step 5: A refined bound for the mean gradients. From Lemma 15, we can obtain the following bound for the mean gradients

$$\lim_{d\to\infty} \mathbb{P}\left(\sup_{0\le n<N_T\wedge J} \left[\|b_k^{(n)}\| - C_{\mathrm{loc}} e^{-c_{\mathrm{loc}}\alpha_d n} \right]_+ > \epsilon \right) = 0 \tag{585}$$

by the boundedness of the Hessian. However, we would require a refined bound given by the following lemma.

Lemma 16. *Under the conditions of Theorem 3, fix $T > 0$ and let J be a stopping time satisfying* (528). *Set $N_T := \left\lfloor \frac{Td}{\alpha_d} \right\rfloor$ and $b_{\mathcal{K}}^{(n)} := \left(b_k^{(n)}\right)_{k\in\mathcal{K}}$, where $b_k^{(n)}$ is defined in Lemma 13. Then there exist constants $C_{\mathrm{tr}} = C_{\mathrm{tr}}(T,\gamma,C_*,\mathfrak{D}) < \infty, c_{\mathrm{tr}} = c_{\mathrm{tr}}(T,\gamma,C_*,\mathfrak{D}) > 0, C_b = C_b(C_*,\mathfrak{D},\mathcal{C},\mathcal{L},K,L) < \infty, c_0^\star = c_0^\star(T,\gamma,C_*,\mathfrak{D}) > 0$ such that, whenever $c_0 \le c_0^\star$,*

$$\lim_{d\to\infty} \mathbb{P}\left(\sup_{0\le n<N_T\wedge J} \left[\|b_{\mathcal{K}}^{(n)}\| - C_{\mathrm{tr}} e^{-c_{\mathrm{tr}}\alpha_d n} \right]_+ > \frac{C_b}{\sqrt{d}} \right) = 0. \tag{586}$$

Proof. Write $q^{(n)} := \left(q^{(1,n)}, q^{(2,n)}\right)$ and $D^{(n)} := \operatorname{diag}\left(t_k^{A,(n)} + t_k^{B,(n)}\right)_{k\in\mathcal{K}}$. All constants below are uniform over (528). By Lemma 12, there exists $c_D = c_D(T,\gamma,c_{\mathrm{bal}}) > 0$ such that the event $\mathcal{B}_d := \left\{ \inf_{\substack{0\le n<N_T\wedge J \\ k\in\mathcal{K}}} \left(t_k^{A,(n)} + t_k^{B,(n)}\right) \ge c_D \right\}$ satisfies $\mathbb{P}(\mathcal{B}_d) \longrightarrow 1$. We work on $\mathcal{B}_d$.

Let $H_\star(q) := \nabla^2_{m_{\mathcal{K}}}\Phi\left(m^\star(q), q\right)$ and $S(q) := H_\star(q)^{1/2}$. By Assumption 10, the implicit function theorem, and Assumption 9.1, the maps $q \longmapsto m_{\mathcal{K}}^\star(q), q \longmapsto H_\star(q), q \longmapsto S(q)$ are C^2 on a neighborhood of the projection of $\mathfrak{D}$ onto the q-coordinates. In particular, on that neighborhood,

$$\lambda_0 I \preceq H_\star(q) \preceq \Lambda I, \qquad \|S(q)\|_{\mathrm{op}} + \|S(q)^{-1}\|_{\mathrm{op}} \le C, \tag{587}$$

and S, S^{-1}, and $m^\star$ are Lipschitz. Define

$$x^{(n)} := m_{\mathcal{K}}^{(n)} - m_{\mathcal{K}}^\star(q^{(n)}), \qquad y^{(n)} := S(q^{(n)})x^{(n)}. \tag{588}$$

Taylor expansion gives

$$g^{(n)} := \nabla_{m_{\mathcal{K}}}\Phi\left(m^{(n)}, q^{(n)}\right) = H_\star(q^{(n)})x^{(n)} + \rho_g^{(n)}, \qquad \|\rho_g^{(n)}\| \le C\|x^{(n)}\|^2. \tag{589}$$

Fix $r_0 > 0$, to be chosen sufficiently small below. For convenience, write

$$S_n := S(q^{(n)}), \qquad H_n := H_\star(q^{(n)}) = S_n^2, \qquad \mathcal{D}_n := D^{(n)}. \tag{590}$$

Define

$$A^{(n)} := S_n\mathcal{D}_n S_n, \qquad u^{(n)} := S_n^{-1}e^{(n)}, \tag{591}$$

where $e^{(n)}$ is the bounded vector appearing in Lemma 14. On the event $\mathcal{B}_d$, we have $c_D I \preceq \mathcal{D}_n \preceq 2C_*^2 I$. Together with (587), this yields

$$a_- I \preceq A^{(n)} \preceq a_+ I, \qquad a_- := \lambda_0 c_D > 0, \qquad a_+ := 2\Lambda C_*^2. \tag{592}$$

Moreover,

$$\sup_{n<J} \|u^{(n)}\| \le \sup_{q} \|S(q)^{-1}\|_{\mathrm{op}} \sup_{n<J} \|e^{(n)}\| =: C_u < \infty. \tag{593}$$

We next derive the recursion for $y^{(n)}$. Define

$$F(m_{\mathcal{K}}, q) := S(q)\left(m_{\mathcal{K}} - m^{\star}_{\mathcal{K}}(q)\right), \tag{594}$$

so that

$$y^{(n)} = F(m^{(n)}_{\mathcal{K}}, q^{(n)}). \tag{595}$$

By Lemma 14,

$$\mathbb{E}_n[\Delta m^{(n)}_{\mathcal{K}}] = -\alpha_d \mathcal{D}_n \left(g^{(n)} + \frac{e^{(n)}}{\sqrt{d}} \right) + r^{(n)}_m, \tag{596}$$

where $\|r^{(n)}_m\| \le C\left(\frac{(1+\gamma)\alpha_d}{d} + \alpha_d^2\right)$. Furthermore,

$$\left\| \mathbb{E}_n[\Delta q^{(n)}] \right\| \le C\left(\frac{\alpha_d}{\sqrt{d}} \|g^{(n)}\| + \frac{(1+\gamma)\alpha_d}{d} \right), \tag{597}$$

and

$$\mathbb{E}_n \|\Delta q^{(n)}\|^2 \le C(1+\gamma)^2 \frac{\alpha_d^2}{d}. \tag{598}$$

Also,

$$\mathbb{E}_n \|\Delta m^{(n)}_{\mathcal{K}}\|^2 \le C\alpha_d^2. \tag{599}$$

Since F is C^2 on a compact neighborhood before the stopping time, a second-order Taylor expansion gives

$$\mathbb{E}_n[y^{(n+1)} - y^{(n)}] = S_n \mathbb{E}_n[\Delta m^{(n)}_{\mathcal{K}}] + \partial_q F(m^{(n)}_{\mathcal{K}}, q^{(n)}) \left[\mathbb{E}_n[\Delta q^{(n)}] \right] + \rho^{(n)}_F, \tag{600}$$

where, uniformly before the stopping time,

$$\|\rho^{(n)}_F\| \le C\mathbb{E}_n \left[\|\Delta m^{(n)}_{\mathcal{K}}\|^2 + \|\Delta q^{(n)}\|^2 \right] \le C(1+\gamma)^2 \alpha_d^2. \tag{601}$$

Using (589),

$$g^{(n)} = H_n x^{(n)} + \rho^{(n)}_g, \qquad \|\rho^{(n)}_g\| \le C\|x^{(n)}\|^2, \tag{602}$$

and the identities $S_n \mathcal{D}_n H_n x^{(n)} = S_n \mathcal{D}_n S_n y^{(n)} = A^{(n)} y^{(n)}$, $S_n \mathcal{D}_n e^{(n)} = A^{(n)} u^{(n)}$, we obtain, whenever $n < J$ and $\|y^{(n)}\| \le r_0$,

$$\mathbb{E}_n[y^{(n+1)} - y^{(n)}] = -\alpha_d A^{(n)} y^{(n)} - \frac{\alpha_d}{\sqrt{d}} A^{(n)} u^{(n)} + r^{(n)}_y, \tag{603}$$

where

$$\|r^{(n)}_y\| \le C\left[\alpha_d \|y^{(n)}\|^2 + \frac{\alpha_d}{\sqrt{d}} \|y^{(n)}\| + \frac{(1+\gamma)\alpha_d}{d} + (1+\gamma)^2 \alpha_d^2 \right]. \tag{604}$$

Here we used (587) and the fact that $\partial_q F$ is uniformly bounded before the stopping time.

Define

$$\xi^{(n+1)} := y^{(n+1)} - \mathbb{E}_n[y^{(n+1)}]. \tag{605}$$

Then $\mathbb{E}_n[\xi^{(n+1)}] = 0$. Since F is Lipschitz before the stopping time, Lemma 14 and (599)– (598) imply

$$\mathbb{E}_n \|\xi^{(n+1)}\|^2 \le C\alpha_d^2. \tag{606}$$

Consequently,

$$y^{(n+1)} = \left(I - \alpha_d A^{(n)} \right) y^{(n)} - \frac{\alpha_d}{\sqrt{d}} A^{(n)} u^{(n)} + \xi^{(n+1)} + r^{(n)}_y. \tag{607}$$

We first use Lemma 15 to enter the region in which (607) can be iterated. Let $C_S := \sup_q \|S(q)\|_{\mathrm{op}} < \infty$, where the supremum is taken over a compact neighborhood of the q-projection of $\mathfrak{D}$. Choose

$t_0 = t_0(T, \gamma, r_0)$ sufficiently large that $C_S C_{\mathrm{loc}} e^{-c_{\mathrm{loc}} t_0} < \frac{r_0}{8}$ and set $n_0 := \left\lceil \frac{t_0}{\alpha_d} \right\rceil$. Applying Lemma 15 with $\epsilon = r_0/(8C_S)$, and using $\alpha_d n_0 \geq t_0$, gives

$$\mathbb{P}\left(\sup_{n_0 \leq n < N_T \wedge J} \|y^{(n)}\| > \frac{r_0}{4} \right) \longrightarrow 0. \tag{608}$$

Here and below, the supremum over an empty index set is understood to be zero.

Starting from the entry time n_0, define the local exit time

$$J_{\mathrm{loc}} := \inf\left\{ n \geq n_0 : \|y^{(n)}\| > r_0 \right\}. \tag{609}$$

All local estimates below are carried out up to $N_T \wedge J \wedge J_{\mathrm{loc}}$. We divide the time interval after n_0 into blocks. Set

$$\ell_d := \left\lceil \frac{1}{a_- \alpha_d} \right\rceil, \qquad s_b := n_0 + b\ell_d, \tag{610}$$

and, whenever $s_b < N_T \wedge J \wedge J_{\mathrm{loc}}$, let $I_b := \{s_b, \dots, \min\left(s_b + \ell_d - 1,\ N_T \wedge J \wedge J_{\mathrm{loc}} - 1\right)\}$ to be the $b-$th block.

We next analyze the dynamics on each block. For a block I_b, set

$$\bar{A}_b := A^{(s_b)}, \qquad P_b := I - \alpha_d \bar{A}_b. \tag{611}$$

By (592), for all sufficiently large d,

$$0 \preceq P_b \preceq (1 - a_- \alpha_d) I, \qquad \|P_b^j\|_{\mathrm{op}} \leq 1, \quad j \geq 0. \tag{612}$$

For $n \in I_b$, rewrite (607) as

$$y^{(n+1)} = P_b y^{(n)} - \frac{\alpha_d}{\sqrt{d}} \bar{A}_b u^{(n)} + \xi^{(n+1)} + \rho_b^{(n)}, \tag{613}$$

where

$$\rho_b^{(n)} := -\alpha_d \big(A^{(n)} - \bar{A}_b\big) y^{(n)} - \frac{\alpha_d}{\sqrt{d}} \big(A^{(n)} - \bar{A}_b\big) u^{(n)} + r_y^{(n)}. \tag{614}$$

For $0 \leq r \leq |I_b|$, define

$$R_{b,r} := \sum_{j=0}^{r-1} P_b^{\,r-1-j} \rho_b^{(s_b+j)} \tag{615}$$

with $R_{b,0} := 0$. Iterating (613) within the block gives

$$y^{(s_b+r)} = P_b^r y^{(s_b)} - \frac{1}{\sqrt{d}} \sum_{j=0}^{r-1} P_b^{r-1-j} \alpha_d \bar{A}_b u^{(s_b+j)} + \sum_{j=0}^{r-1} P_b^{r-1-j} \xi^{(s_b+j+1)} + R_{b,r}. \tag{616}$$

We first control the second term of (616). Diagonalize $\bar{A}_b = Q_b \,\mathrm{diag}(\lambda_1, \dots, \lambda_p) Q_b^T$, where $p := |\mathcal{K}|$. By (592), $\lambda_i \in [a_-, a_+]$. For every coordinate i, using (593),

$$\left| \sum_{j=0}^{r-1} (1 - \alpha_d \lambda_i)^{r-1-j} \alpha_d \lambda_i \left(Q_b^T u^{(s_b+j)} \right)_i \right| \leq C_u \sum_{j=0}^{r-1} (1 - \alpha_d \lambda_i)^{r-1-j} \alpha_d \lambda_i \leq C_u. \tag{617}$$

Consequently,

$$\max_{0 \leq r \leq |I_b|} \left\| \sum_{j=0}^{r-1} P_b^{\,r-1-j} \alpha_d \bar{A}_b u^{(s_b+j)} \right\| \leq \sqrt{p}\, C_u. \tag{618}$$

We next show that $A^{(n)}$ is approximately frozen on each block. Write

$$\Delta\vartheta^{(n)} = \mathbb{E}_n[\Delta\vartheta^{(n)}] + \zeta^{(n+1)}, \qquad \mathbb{E}_n[\zeta^{(n+1)}] = 0. \tag{619}$$

By Lemma 14, uniformly before the stopping time,

$$\left\| \mathbb{E}_n[\Delta\vartheta^{(n)}] \right\| \leq C \left(\frac{\alpha_d}{\sqrt{d}} + \frac{(1+\gamma)\alpha_d}{d} \right), \tag{620}$$

where we used the boundedness of $g^{(n)}$ on (528), while

$$\mathbb{E}_n\|\zeta^{(n+1)}\|^2 \le C(1+\gamma)^2\frac{\alpha_d^2}{d}. \tag{621}$$

Since $\ell_d = \left\lceil \frac{1}{a_-\alpha_d} \right\rceil = O\big((a_-\alpha_d)^{-1}\big)$, the accumulated conditional drift over a single block is $o(1)$. Moreover, Doob's maximal inequality applied on each block, followed by a union bound over the $\mathcal{O}(d)$ blocks, yields

$$\max_b \sup_{n\in I_b} \left\|\vartheta^{(n)} - \vartheta^{(s_b)}\right\| \xrightarrow{\mathbb{P}} 0. \tag{622}$$

Since $A(\vartheta) = S(q)D(\vartheta)S(q)$ is Lipschitz on the compact stopped region (528), it follows that

$$\delta_{A,d} := \max_b \sup_{n\in I_b} \left\|A^{(n)} - \bar{A}_b\right\| \xrightarrow{\mathbb{P}} 0. \tag{623}$$

We now bound the block remainder $R_{b,r}$ (the last term of (616)). Recall from (604) that, on the region $\|y^{(n)}\| \le r_0$,

$$\|r_y^{(n)}\| \le C\left[\alpha_d\|y^{(n)}\|^2 + \frac{\alpha_d}{\sqrt{d}}\|y^{(n)}\| + \frac{(1+\gamma)\alpha_d}{d} + (1+\gamma)^2\alpha_d^2\right]. \tag{624}$$

Using (615), (612), (614), (593), and (624), we obtain, for every $0 \le r \le |I_b|$,

$$\|R_{b,r}\| \le Cr\alpha_d\left(\delta_{A,d} + r_0 + \frac{1}{\sqrt{d}}\right)\max_{0\le j\le r}\|y^{(s_b+j)}\| + Cr\alpha_d\left[\frac{\delta_{A,d}}{\sqrt{d}} + \frac{1+\gamma}{d} + (1+\gamma)^2\alpha_d\right]. \tag{625}$$

Since $r \le |I_b| \le \ell_d$ and $\alpha_d\ell_d \le C/a_-$, this implies

$$\|R_{b,r}\| \le \frac{C}{a_-}\left(r_0 + \delta_{A,d} + \frac{1}{\sqrt{d}}\right)\max_{0\le j\le r}\|y^{(s_b+j)}\| + \frac{C}{a_-}\left[\frac{\delta_{A,d}}{\sqrt{d}} + \frac{1+\gamma}{d} + (1+\gamma)^2\alpha_d\right]. \tag{626}$$

Choose $r_0 = r_0(T,\gamma) > 0$ sufficiently small so that $\frac{C}{a_-}r_0 \le \frac{1}{16}$. By (623), $\delta_{A,d} \xrightarrow{\mathbb{P}} 0$, and therefore

$$\frac{C}{a_-}\left(\delta_{A,d} + \frac{1}{\sqrt{d}}\right) \le \frac{1}{16} \tag{627}$$

with probability tending to one. Furthermore, for fixed T,γ, the learning-rate condition $\alpha_d \le c_0(d\log d)^{-1}$ gives

$$\frac{\delta_{A,d}}{\sqrt{d}} + \frac{1+\gamma}{d} + (1+\gamma)^2\alpha_d = \frac{o_{\mathbb{P}}(1)}{\sqrt{d}}. \tag{628}$$

Consequently, there exists a nonnegative random sequence $\varepsilon_d \xrightarrow{\mathbb{P}} 0$ such that, with probability tending to one,

$$\|R_{b,r}\| \le \frac{1}{8}\max_{0\le j\le r}\|y^{(s_b+j)}\| + \frac{\varepsilon_d}{\sqrt{d}}, \qquad 0 \le r \le |I_b|, \tag{629}$$

uniformly over all blocks b.

Finally, We control the martingale term (the third term of (616)). Truncate the fresh sample at step n by $\mathcal{A}_n = \big\{\|x^{(n)}\|^2 \le c_x d\big\} \cap \Big\{\max_{i,j,k}|G_{ijk}^{(n)}| \le c_G\log d\Big\}$. The probability of any truncation failure before $N_T \wedge J$ tends to zero. On the truncated event,

$$\|\xi^{(n+1)}\| \le C\alpha_d(1 + (\log d)^r) \tag{630}$$

for some fixed $r < \infty$. For each block, Freedman's inequality applied to (616) gives

$$\mathbb{P}\left(\left|\sum_{j=0}^{r-1}(1-\alpha_d\lambda_i)^{r-1-j}\left(Q_b^T\xi^{(s_b+j+1)}\right)_i\right| > \frac{\eta}{\sqrt{d}}\right) \le 2\exp\left[-\frac{c\eta^2 a_-}{d\alpha_d}\right] \tag{631}$$

for every fixed $\eta > 0$ and all sufficiently large d. Here we used

$$\sum_{j\geq 0}(1-\alpha_d\lambda_i)^{2j}\mathbb{E}_{s_b+j}\left[\left|\left(Q_b^T\xi^{(s_b+j+1)}\right)_i\right|^2\right] \leq C\frac{\alpha_d}{a_-}. \tag{632}$$

We now make the estimate uniform over all blocks and all terminal times within each block. Let $\eta_0 := \frac{1}{\sqrt{p}}$ and $p := |\mathcal{K}|$. The total number of pairs (b, r), with b ranging over the blocks and $0 \leq r \leq |I_b|$, is at most $N_T + 1$. Therefore, a union bound gives

$$\mathbb{P}\left(\max_b \max_{0\leq r\leq |I_b|}\max_{1\leq i\leq p}\left|\sum_{j=0}^{r-1}(1-\alpha_d\lambda_i)^{r-1-j}\left(Q_b^T\xi^{(s_b+j+1)}\right)_i\right| > \frac{\eta_0}{\sqrt{d}}\right) \leq 2p(N_T+1)\exp\left[-\frac{c\eta_0^2 a_-}{d\alpha_d}\right]. \tag{633}$$

By the lower learning-rate bound $\alpha_d = \Omega(d^{-\iota})$, $N_T \leq C_T d^{1+\iota}$, while $\frac{1}{d\alpha_d} \geq \frac{\log d}{c_0}$ by the upper learning-rate bound. Hence

$$2p(N_T+1)\exp\left[-\frac{c\eta_0^2 a_-}{d\alpha_d}\right] \leq C_T d^{\,1+\iota-c\eta_0^2 a_-/c_0}. \tag{634}$$

We may therefore choose $c_0^\star = c_0^\star(T, \gamma, C_*, \mathfrak{D}) > 0$ sufficiently small so that

$$\frac{c\eta_0^2 a_-}{c_0^\star} > 1 + \iota. \tag{635}$$

Whenever $c_0 \leq c_0^\star$, the right-hand side of (633) converges to zero. Let $v_{b,r} := \sum_{j=0}^{r-1} P_b^{r-1-j}\xi^{(s_b+j+1)}$. Since $P_b = Q_b\,\mathrm{diag}(1-\alpha_d\lambda_1, \ldots, 1-\alpha_d\lambda_p)Q_b^T$, we have

$$(Q_b^T v_{b,r})_i = \sum_{j=0}^{r-1}(1-\alpha_d\lambda_i)^{r-1-j}(Q_b^T\xi^{(s_b+j+1)})_i. \tag{636}$$

Hence, on the high-probability event obtained by (633),

$$\max_{b,r,i}|(Q_b^T v_{b,r})_i| \leq \frac{\eta_0}{\sqrt{d}}. \tag{637}$$

Choosing $\eta_0 = p^{-1/2}$ and using the orthogonality of Q_b,

$$\|v_{b,r}\| = \|Q_b^T v_{b,r}\| \leq \sqrt{p}\max_{1\leq i\leq p}|(Q_b^T v_{b,r})_i| \leq \frac{1}{\sqrt{d}}. \tag{638}$$

Therefore,

$$\max_b \max_{0\leq r\leq |I_b|}\left\|\sum_{j=0}^{r-1} P_b^{r-1-j}\xi^{(s_b+j+1)}\right\| \leq \frac{1}{\sqrt{d}} \tag{639}$$

with probability tending to one.

Combining (629), (618), and (639), we obtain, on an event whose probability tends to one,

$$\max_{0\leq r\leq |I_b|}\|y^{(s_b+r)}\| \leq 2\|y^{(s_b)}\| + \frac{C_y}{\sqrt{d}}, \tag{640}$$

where $C_y = C_y(C_*, \mathfrak{D}, \mathcal{C}, \mathcal{L}, K, L)$ is independent of γ.

Moreover, for a full block, i.e. whenever $s_b + \ell_d < N_T \wedge J \wedge J_{\mathrm{loc}}$, we additionally have $\|P_b^{\ell_d}\|_{\mathrm{op}} \leq (1-a_-\alpha_d)^{\ell_d} \leq e^{-1}$. Evaluating (616) at the block endpoint and using (640) therefore gives

$$\|y^{(s_b+\ell_d)}\| \leq \|P_b^{\ell_d}\|_{\mathrm{op}}\|y^{(s_b)}\| + \frac{1}{8}\max_{0\leq j\leq \ell_d}\|y^{(s_b+j)}\| + \frac{C}{\sqrt{d}} \leq \theta\|y^{(s_b)}\| + \frac{C}{\sqrt{d}}, \tag{641}$$

where we use (629) for the first inequality and choose $\theta := e^{-1} + \frac{1}{4} \in (0, 1)$ for the second inequality. Iteration gives

$$\|y^{(s_b)}\| \leq \theta^b\|y^{(s_0)}\| + \frac{C_y}{(1-\theta)\sqrt{d}}. \tag{642}$$

We now close the local stopping-time argument. At the entry time, $\|y^{(n_0)}\| \le \frac{r_0}{4}$. By (640), for all sufficiently large d, $\max_{0\le r\le |I_0|} \|y^{(n_0+r)}\| < r_0$. Moreover, by (642), the value at the next block endpoint is again at most $r_0/4$ for all sufficiently large d. Repeating the same argument block by block shows that $J_{\mathrm{loc}} > N_T \wedge J$ with probability tending to one. Hence the local stopping time J_{loc} may be removed.

The number of blocks between n_0 and n is bounded below by $a_-\alpha_d(n-n_0)-2$. Hence (640)–(642) imply

$$\|y^{(n)}\| \le C'_{\mathrm{tr}} e^{-c'_{\mathrm{tr}}\alpha_d n} + \frac{C'_y}{\sqrt{d}}, \qquad n_0 \le n < N_T \wedge J, \tag{643}$$

where $c'_{\mathrm{tr}} = -a_- \log\theta > 0$. The constants C'_{tr} and c'_{tr} may depend on T, γ, whereas C'_y is independent of γ.

We finally convert the estimate on $y^{(n)}$ into the desired estimate on $b_{\mathcal{K}}^{(n)}$. By (589) and Lemma 13,

$$\|b_{\mathcal{K}}^{(n)}\| \le \|g^{(n)}\| + \frac{C}{\sqrt{d}} \le C\|y^{(n)}\| + C\|y^{(n)}\|^2 + \frac{C}{\sqrt{d}}. \tag{644}$$

Combining this with (643), there exist constants $C_1 < \infty$, $C_b < \infty$, and $c_{\mathrm{tr}} := c'_{\mathrm{tr}} > 0$, such that, with probability tending to one,

$$\|b_{\mathcal{K}}^{(n)}\| \le C_1 e^{-c_{\mathrm{tr}}\alpha_d n} + \frac{C_b}{\sqrt{d}}, \qquad n_0 \le n < N_T \wedge J. \tag{645}$$

Here C_b can be chosen independently of γ.

It remains to cover the initial interval $0 \le n < n_0 \wedge J$. By Lemma 13, there exists $B_0 < \infty$, independent of d, such that

$$\sup_{0\le n<J} \|b_{\mathcal{K}}^{(n)}\| \le B_0. \tag{646}$$

For all sufficiently large d, $\alpha_d \le 1$, and hence

$$n < n_0 \quad \Longrightarrow \quad \alpha_d n < \alpha_d n_0 \le t_0 + \alpha_d \le t_0 + 1. \tag{647}$$

Choose $C_{\mathrm{tr}} \ge \max\left\{C_1,\, B_0 e^{c_{\mathrm{tr}}(t_0+1)}\right\}$. Then, for every $0 \le n < n_0 \wedge J$, $B_0 \le C_{\mathrm{tr}} e^{-c_{\mathrm{tr}}\alpha_d n}$. Combining this with (645) yields

$$\|b_{\mathcal{K}}^{(n)}\| \le C_{\mathrm{tr}} e^{-c_{\mathrm{tr}}\alpha_d n} + \frac{C_b}{\sqrt{d}}, \qquad 0 \le n < N_T \wedge J, \tag{648}$$

on an event whose probability tends to one. This proves (586). Removing $\mathcal{B}_d$ is legitimate because $\mathbb{P}(\mathcal{B}_d^c) \to 0$. □

Step 6: Final spectral bounds

Lemma 17. *Fix $T > 0$. Let $\mathfrak{D} \Subset \mathcal{M} \times \mathcal{Q}$ be a compact set containing the limiting trajectory, and define*

$$J_{\mathrm{mac}} := \inf\left\{n \ge 0 : \left(m_{\mathcal{K}}^{(n)}, q^{(1,n)}, q^{(2,n)}\right) \notin \mathfrak{D}\right\}. \tag{649}$$

Under the conditions of Theorem 3, there exists a constant $C_* > C_0$, independent of d, such that

$$\lim_{d\to\infty} \mathbb{P}\left(\sup_{0\le n<N_T\wedge J_{\mathrm{mac}}} \max_{1\le k\le K} \left\{\|U_k^{(n)}\|_{\mathrm{op}}, \|V_k^{(n)}\|_{\mathrm{op}}\right\} \le C_*\right) = 1, \tag{650}$$

where $N_T := \left\lfloor \frac{Td}{\alpha_d} \right\rfloor$.

Proof. The weights with $k \notin \mathcal{K}$ remain fixed throughout training, so it suffices to consider $k \in \mathcal{K}$.

Fix $C_* > C_0$, to be chosen below, and define the spectral stopping time

$$J_{\mathrm{sp}} := \inf\left\{n \ge 0 : \max_{1\le k\le K} \left\{\|U_k^{(n)}\|_{\mathrm{op}}, \|V_k^{(n)}\|_{\mathrm{op}}\right\} > C_*\right\}, \tag{651}$$

and set $J := J_{\rm sp} \wedge J_{\rm mac}$. All estimates below are uniform on $\{n < J\}$.

For $k \in \mathcal{K}$, denote the stochastic gradients by

$$g_{U,k}^{(n)} := \nabla_{U_k} \mathcal{L}(G^{(n)}), \qquad g_{V,k}^{(n)} := \nabla_{V_k} \mathcal{L}(G^{(n)}), \tag{652}$$

and their conditional means by $\bar{g}_{U,k}^{(n)} := \mathbb{E}[g_{U,k}^{(n)} \mid \mathcal{F}_n]$ and $\bar{g}_{V,k}^{(n)} := \mathbb{E}[g_{V,k}^{(n)} \mid \mathcal{F}_n]$. Lemma 13 gives

$$\bar{g}_{U,k}^{(n)} = \frac{b_k^{(n)}}{\sqrt{d}} V_k^{(n)} + \frac{1}{d} H_{U,k}^{(n)} V_k^{(n)} + \mathcal{E}_{U,k}^{(n)}, \tag{653}$$

$$\bar{g}_{V,k}^{(n)} = \frac{b_k^{(n)}}{\sqrt{d}} U_k^{(n)} + \frac{1}{d} H_{V,k}^{(n)} U_k^{(n)} + \mathcal{E}_{V,k}^{(n)} \tag{654}$$

with $\max_{k\in\mathcal{K}} \left\{ \|H_{U,k}^{(n)}\|_{\rm op}, \|H_{V,k}^{(n)}\|_{\rm op} \right\} \le C_H$ and $\max_{k\in\mathcal{K}} \left\{ \|\mathcal{E}_{U,k}^{(n)}\|_{\rm op}, \|\mathcal{E}_{V,k}^{(n)}\|_{\rm op} \right\} \le C_E d^{-3/2}$ uniformly on $\{n < J\}$. Here C_H, C_E are constants independent of d and n.

By Lemma 16: there exist constants $C_{\rm tr}(T,\gamma), c_{\rm tr}(T,\gamma) > 0$ and $C_b(T) < \infty$ such that, with probability tending to one,

$$\max_{k\in\mathcal{K}} |b_k^{(n)}| \le C_{\rm tr}(T,\gamma) e^{-c_{\rm tr}(T,\gamma)\alpha_d n} + \frac{C_b(T)}{\sqrt{d}}, \qquad 0 \le n < N_T \wedge J. \tag{655}$$

Define

$$\mathcal{A}_n := \left\{ \|x^{(n)}\|^2 \le c_x d \right\} \cap \left\{ \max_{i,j,k} |G_{ijk}^{(n)}| \le c_G \log d \right\}, \tag{656}$$

where c_x, c_G are sufficiently large constants. Thus we have

$$\mathbb{P}\left(\exists\, n < N_T \wedge J : \mathcal{A}_n^c\right) \longrightarrow 0. \tag{657}$$

Denote the complementary high-probability event by $\mathcal{A}_{\rm data}$.

Define the stopped and truncated gradients $\widehat{g}_{U,k}^{(n)} := \mathbf{1}_{\{n<J\}} g_{U,k}^{(n)} \mathbf{1}_{\mathcal{A}_n}$, $\widehat{g}_{V,k}^{(n)} := \mathbf{1}_{\{n<J\}} g_{V,k}^{(n)} \mathbf{1}_{\mathcal{A}_n}$, their conditional means $\widehat{\bar{g}}_{U,k}^{(n)}, \widehat{\bar{g}}_{V,k}^{(n)}$, and the martingale differences $\widehat{\xi}_{U,k}^{(n)} := \widehat{g}_{U,k}^{(n)} - \widehat{\bar{g}}_{U,k}^{(n)}$, $\widehat{\xi}_{V,k}^{(n)} := \widehat{g}_{V,k}^{(n)} - \widehat{\bar{g}}_{V,k}^{(n)}$.

Set $a_d := 1 - \frac{\alpha_d \gamma}{d}$. Define the discounted martingale sums

$$Y_{U,k,m} := -\alpha_d \sum_{n=0}^{m-1} a_d^{m-1-n} \widehat{\xi}_{U,k}^{(n)}, \quad Y_{V,k,m} := -\alpha_d \sum_{n=0}^{m-1} a_d^{m-1-n} \widehat{\xi}_{V,k}^{(n)}. \tag{658}$$

Similarly to Lemma 3 we have

$$\mathbb{P}\left(\max_{\substack{k\in\mathcal{K} \\ 1\le m\le N_T}} \max\left\{ \|Y_{U,k,m}\|_{\rm op}, \|Y_{V,k,m}\|_{\rm op} \right\} > \delta \right) \longrightarrow 0. \tag{659}$$

Denote the complementary event by $\mathcal{A}_{\rm noise}$. Also the truncation-tail corrections are small:

$$\mathbf{1}_{\{n<J\}} \max\left\{ \|\mathbb{E}\left[g_{U,k}^{(n)} \mathbf{1}_{\mathcal{A}_n^c} \,\middle|\, \mathcal{F}_n \right]\|_{\rm op}, \|\mathbb{E}\left[g_{V,k}^{(n)} \mathbf{1}_{\mathcal{A}_n^c} \,\middle|\, \mathcal{F}_n \right]\|_{\rm op} \right\} \le r_d, \tag{660}$$

where $r_d = o(d^{-M})$ for every fixed $M > 0$.

On $\mathcal{A}_{\rm data}$, for every $m \le J$ and $n < m$ unrolling the SGD recursion gives

$$U_k^{(m)} = a_d^m U_k^{(0)} - \frac{\alpha_d}{\sqrt{d}} \sum_{n=0}^{m-1} a_d^{m-1-n} b_k^{(n)} V_k^{(n)} - \frac{\alpha_d}{d} \sum_{n=0}^{m-1} a_d^{m-1-n} H_{U,k}^{(n)} V_k^{(n)} + Y_{U,k,m} + R_{U,k,m}, \tag{661}$$

where, uniformly for $m \le J \wedge N_T$, $\|R_{U,k,m}\|_{\rm op} \le \alpha_d \sum_{n=0}^{m-1} a_d^{m-1-n} \left(C_E d^{-3/2} + r_d \right) = o(1)$. The analogous expansion holds for $V_k^{(m)}$.

We now bound the second term of (661). On the high-probability event where (655) holds, the first term of (655) contributes as

$$\frac{\alpha_d}{\sqrt{d}} \sum_{n=0}^{m-1} a_d^{m-1-n} C_{\mathrm{tr}}(T,\gamma) e^{-c_{\mathrm{tr}}(T,\gamma)\alpha_d n} \|V_k^{(n)}\|_{\mathrm{op}} \le \frac{C(T,\gamma)}{\sqrt{d}} = o(1), \tag{662}$$

uniformly in $m \le J \wedge N_T$, where we use $\sum_{n\ge 0} e^{-c_{\mathrm{tr}}(T,\gamma)\alpha_d n} \le C(T,\gamma)/\alpha_d$. Using $\alpha_d \sum_{r=0}^{\infty} a_d^r = \frac{d}{\gamma}$, the second term of (655) contributes as

$$\frac{\alpha_d}{\sqrt{d}} \sum_{n=0}^{m-1} a_d^{m-1-n} \frac{C_b(T)}{\sqrt{d}} \|V_k^{(n)}\|_{\mathrm{op}} \le \frac{C_b(T) C_*}{\gamma}. \tag{663}$$

The third term of (661) can be bounded as

$$\frac{\alpha_d}{d} \sum_{n=0}^{m-1} a_d^{m-1-n} \|H_{U,k}^{(n)} V_k^{(n)}\|_{\mathrm{op}} \le \frac{C_H C_*}{\gamma}. \tag{664}$$

Combining (661)– (664), on $\mathcal{A}_{\mathrm{data}} \cap \mathcal{A}_{\mathrm{noise}}$ and the high-probability event (655),

$$\|U_k^{(m)}\|_{\mathrm{op}} \le C_0 + \frac{(C_b(T) + C_H) C_*}{\gamma} + \delta + o(1), \qquad m \le J \wedge N_T. \tag{665}$$

The same estimate holds for $V_k^{(m)}$.

Choose $C_* := 2C_0 + 1$ and $\delta := \frac{C_* - C_0}{4}$. By Assumption 9.5, choose γ sufficiently large (which might depend on T) such that

$$\frac{(C_b(T) + C_H) C_*}{\gamma} \le \frac{C_* - C_0}{4}. \tag{666}$$

Therefore, for all sufficiently large d,

$$\max_{k \in \mathcal{K}} \max \left\{ \|U_k^{(m)}\|_{\mathrm{op}}, \|V_k^{(m)}\|_{\mathrm{op}} \right\} < C_* \tag{667}$$

for every $m \le J \wedge N_T$ on an event whose probability tends to one. Hence

$$\mathbb{P}\left(J_{\mathrm{sp}} \le J_{\mathrm{mac}} \wedge N_T\right) \longrightarrow 0. \tag{668}$$

This is precisely (650). □

F.2.4 Fast-dynamics convergence

We now prove the convergence on the fast timescale. Fix a finite fast-time horizon $\widetilde{T} > 0$, and set $N_{\widetilde{T}}^{\mathrm{f}} := \left\lfloor \frac{\widetilde{T}}{\alpha_d} \right\rfloor$. Let $D_0 := \mathrm{diag}\left(\bar{t}_k^A(0) + \bar{t}_k^B(0)\right)_{k \in \mathcal{K}}$, $q_0 := \left(\bar{q}^{(1)}(0), \bar{q}^{(2)}(0)\right)$ and let $\tilde{m}_{\mathcal{K}}(t)$ be the solution of

$$\frac{d\tilde{m}_{\mathcal{K}}}{dt} = -D_0 \nabla_{m_{\mathcal{K}}} \Phi\left(\tilde{m}(t), q_0\right), \qquad \tilde{m}_{\mathcal{K}}(0) = \lim_{d \to \infty} m_{\mathcal{K}}^{(d)}(0), \tag{669}$$

with the non-trainable components fixed at their initial values.

By Assumption 10,

$$\frac{d}{dt} \Phi(\tilde{m}(t), q_0) = -\nabla_{m_{\mathcal{K}}} \Phi(\tilde{m}(t), q_0)^T D_0 \nabla_{m_{\mathcal{K}}} \Phi(\tilde{m}(t), q_0) \le 0. \tag{670}$$

Hence the trajectory $\{\tilde{m}_{\mathcal{K}}(t) : 0 \le t \le \widetilde{T}\}$ remains in a compact subset of $\mathcal{M}$. Choose compact neighborhoods $\mathfrak{D}_{\mathrm{f}} \Subset \mathcal{M} \times \mathcal{Q}$ and C_* in Lemma 17 such that the deterministic trajectory $(\tilde{m}_{\mathcal{K}}(t), q_0)$, $0 \le t \le \widetilde{T}$, lies in the interior of $\mathfrak{D}_{\mathrm{f}}$. Define the stopping times

$$J_{\mathrm{f}} := \inf \left\{ n \ge 0 : \left(m_{\mathcal{K}}^{(n)}, q^{(1,n)}, q^{(2,n)} \right) \notin \mathfrak{D}_{\mathrm{f}} \right\}, \tag{671}$$

$$J_{sp} := \inf \left\{ n \ge 0 : \max_k \left\{ \|U_k^{(n)}\|_{\mathrm{op}}, \|V_k^{(n)}\|_{\mathrm{op}} \right\} > C_* \right\}, \tag{672}$$

and set $J := J_{\mathrm{f}} \wedge J_{sp}$. The spectral estimate of Lemma 17 gives

$$\mathbb{P}\left(J_{\mathrm{sp}} \le J_{\mathrm{f}} \wedge N^{\mathrm{f}}_{\widetilde{T}}\right) \longrightarrow 0. \tag{673}$$

We first show that all slow coordinates remain frozen on the fast timescale. Recall $\vartheta^{(n)} := \left(t^{A,(n)}, t^{B,(n)}, q^{(1,n)}, q^{(2,n)}\right)$. Write

$$\Delta\vartheta^{(n)} = \mathbb{E}_n[\Delta\vartheta^{(n)}] + \zeta_\vartheta^{(n+1)}, \qquad \mathbb{E}_n[\zeta_\vartheta^{(n+1)}] = 0. \tag{674}$$

On $\{n < J\}$, Lemma 14 and the boundedness of $\nabla_{m_\mathcal{K}}\Phi$ on $\mathfrak{D}_{\mathrm{f}}$ give

$$\left\|\mathbb{E}_n[\Delta\vartheta^{(n)}]\right\| \le C\left(\frac{\alpha_d}{\sqrt{d}} + \frac{(1+\gamma)\alpha_d}{d}\right) \tag{675}$$

and

$$\mathbb{E}_n\left[\|\zeta_\vartheta^{(n+1)}\|^2\right] \le C(1+\gamma)^2\frac{\alpha_d^2}{d}. \tag{676}$$

Therefore the accumulated conditional drift over $N^{\mathrm{f}}_{\widetilde{T}} = \mathcal{O}(\alpha_d^{-1})$ steps satisfies

$$\sup_{m \le N^{\mathrm{f}}_{\widetilde{T}}} \left\|\sum_{n=0}^{m-1} \mathbf{1}_{\{n<J\}}\mathbb{E}_n[\Delta\vartheta^{(n)}]\right\| \le C_{\widetilde{T}}\left(d^{-1/2} + \frac{1+\gamma}{d}\right) = o(1). \tag{677}$$

On the other hand, Doob's maximal inequality and (676) imply

$$\mathbb{E}\left[\sup_{m \le N^{\mathrm{f}}_{\widetilde{T}}} \left\|\sum_{n=0}^{m-1} \mathbf{1}_{\{n<J\}}\zeta_\vartheta^{(n+1)}\right\|^2\right] \le CN^{\mathrm{f}}_{\widetilde{T}}\frac{\alpha_d^2}{d} \le C_{\widetilde{T}}\frac{\alpha_d}{d} \longrightarrow 0. \tag{678}$$

Consequently,

$$\sup_{0 \le n \le N^{\mathrm{f}}_{\widetilde{T}} \wedge J} \left\|\vartheta^{(n)} - \vartheta^{(0)}\right\| \xrightarrow{\mathbb{P}} 0. \tag{679}$$

Together with Assumption 9.3, this yields

$$\sup_{0 \le n \le N^{\mathrm{f}}_{\widetilde{T}} \wedge J} \left\|\vartheta^{(n)} - \bar{\vartheta}(0)\right\| \xrightarrow{\mathbb{P}} 0. \tag{680}$$

In particular,

$$\sup_{0 \le n \le N^{\mathrm{f}}_{\widetilde{T}} \wedge J} \left(\|D^{(n)} - D_0\| + \|q^{(n)} - q_0\|\right) \xrightarrow{\mathbb{P}} 0. \tag{681}$$

We next turn to the mean variables. Define

$$F(m) := -D_0\nabla_{m_\mathcal{K}}\Phi(m, q_0). \tag{682}$$

By Lemma 14, on $\{n < J\}$,

$$\mathbb{E}_n[\Delta m_\mathcal{K}^{(n)}] = \alpha_d F(m^{(n)}) + \alpha_d\rho_m^{(n)}, \tag{683}$$

where, using (681), the local Lipschitz continuity of $\nabla_{m_\mathcal{K}}\Phi$, and the bound $\|e^{(n)}\| \le C$,

$$\sup_{0 \le n < N^{\mathrm{f}}_{\widetilde{T}} \wedge J} \|\rho_m^{(n)}\| \xrightarrow{\mathbb{P}} 0. \tag{684}$$

This is because

$$\|\rho_m^{(n)}\| \le C\|D^{(n)} - D_0\| + C\|q^{(n)} - q_0\| + \frac{C}{\sqrt{d}} + C\left(\frac{1+\gamma}{d} + \alpha_d\right). \tag{685}$$

Let

$$\zeta_m^{(n+1)} := \Delta m_\mathcal{K}^{(n)} - \mathbb{E}_n[\Delta m_\mathcal{K}^{(n)}]. \tag{686}$$

Again by Lemma 14, $\mathbb{E}_n\left[\|\zeta_m^{(n+1)}\|^2\right] \le C\alpha_d^2$. Hence Doob's maximal inequality gives

$$\sup_{m \le N_{\widetilde{T}}^{\mathrm{f}}} \left\| \sum_{n=0}^{m-1} \mathbf{1}_{\{n<J\}} \zeta_m^{(n+1)} \right\| \xrightarrow{\mathbb{P}} 0, \tag{687}$$

because its quadratic variation is bounded by $CN_{\widetilde{T}}^{\mathrm{f}}\alpha_d^2 \le C_{\widetilde{T}}\alpha_d \longrightarrow 0$.

Summing (683), we therefore obtain, uniformly for $m \le N_{\widetilde{T}}^{\mathrm{f}} \wedge J$,

$$m_{\mathcal{K}}^{(m)} = m_{\mathcal{K}}^{(0)} + \alpha_d \sum_{n=0}^{m-1} F(m^{(n)}) + o_{\mathbb{P}}(1), \tag{688}$$

where the error is uniform in m. Since F is Lipschitz on the compact neighborhood $\mathfrak{D}_{\mathrm{f}}$, the standard discrete error estimate, together with Assumption 9.3 and the Grönwall inequality (see the proof of Theorem 4), gives

$$\sup_{0 \le n \le N_{\widetilde{T}}^{\mathrm{f}} \wedge J} \left\| m_{\mathcal{K}}^{(n)} - \tilde{m}_{\mathcal{K}}(n\alpha_d) \right\| \xrightarrow{\mathbb{P}} 0. \tag{689}$$

It remains to remove the stopping time. By construction, the deterministic trajectory $\{(\tilde{m}_{\mathcal{K}}(t), q_0) : 0 \le t \le \widetilde{T}\}$ has positive distance from $\partial\mathfrak{D}_{\mathrm{f}}$. Equations (681) and (689) therefore imply

$$\mathbb{P}\left(J_{\mathrm{f}} \le N_{\widetilde{T}}^{\mathrm{f}} \wedge J_{sp}\right) \longrightarrow 0. \tag{690}$$

Combining (673) and (690), we conclude that

$$\mathbb{P}\left(J \le N_{\widetilde{T}}^{\mathrm{f}}\right) \longrightarrow 0. \tag{691}$$

Hence the stopping may be removed from (689). Under the piecewise-constant interpolation, we obtain

$$\sup_{t \in [0,\widetilde{T}]} \left\| \tilde{m}^{(d)}(t) - \tilde{m}(t) \right\|_{\infty} \xrightarrow{\mathbb{P}} 0, \tag{692}$$

which proves (20).

F.2.5 Moment expansion

Step 1: One-step expansion For $k \in \mathcal{K}$, define the centered gradient noises $\xi_{U,k}^{(n)} := g_{U,k}^{(n)} - \bar{g}_{U,k}^{(n)}$, $\xi_{V,k}^{(n)} := g_{V,k}^{(n)} - \bar{g}_{V,k}^{(n)}$, where $\bar{g}_{U,k}^{(n)} := \mathbb{E}_n[g_{U,k}^{(n)}]$, $\bar{g}_{V,k}^{(n)} = \mathbb{E}_n[g_{V,k}^{(n)}]$. Let $a_d := 1 - \frac{\alpha_d \gamma}{d}$. For $Z_k \in \{M_k, M_k^T, A_k, B_k\}$, define

$$\mathscr{N}_k^{(n)}(Z_k) := \begin{cases} -a_d\left(\xi_{U,k}^{(n)} V_k^T + U_k(\xi_{V,k}^{(n)})^T\right), & Z_k = M_k, \\ -a_d\left(V_k(\xi_{U,k}^{(n)})^T + \xi_{V,k}^{(n)} U_k^T\right), & Z_k = M_k^T, \\ -a_d\left(\xi_{U,k}^{(n)} U_k^T + U_k(\xi_{U,k}^{(n)})^T\right), & Z_k = A_k, \\ -a_d\left(\xi_{V,k}^{(n)} V_k^T + V_k(\xi_{V,k}^{(n)})^T\right), & Z_k = B_k. \end{cases} \tag{693}$$

For $k \notin \mathcal{K}$, set $\mathscr{N}_k^{(n)}(Z_k) = 0$. Define the finite-dimensional raw vector field as

$$[\mathcal{V}_{\mathrm{raw}}^{(d,n)}]_\alpha := -\sum_{\substack{i=1 \\ k_i \in \mathcal{K}}}^{w} \sqrt{d}\, b_{k_i}^{(n)} \mathcal{T}_i^{\mathrm{bias}}(\mu^{(n)}) - 2\sum_{\substack{i=1 \\ k_i \in \mathcal{K}}}^{w} \sum_{l=1}^{K} \sum_{r=1}^{2} a_{k_i l}^{(r,d,n)} \mathcal{T}_{i,l,r}^{\mathrm{diff}}(\mu^{(n)}) - 2\gamma|\alpha|_{\mathcal{K}} \mu_\alpha^{(n)}. \tag{694}$$

Lemma 18. *Under the conditions of Theorem 3, let J be a stopping time satisfying* (528)*. Fix a finite word $\alpha = (Z_1, \dots, Z_w) \in \mathscr{Z}^w$, and let k_i denote the index carried by the letter Z_i. Then, uniformly on $\{n < J\}$,*

$$\mu_\alpha^{(n+1)} - \mu_\alpha^{(n)} = \frac{\alpha_d}{d}[\mathcal{V}_{\mathrm{raw}}^{(d,n)}]_\alpha + \alpha_d \eta_\alpha^{(n)} + R_\alpha^{(n)}, \tag{695}$$

where

$$\eta_\alpha^{(n)} := \sum_{\substack{i=1 \\ k_i \in \mathcal{K}}}^{w} \frac{1}{d} \mathrm{Tr}\left[Z_1^{(n)} \cdots Z_{i-1}^{(n)} \mathscr{N}_{k_i}^{(n)}(Z_i) Z_{i+1}^{(n)} \cdots Z_w^{(n)} \right] \tag{696}$$

is a martingale difference: $\mathbb{E}_n[\eta_\alpha^{(n)}] = 0$*. Moreover, for every fixed word* α*, there exists* $C_\alpha < \infty$*, independent of* d *and* n*, such that*

$$\mathbf{1}_{\{n<J\}} \mathbb{E}_n\left[|\eta_\alpha^{(n)}|^2 \right] \le \frac{C_\alpha}{d}, \tag{697}$$

and

$$\mathbf{1}_{\{n<J\}} \mathbb{E}_n\left[|R_\alpha^{(n)}| \right] \le C_\alpha \left[(1+\gamma)^2 \alpha_d^2 + \frac{\alpha_d}{d^{3/2}} \right]. \tag{698}$$

Proof. Throughout the proof we condition on $\mathcal{F}_n$ and suppress the superscript (n). Thus U_k, V_k, M_k, A_k, B_k are deterministic conditional on $\mathcal{F}_n$, while the stochastic gradients depend on the fresh sample $x^{(n)}$. All estimates below are uniform over (528).

For $k \in \mathcal{K}$, the SGD recursion can be written as $U_k^+ = a_d U_k - \alpha_d g_{U,k}$ and $V_k^+ = a_d V_k - \alpha_d g_{V,k}$. Therefore

$$M_k^+ = a_d^2 M_k - \alpha_d a_d \left(g_{U,k} V_k^T + U_k g_{V,k}^T \right) + \alpha_d^2 g_{U,k} g_{V,k}^T, \tag{699}$$

$$A_k^+ = a_d^2 A_k - \alpha_d a_d \left(g_{U,k} U_k^T + U_k g_{U,k}^T \right) + \alpha_d^2 g_{U,k} g_{U,k}^T, \tag{700}$$

$$B_k^+ = a_d^2 B_k - \alpha_d a_d \left(g_{V,k} V_k^T + V_k g_{V,k}^T \right) + \alpha_d^2 g_{V,k} g_{V,k}^T. \tag{701}$$

Write $g_{U,k} = \bar{g}_{U,k} + \xi_{U,k}$ and $g_{V,k} = \bar{g}_{V,k} + \xi_{V,k}$. By Lemma 13,

$$\bar{g}_{U,k} = \frac{b_k}{\sqrt{d}} V_k + \frac{2}{d} \sum_{l=1}^{K} \sum_{r=1}^{2} a_{kl}^{(r,d)} M_l^{(r)} V_k + \mathcal{E}_{U,k}, \tag{702}$$

$$\bar{g}_{V,k} = \frac{b_k}{\sqrt{d}} U_k + \frac{2}{d} \sum_{l=1}^{K} \sum_{r=1}^{2} a_{kl}^{(r,d)} M_l^{(\bar{r})} U_k + \mathcal{E}_{V,k}, \tag{703}$$

where

$$|b_k| + \max_{l,r} |a_{kl}^{(r,d)}| \le C, \qquad \|\mathcal{E}_{U,k}\|_{\mathrm{op}} + \|\mathcal{E}_{V,k}\|_{\mathrm{op}} \le C d^{-3/2}. \tag{704}$$

Substituting (702)–(703) into (699), using $(M_l^{(\bar{r})})^T = M_l^{(r)}$, we obtain

$$\Delta M_k = -\frac{\alpha_d b_k}{\sqrt{d}} (A_k + B_k) - \frac{2\alpha_d}{d} \sum_{l=1}^{K} \sum_{r=1}^{2} a_{kl}^{(r,d)} \left(M_l^{(r)} B_k + A_k M_l^{(r)} \right) - \frac{2\alpha_d \gamma}{d} M_k + \alpha_d \mathscr{N}_k(M_k) + \mathscr{R}_k(M_k). \tag{705}$$

Taking the transpose gives

$$\Delta M_k^T = -\frac{\alpha_d b_k}{\sqrt{d}} (A_k + B_k) - \frac{2\alpha_d}{d} \sum_{l=1}^{K} \sum_{r=1}^{2} a_{kl}^{(r,d)} \left(B_k M_l^{(\bar{r})} + M_l^{(\bar{r})} A_k \right) - \frac{2\alpha_d \gamma}{d} M_k^T + \alpha_d \mathscr{N}_k(M_k^T) + \mathscr{R}_k(M_k^T). \tag{706}$$

Likewise, substituting (702) into (700) gives

$$\Delta A_k = -\frac{\alpha_d b_k}{\sqrt{d}} (M_k + M_k^T) - \frac{2\alpha_d}{d} \sum_{l=1}^{K} \sum_{r=1}^{2} a_{kl}^{(r,d)} \left(M_l^{(r)} M_k^T + M_k M_l^{(\bar{r})} \right) - \frac{2\alpha_d \gamma}{d} A_k + \alpha_d \mathscr{N}_k(A_k) + \mathscr{R}_k(A_k), \tag{707}$$

whereas (703) and (701) give

$$\Delta B_k = -\frac{\alpha_d b_k}{\sqrt{d}} (M_k + M_k^T) - \frac{2\alpha_d}{d} \sum_{l=1}^{K} \sum_{r=1}^{2} a_{kl}^{(r,d)} \left(M_l^{(\bar{r})} M_k + M_k^T M_l^{(r)} \right) - \frac{2\alpha_d \gamma}{d} B_k + \alpha_d \mathscr{N}_k(B_k) + \mathscr{R}_k(B_k). \tag{708}$$

We next make the remainders $\mathscr{R}_k(Z_k)$ explicit. For $Z_k = M_k$,

$$\begin{aligned}\mathscr{R}_k(M_k) =&\alpha_d(1-a_d)\left[\frac{b_k}{\sqrt{d}}(A_k+B_k)+\frac{2}{d}\sum_{l=1}^{K}\sum_{r=1}^{2}a_{kl}^{(r,d)}\left(M_l^{(r)}B_k+A_kM_l^{(r)}\right)\right]\\&-\alpha_d a_d\left(\mathcal{E}_{U,k}V_k^T+U_k\mathcal{E}_{V,k}^T\right)+\frac{\alpha_d^2\gamma^2}{d^2}M_k+\alpha_d^2 g_{U,k}g_{V,k}^T,\end{aligned}\tag{709}$$

which follows directly from $a_d^2-1=-\frac{2\alpha_d\gamma}{d}+\frac{\alpha_d^2\gamma^2}{d^2}$ and $1-a_d=\frac{\alpha_d\gamma}{d}$. Moreover, $\mathscr{R}_k(M_k^T)=\mathscr{R}_k(M_k)^T$. Similarly,

$$\begin{aligned}\mathscr{R}_k(A_k) =\alpha_d(1-a_d)&\left[\frac{b_k}{\sqrt{d}}(M_k+M_k^T)+\frac{2}{d}\sum_{l=1}^{K}\sum_{r=1}^{2}a_{kl}^{(r,d)}\left(M_l^{(r)}M_k^T+M_kM_l^{(\bar{r})}\right)\right]\\&-\alpha_d a_d\left(\mathcal{E}_{U,k}U_k^T+U_k\mathcal{E}_{U,k}^T\right)+\frac{\alpha_d^2\gamma^2}{d^2}A_k+\alpha_d^2 g_{U,k}g_{U,k}^T,\end{aligned}\tag{710}$$

and

$$\begin{aligned}\mathscr{R}_k(B_k) =\alpha_d(1-a_d)&\left[\frac{b_k}{\sqrt{d}}(M_k+M_k^T)+\frac{2}{d}\sum_{l=1}^{K}\sum_{r=1}^{2}a_{kl}^{(r,d)}\left(M_l^{(\bar{r})}M_k+M_k^TM_l^{(r)}\right)\right]\\&-\alpha_d a_d\left(\mathcal{E}_{V,k}V_k^T+V_k\mathcal{E}_{V,k}^T\right)+\frac{\alpha_d^2\gamma^2}{d^2}B_k+\alpha_d^2 g_{V,k}g_{V,k}^T.\end{aligned}\tag{711}$$

We now estimate these remainders in the form needed for the moment expansion. Let P be any $\mathcal{F}_n$-measurable $d\times d$ matrix satisfying $\|P\|_{\mathrm{op}}\le C_P$. On the stopped (528), all letters satisfy

$$\max_{1\le l\le K}\{\|M_l\|_{\mathrm{op}},\|A_l\|_{\mathrm{op}},\|B_l\|_{\mathrm{op}}\}\le C_*^2.\tag{712}$$

Hence every fixed product of the letters appearing in (709)– (711) has uniformly bounded operator norm, i.e., $\frac{1}{d}|\mathrm{Tr}(PQ)|\le\|P\|_{\mathrm{op}}\|Q\|_{\mathrm{op}}\le C$ for every such product Q.

Since $1-a_d=\frac{\alpha_d\gamma}{d}$, the terms are therefore bounded by

$$\alpha_d(1-a_d)\left(\frac{C}{\sqrt{d}}+\frac{C}{d}\right)\le C\alpha_d^2\gamma\left(d^{-3/2}+d^{-2}\right).\tag{713}$$

Likewise,

$$\frac{\alpha_d^2\gamma^2}{d^2}\frac{1}{d}|\mathrm{Tr}(PZ_k)|\le C\frac{\alpha_d^2\gamma^2}{d^2}.\tag{714}$$

For the remainders, Lemma 13 gives $\|\mathcal{E}_{U,k}\|_{\mathrm{op}}+\|\mathcal{E}_{V,k}\|_{\mathrm{op}}\le Cd^{-3/2}$. Since $\|U_k\|_{\mathrm{op}},\|V_k\|_{\mathrm{op}}\le C_*$, we obtain, for example,

$$\frac{1}{d}\left|\mathrm{Tr}(P\mathcal{E}_{U,k}V_k^T)\right|\le\|P\mathcal{E}_{U,k}V_k^T\|_{\mathrm{op}}\le C_P\|\mathcal{E}_{U,k}\|_{\mathrm{op}}\|V_k\|_{\mathrm{op}}\le Cd^{-3/2}.\tag{715}$$

The same estimate applies to all the other Stein-remainder terms, and hence their total contribution is bounded by $C\alpha_d d^{-3/2}$.

Finally, consider the quadratic-gradient terms. For instance, by Cauchy–Schwarz and (519),

$$\mathbb{E}_n\left[\frac{1}{d}\left|\mathrm{Tr}\left[Pg_{U,k}g_{V,k}^T\right]\right|\right]\le\frac{C_P}{d}\left(\mathbb{E}_n\|g_{U,k}\|_F^2\right)^{1/2}\left(\mathbb{E}_n\|g_{V,k}\|_F^2\right)^{1/2}\le C.\tag{716}$$

The same argument gives

$$\mathbb{E}_n\left[\frac{1}{d}\left|\mathrm{Tr}\left[Pg_{U,k}g_{U,k}^T\right]\right|+\frac{1}{d}\left|\mathrm{Tr}\left[Pg_{V,k}g_{V,k}^T\right]\right|\right]\le C.\tag{717}$$

Thus the quadratic-gradient part of the letter remainder contributes at most $C\alpha_d^2$.

Combining (713)– (716), we conclude that, uniformly on $\{n<J\}$,

$$\mathbb{E}_n\left[\left|\frac{1}{d}\mathrm{Tr}\left[P\mathscr{R}_k(Z_k)\right]\right|\right]\le C_P\left[(1+\gamma)^2\alpha_d^2+\frac{\alpha_d}{d^{3/2}}\right],\qquad Z_k\in\{M_k,M_k^T,A_k,B_k\}.\tag{718}$$

Fix $\alpha = (Z_1, \dots, Z_w)$. Writing $Z_i^+ = Z_i + \Delta Z_i$ in $\mu_\alpha^+ = \frac{1}{d}\mathrm{Tr}[Z_1^+ \cdots Z_w^+]$ and expanding the product gives

$$\mu_\alpha^+ - \mu_\alpha = \sum_{i=1}^{w} \frac{1}{d}\mathrm{Tr}\left[Z_1 \cdots Z_{i-1} \Delta Z_i Z_{i+1} \cdots Z_w\right] + R_{\alpha,\mathrm{mult}}, \tag{719}$$

where $R_{\alpha,\mathrm{mult}}$ is the sum of all terms containing increments at two or more positions.

Insert (705)– (708) into the first-order sum in (719). By the definition (18), we obtain

$$\mu_\alpha^+ - \mu_\alpha = \frac{\alpha_d}{d}[\mathcal{V}_{\mathrm{raw}}^{(d)}]_\alpha + \alpha_d \eta_\alpha + R_\alpha. \tag{720}$$

$\eta_\alpha^{(n)}$ is defined in (696). By $\mathbb{E}_n[\mathscr{N}_{k_i}(Z_i)] = 0$ we immediately obtain $\mathbb{E}_n[\eta_\alpha] = 0$. Moreover, on (528) all letters satisfy $\max_{Z \in \mathscr{Z}} \|Z\|_{\mathrm{op}} \le C_*^2$, which gives $\mathbb{E}_n|\eta_\alpha|^2 \le \frac{C_\alpha}{d}$ and proves (697). R_α is the sum of the first-order letter remainders and $R_{\alpha,\mathrm{mult}}$.

The final step is to control the first-order letter remainders and $R_{\alpha,\mathrm{mult}}$. By (718), the sum of all first-order letter remainders satisfies

$$\mathbb{E}_n \left| \sum_{\substack{i=1 \\ k_i \in \mathcal{K}}}^{w} \frac{1}{d}\mathrm{Tr}\left[Z_1 \cdots Z_{i-1} \mathscr{R}_{k_i}(Z_i) Z_{i+1} \cdots Z_w\right] \right| \le C_\alpha \left[(1+\gamma)^2 \alpha_d^2 + \frac{\alpha_d}{d^{3/2}}\right]. \tag{721}$$

We next consider products containing at least two letter increments. The spectral bounds (528), polynomial growth of the derivatives of $\mathcal{L}$, and Gaussian moment estimates imply, for every fixed $p < \infty$,

$$\mathbb{E}_n\left[\|\Delta Z_k\|_F^p\right] \le C_p (1+\gamma)^p \alpha_d^p d^{p/2}, \qquad Z_k \in \{M_k, M_k^T, A_k, B_k\}. \tag{722}$$

The same estimate holds with the Frobenius norm replaced by the operator norm, since $\|\cdot\|_{\mathrm{op}} \le \|\cdot\|_F$.

Consider a term in $R_{\alpha,\mathrm{mult}}$ containing $r \ge 2$ increments. Hölder's inequality gives

$$\frac{1}{d}\mathbb{E}_n\left|\mathrm{Tr}\left[P_0 \Delta Z_{i_1} P_1 \cdots \Delta Z_{i_r} P_r\right]\right| \le C_\alpha (1+\gamma)^r \alpha_d^r d^{(r-2)/2}, \tag{723}$$

where the P_j's are products of unchanged letters. Since the length w is fixed and $\alpha_d\sqrt{d} \longrightarrow 0$, under the learning-rate condition of Theorem 3, for every $2 \le r \le w$,

$$\alpha_d^r d^{(r-2)/2} = \alpha_d^2 (\alpha_d \sqrt{d})^{r-2} \le C \alpha_d^2 \tag{724}$$

for all sufficiently large d. Summing over the finitely many multi-increment terms therefore gives

$$\mathbb{E}_n|R_{\alpha,\mathrm{mult}}| \le C_\alpha (1+\gamma)^2 \alpha_d^2. \tag{725}$$

Combining (721) and (725), and defining $R_\alpha^{(n)}$ to be their sum, yields

$$\mathbb{E}_n|R_\alpha^{(n)}| \le C_\alpha \left[(1+\gamma)^2 \alpha_d^2 + \frac{\alpha_d}{d^{3/2}}\right]. \tag{726}$$

This proves (698), and completes the proof. □

Step 2: Vanishing errors For a fixed word $\alpha \in \mathscr{Z}^w$, define the stopped interpolation

$$\tilde{\mu}_{\alpha,J}^{(d)}(\tau) := \mu_\alpha^{(m_d(\tau)\wedge J)}, \qquad m_d(\tau) := \left\lfloor \frac{\tau}{\tau_d} \right\rfloor, \qquad 0 \le \tau \le T. \tag{727}$$

Also define the piecewise-constant stopped finite-dimensional vector field by

$$[\widetilde{\mathcal{V}}_{\mathrm{raw},J}^{(d)}(\tau)]_\alpha := \mathbf{1}_{\{m_d(\tau) < J\}} [\mathcal{V}_{\mathrm{raw}}^{(d, m_d(\tau))}]_\alpha. \tag{728}$$

The following lemma proves the integral equation for the interpolated moments, in the spirit of Lemma 4.

Lemma 19. *Under the conditions of Theorem 3, let J be a stopping time satisfying* (528)*. Fix $T > 0$, and set $\tau_d := \frac{\alpha_d}{d}$, $N_T := \left\lfloor \frac{T}{\tau_d} \right\rfloor = \left\lfloor \frac{Td}{\alpha_d} \right\rfloor$. Then there exists $E_\alpha^{(d)} : [0,T] \to \mathbb{R}$ such that*

$$\tilde{\mu}_{\alpha,J}^{(d)}(\tau) = \mu_\alpha^{(0)} + \int_0^\tau [\widetilde{\mathcal{V}}_{\mathrm{raw},J}^{(d)}(s)]_\alpha \, ds + E_\alpha^{(d)}(\tau), \qquad 0 \le \tau \le T, \tag{729}$$

and $\sup_{\tau\in[0,T]} |E_\alpha^{(d)}(\tau)| \xrightarrow{\mathbb{P}} 0$.

Proof. Fix a word

$$\alpha = (Z_1, \dots, Z_w).$$

Throughout the proof all estimates are uniform on (528). Recall from Lemma 18 that, on $\{n < J\}$,

$$\mu_\alpha^{(n+1)} - \mu_\alpha^{(n)} = \tau_d [\mathcal{V}_{\mathrm{raw}}^{(d,n)}]_\alpha + \alpha_d \eta_\alpha^{(n)} + R_\alpha^{(n)}, \tag{730}$$

where

$$\mathbb{E}_n[\eta_\alpha^{(n)}] = 0, \qquad \mathbb{E}_n|\eta_\alpha^{(n)}|^2 \le \frac{C_\alpha}{d}, \tag{731}$$

and

$$\mathbb{E}_n|R_\alpha^{(n)}| \le C_\alpha \left[(1+\gamma)^2 \alpha_d^2 + \frac{\alpha_d}{d^{3/2}}\right]. \tag{732}$$

For $m \le N_T$, the stopped process satisfies

$$\mu_\alpha^{(m\wedge J)} - \mu_\alpha^{(0)} = \sum_{n=0}^{m-1} \mathbf{1}_{\{n<J\}} \left(\mu_\alpha^{(n+1)} - \mu_\alpha^{(n)}\right). \tag{733}$$

Substituting (730) into (733) gives

$$\mu_\alpha^{(m\wedge J)} = \mu_\alpha^{(0)} + \tau_d \sum_{n=0}^{m-1} \mathbf{1}_{\{n<J\}} [\mathcal{V}_{\mathrm{raw}}^{(d,n)}]_\alpha + \sum_{n=0}^{m-1} \alpha_d \mathbf{1}_{\{n<J\}} \eta_\alpha^{(n)} + \sum_{n=0}^{m-1} \mathbf{1}_{\{n<J\}} R_\alpha^{(n)}. \tag{734}$$

Accordingly, define $E_{\mathrm{noise},\alpha}^{(d)}(\tau) := \sum_{n=0}^{m_d(\tau)-1} \alpha_d \mathbf{1}_{\{n<J\}} \eta_\alpha^{(n)}$ and $E_{\mathrm{rem},\alpha}^{(d)}(\tau) := \sum_{n=0}^{m_d(\tau)-1} \mathbf{1}_{\{n<J\}} R_\alpha^{(n)}$.

Since J is a stopping time, $\mathbf{1}_{\{n<J\}}$ is $\mathcal{F}_n$-measurable. Hence $\alpha_d \mathbf{1}_{\{n<J\}} \eta_\alpha^{(n)}$ is a martingale difference sequence. Let

$$M_{\alpha,m}^{(d)} := \sum_{n=0}^{m-1} \alpha_d \mathbf{1}_{\{n<J\}} \eta_\alpha^{(n)}, \qquad 0 \le m \le N_T. \tag{735}$$

Using (731), its predictable quadratic variation satisfies

$$\mathbb{E}\left[\left\langle M_\alpha^{(d)} \right\rangle_{N_T}\right] \le \sum_{n=0}^{N_T-1} \alpha_d^2 \mathbb{E}\left[\mathbf{1}_{\{n<J\}} \mathbb{E}_n|\eta_\alpha^{(n)}|^2\right] \le N_T C_\alpha \frac{\alpha_d^2}{d} \le C_{\alpha,T}\alpha_d. \tag{736}$$

Since $\alpha_d \to 0$, Doob's maximal inequality gives, for every $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_{\tau\in[0,T]} |E_{\mathrm{noise},\alpha}^{(d)}(\tau)| > \varepsilon\right) \le \mathbb{P}\left(\max_{m\le N_T} |M_{\alpha,m}^{(d)}| > \varepsilon\right) \le \frac{C_{\alpha,T}\alpha_d}{\varepsilon^2} \longrightarrow 0. \tag{737}$$

Therefore

$$\sup_{\tau\in[0,T]} |E_{\mathrm{noise},\alpha}^{(d)}(\tau)| \xrightarrow{\mathbb{P}} 0. \tag{738}$$

By (732),

$$\mathbb{E}\left[\sum_{n=0}^{N_T-1} \mathbf{1}_{\{n<J\}} |R_\alpha^{(n)}|\right] \le N_T C_\alpha \left[(1+\gamma)^2\alpha_d^2 + \frac{\alpha_d}{d^{3/2}}\right] \le C_{\alpha,T}\left[(1+\gamma)^2 d\alpha_d + d^{-1/2}\right]. \tag{739}$$

For the fixed time horizon T, γ is independent of d. Moreover, $d\alpha_d \le \frac{c_0}{\log d} \longrightarrow 0$ under the learning-rate assumption. Hence the right-hand side of (739) tends to zero.

Since $\sup_{\tau\in[0,T]} |E^{(d)}_{\mathrm{rem},\alpha}(\tau)| \le \sum_{n=0}^{N_T-1} \mathbf{1}_{\{n<J\}} |R^{(n)}_\alpha|$, Markov's inequality yields

$$\sup_{\tau\in[0,T]} |E^{(d)}_{\mathrm{rem},\alpha}(\tau)| \xrightarrow{\mathbb{P}} 0. \tag{740}$$

For $\tau \in [0,T]$, write $m := m_d(\tau)$. By the definition (728),

$$\int_0^\tau [\widetilde{\mathcal{V}}^{(d)}_{\mathrm{raw},J}(s)]_\alpha \, ds = \tau_d \sum_{n=0}^{m-1} \mathbf{1}_{\{n<J\}} [\mathcal{V}^{(d,n)}_{\mathrm{raw}}]_\alpha + (\tau - m\tau_d)\mathbf{1}_{\{m<J\}} [\mathcal{V}^{(d,m)}_{\mathrm{raw}}]_\alpha. \tag{741}$$

We therefore define the discretization error

$$E^{(d)}_{\mathrm{disc},\alpha}(\tau) := -(\tau - m\tau_d)\mathbf{1}_{\{m<J\}} [\mathcal{V}^{(d,m)}_{\mathrm{raw}}]_\alpha. \tag{742}$$

By (528), all letters satisfy $\max_{Z\in\mathscr{Z}} \|Z\|_{\mathrm{op}} \le C_*^2$. Consequently, for the fixed α, T_i^{bias} and T_i^{diff} in (694) are uniformly bounded. In addition, Lemma 13 gives $\max_k |b_k^{(n)}| + \max_{k,l,r} |a_{kl}^{(r,d,n)}| \le C$ uniformly on $\{n < J\}$. Hence

$$\mathbf{1}_{\{n<J\}} \left| [\mathcal{V}^{(d,n)}_{\mathrm{raw}}]_\alpha \right| \le C_\alpha \left(\sqrt{d} + 1 + \gamma \right). \tag{743}$$

Since $0 \le \tau - m\tau_d < \tau_d = \frac{\alpha_d}{d}$, we obtain

$$\sup_{\tau\in[0,T]} |E^{(d)}_{\mathrm{disc},\alpha}(\tau)| \le \frac{\alpha_d}{d} C_\alpha (\sqrt{d} + 1 + \gamma) \le C_\alpha \left(\frac{\alpha_d}{\sqrt{d}} + \frac{(1+\gamma)\alpha_d}{d} \right) \longrightarrow 0. \tag{744}$$

Finally, combining (734) and (741) gives

$$\tilde{\mu}^{(d)}_{\alpha,J}(\tau) = \mu^{(0)}_\alpha + \int_0^\tau [\widetilde{\mathcal{V}}^{(d)}_{\mathrm{raw},J}(s)]_\alpha \, ds + E^{(d)}_{\mathrm{noise},\alpha}(\tau) + E^{(d)}_{\mathrm{rem},\alpha}(\tau) + E^{(d)}_{\mathrm{disc},\alpha}(\tau). \tag{745}$$

Defining $E^{(d)}_\alpha = E^{(d)}_{\mathrm{noise},\alpha} + E^{(d)}_{\mathrm{rem},\alpha} + E^{(d)}_{\mathrm{disc},\alpha}$ and using (738), (740), (744) finishes the proof. □

Step 3: Effective bias of the fast flow For $k \in \mathcal{K}$, define $D_k^{(n)} := t_k^{A,(n)} + t_k^{B,(n)}$ and

$$\psi_k^{(d,n)} := 2 \sum_{l=1}^{K} \sum_{r=1}^{2} a_{kl}^{(r,d,n)} \left[\mu^{(n)}_{(M_l^{(r)},B_k)} + \mu^{(n)}_{(A_k,M_l^{(r)})} \right]. \tag{746}$$

Define the finite-dimensional effective bias by

$$\beta_k^{(d,n)} := -\frac{\psi_k^{(d,n)}}{D_k^{(n)}} \tag{747}$$

whenever $D_k^{(n)} > 0$, and set $\beta_k^{(d,n)} := 0$ when $D_k^{(n)} = 0$. The following lemma proves its convergence.

Lemma 20. *Under the conditions of Theorem 3, let J be a stopping time satisfying* (528)*, and fix $T > 0$. Set $\tau_d := \frac{\alpha_d}{d}$ and $N_T := \left\lfloor \frac{Td}{\alpha_d} \right\rfloor$. Then,*

$$\sup_{0\le\tau\le T} \left| \tau_d \sum_{n=0}^{m_d(\tau)-1} \mathbf{1}_{\{n<J\}} \sum_{\substack{i=1 \\ k_i\in\mathcal{K}}}^{w} \left(\sqrt{d}\, b_{k_i}^{(n)} - \beta_{k_i}^{(d,n)} \right) \mathcal{T}_i^{\mathrm{bias}}(\mu^{(n)}) \right| \xrightarrow{\mathbb{P}} 0, \tag{748}$$

where $m_d(\tau) := \left\lfloor \frac{\tau d}{\alpha_d} \right\rfloor$.

Proof. Fix a trainable position i and write $k := k_i$, $H_n := H_{\alpha,i}^{(n)} := \mathcal{T}_i^{\mathrm{bias}}(\mu^{(n)})$, $D_n := D_k^{(n)}$, $b_n := b_k^{(n)}$ and $\beta_n := \beta_k^{(d,n)}$. All constants below may depend on the fixed word α, but are independent of d and n.

Returning to the proof of Lemma 14, the expansion (545) gives

$$\mathbb{E}_n[\Delta m_k^{(n)}] = -\alpha_d \mathcal{D}_n b_n - \frac{\alpha_d}{\sqrt{d}} \psi_k^{(d,n)} + r_k^{(n)}, \tag{749}$$

where, uniformly on $\{n < J\}$,

$$|r_k^{(n)}| \le C \left(\frac{(1+\gamma)\alpha_d}{d} + \alpha_d^2 \right). \tag{750}$$

Using $\psi_k^{(d,n)} = -D_n \beta_n$, we may rewrite (749) as

$$\mathbb{E}_n[\Delta m_k^{(n)}] = -\frac{\alpha_d \mathcal{D}_n}{\sqrt{d}} \left(\sqrt{d}\, b_n - \beta_n \right) + r_k^{(n)}. \tag{751}$$

Let $\zeta_{m,k}^{(n+1)} := \Delta m_k^{(n)} - \mathbb{E}_n[\Delta m_k^{(n)}]$. By Lemma 14,

$$\mathbb{E}_n \left[|\zeta_{m,k}^{(n+1)}|^2 \right] \le C \alpha_d^2. \tag{752}$$

Therefore (751) gives

$$\tau_d H_n \left(\sqrt{d}\, b_n - \beta_n \right) = -\frac{1}{\sqrt{d}} \frac{H_n}{D_n} \Delta m_k^{(n)} + \frac{1}{\sqrt{d}} \frac{H_n}{D_n} \zeta_{m,k}^{(n+1)} + \frac{1}{\sqrt{d}} \frac{H_n}{D_n} r_k^{(n)}. \tag{753}$$

Thus it remains to show that the three terms on the right vanish after summation.

By Lemma 12, there exists $c_D = c_D(T, \gamma, c_{\mathrm{bal}}) > 0$ such that

$$\mathbb{P} \left(\inf_{\substack{0 \le n < N_T \wedge J \\ k \in \mathcal{K}}} D_k^{(n)} \ge c_D \right) \longrightarrow 1. \tag{754}$$

Moreover, Lemma 16 gives constants $C_{\mathrm{tr}}, c_{\mathrm{tr}}, C_b > 0$ such that

$$\mathbb{P} \left(\sup_{0 \le n < N_T \wedge J} \left[\|b_{\mathcal{K}}^{(n)}\| - C_{\mathrm{tr}} e^{-c_{\mathrm{tr}} \alpha_d n} \right]_+ > \frac{C_b}{\sqrt{d}} \right) \longrightarrow 0. \tag{755}$$

Now define

$$J_D := \inf \left\{ n \ge 0 : \min_{k \in \mathcal{K}} D_k^{(n)} < c_D \right\}, \tag{756}$$

$$J_b := \inf \left\{ n \ge 0 : \|b_{\mathcal{K}}^{(n)}\| > C_{\mathrm{tr}} e^{-c_{\mathrm{tr}} \alpha_d n} + \frac{2C_b}{\sqrt{d}} \right\}, \tag{757}$$

and set $\widehat{J} := J \wedge J_D \wedge J_b$. Equations (754) and (755) imply

$$\mathbb{P} \left(\widehat{J} < N_T \wedge J \right) \longrightarrow 0. \tag{758}$$

It therefore suffices to prove the desired convergence with J replaced by $\widehat{J}$. Before $\widehat{J}$, $D_n \ge c_D$, $|b_n| \le C_{\mathrm{tr}} e^{-c_{\mathrm{tr}} \alpha_d n} + \frac{2C_b}{\sqrt{d}}$, and also $|H_n| \le C_\alpha$, $|m_k^{(n)}| \le C$ by (528).

Define $C_n := \frac{H_n}{D_n}$ satisfying $|C_n| \le C_\alpha$ for $n < \widehat{J}$. To control the second term of (753), consider

$$M_m := \frac{1}{\sqrt{d}} \sum_{n=0}^{m-1} \mathbf{1}_{\{n < \widehat{J}\}} C_n \zeta_{m,k}^{(n+1)}. \tag{759}$$

Since $C_n \mathbf{1}_{\{n < \widehat{J}\}}$ is $\mathcal{F}_n$-measurable, M_m is a martingale. By (752),

$$\mathbb{E}[\langle M \rangle_{N_T}] \le \frac{C_\alpha}{d} N_T \alpha_d^2 \le C_{\alpha,T} \alpha_d. \tag{760}$$

Hence Doob's maximal inequality yields

$$\max_{m \le N_T} |M_m| \xrightarrow{\mathbb{P}} 0. \tag{761}$$

For the third term of (753), using (750),

$$\frac{1}{\sqrt{d}} \sum_{n=0}^{N_T-1} \mathbf{1}_{\{n<\widehat{J}\}} |C_n r_k^{(n)}| \le \frac{C_\alpha}{\sqrt{d}} N_T \left(\frac{(1+\gamma)\alpha_d}{d} + \alpha_d^2 \right) \le C_{\alpha,T} \left(\frac{1+\gamma}{\sqrt{d}} + \sqrt{d}\, \alpha_d \right) \longrightarrow 0, \tag{762}$$

because $\alpha_d \le c_0 (d \log d)^{-1}$.

For the following we control the first term of (753). Since H_n is a finite linear combination of moments of the same degree, Lemma 18 applies to H_n. The same lemma applies to $D_n = t_k^{A,(n)} + t_k^{B,(n)}$. Using the smooth map $(H, D) \longmapsto \frac{H}{D}$ on the region $|H| \le C_\alpha$, $D \ge c_D$, together with a first-order Taylor expansion, we obtain, for $n < \widehat{J}$,

$$C_{n+1} - C_n = A_C^{(n)} + \zeta_C^{(n+1)} + \rho_C^{(n)}, \tag{763}$$

where $A_C^{(n)}$ is $\mathcal{F}_n$-measurable, $\zeta_C^{(n+1)}$ is a martingale difference, and

$$|A_C^{(n)}| \le C_\alpha \left[\frac{\alpha_d}{\sqrt{d}} \|b_{\mathcal{K}}^{(n)}\| + \frac{(1+\gamma)\alpha_d}{d} + (1+\gamma)^2 \alpha_d^2 \right], \tag{764}$$

$$\mathbb{E}_n \left[|\zeta_C^{(n+1)}|^2 \right] \le C_\alpha \frac{\alpha_d^2}{d}, \tag{765}$$

$$\mathbb{E}_n \left[|\rho_C^{(n)}| \right] \le C_\alpha \left[(1+\gamma)^2 \alpha_d^2 + \frac{\alpha_d}{d^{3/2}} \right]. \tag{766}$$

The first term on the right-hand side of (764) comes from (694), while (765) and (766) follow from (697) and (698), together with the quadratic Taylor remainder of H/D.

For $n < J_b$ we have

$$\sum_{n=0}^{N_T-1} \mathbf{1}_{\{n<\widehat{J}\}} |A_C^{(n)}| \le C_\alpha \frac{\alpha_d}{\sqrt{d}} \sum_{n \ge 0} \left(C_{\mathrm{tr}} e^{-c_{\mathrm{tr}} \alpha_d n} + \frac{2C_b}{\sqrt{d}} \right) + C_{\alpha,T} \left(1 + \gamma + d\alpha_d\right) = \mathcal{O}(1), \tag{767}$$

uniformly in d. Here we use $\alpha_d \sum_{n \ge 0} e^{-c_{\mathrm{tr}} \alpha_d n} \le C$ and $N_T \frac{\alpha_d}{d} = \mathcal{O}(1)$, $N_T \alpha_d^2 = \mathcal{O}(d\alpha_d) = o(1)$. Moreover, (766) gives

$$\sum_{n=0}^{N_T-1} \mathbf{1}_{\{n<\widehat{J}\}} |\rho_C^{(n)}| \xrightarrow{\mathbb{P}} 0, \tag{768}$$

For every $m \le N_T$,

$$\sum_{n=0}^{m-1} \mathbf{1}_{\{n<\widehat{J}\}} C_n \Delta m_k^{(n)} = C_{m_*-1} m_k^{(m_*)} - C_0 m_k^{(0)} - \sum_{n=1}^{m_*-1} m_k^{(n)} \left(C_n - C_{n-1} \right), \tag{769}$$

where $m_* := m \wedge \widehat{J}$ with the first term interpreted as zero when $m_* = 0$.

The boundary terms in (769) are uniformly $\mathcal{O}(1)$, and hence become $o(1)$ after multiplication by $d^{-1/2}$.

For the $C_n - C_{n-1}$ term, using the boundedness of $m_k^{(n)}$ and (767),

$$\frac{1}{\sqrt{d}} \sup_{m \le N_T} \left| \sum_{n=1}^{m_*-1} m_k^{(n)} A_C^{(n-1)} \right| \longrightarrow 0. \tag{770}$$

The contribution of $\rho_C^{(n)}$ vanishes by (768).

It remains to treat the martingale part (the second term of (763)). Write $m_k^{(n)} = m_k^{(n-1)} + \Delta m_k^{(n-1)}$. Then

$$\sum_{n=1}^{m_*-1} m_k^{(n)} \zeta_C^{(n)} = \sum_{n=1}^{m_*-1} m_k^{(n-1)} \zeta_C^{(n)} + \sum_{n=1}^{m_*-1} \Delta m_k^{(n-1)} \zeta_C^{(n)}. \tag{771}$$

The first term of (771) is a martingale with predictable quadratic variation bounded by $C_{\alpha,T}\alpha_d$, because $m_k^{(n-1)}$ is uniformly bounded. Hence it is upper bounded by $O_{\mathbb{P}}(\sqrt{\alpha_d})$. For the second term of (771), Lemma 14 gives $\mathbb{E}_{n-1}\left[|\Delta m_k^{(n-1)}|^2\right] \le C\alpha_d^2$ before the stopping time. Together with (765) and Cauchy–Schwarz,

$$\mathbb{E}\left[\sum_{n=1}^{N_T\wedge\widehat{J}-1} \left|\Delta m_k^{(n-1)} \zeta_C^{(n)}\right|\right] \le C N_T \frac{\alpha_d^2}{\sqrt{d}} \le C_T \alpha_d \sqrt{d}. \tag{772}$$

Therefore

$$\frac{1}{\sqrt{d}} \sum_{n=1}^{N_T\wedge\widehat{J}-1} \left|\Delta m_k^{(n-1)} \zeta_C^{(n)}\right| \xrightarrow{\mathbb{P}} 0, \tag{773}$$

since $\alpha_d \to 0$. Combining these estimates with (769), we obtain

$$\sup_{m\le N_T} \left| \frac{1}{\sqrt{d}} \sum_{n=0}^{m-1} \mathbf{1}_{\{n<\widehat{J}\}} C_n \Delta m_k^{(n)} \right| \xrightarrow{\mathbb{P}} 0. \tag{774}$$

Summing (753) from $n=0$ to $m_d(\tau)-1$, and using (761), (762), and (774), gives

$$\sup_{0\le\tau\le T} \left| \tau_d \sum_{n=0}^{m_d(\tau)-1} \mathbf{1}_{\{n<\widehat{J}\}} H_n\left(\sqrt{d}\, b_n - \beta_n\right) \right| \xrightarrow{\mathbb{P}} 0. \tag{775}$$

Finally, (758) allows us to replace $\widehat{J}$ by J. Summing over the trainable positions proves (748). □

Step 4: Final results Finally we can write an approximate integral equation for the moments. For a word $\alpha = (Z_1, \dots, Z_w) \in \mathscr{Z}^w$, define the finite-dimensional effective vector field by

$$[\mathcal{V}_{\text{eff}}^{(d,n)}]_\alpha := -\sum_{\substack{i=1\\ k_i\in\mathcal{K}}}^{w} \beta_{k_i}^{(d,n)} \mathcal{T}_i^{\text{bias}}(\mu^{(n)}) - 2\sum_{\substack{i=1\\ k_i\in\mathcal{K}}}^{w} \sum_{l=1}^{K} \sum_{r=1}^{2} a_{k_i l}^{(r,d,n)} \mathcal{T}_{i,l,r}^{\text{diff}}(\mu^{(n)}) - 2\gamma|\alpha|_{\mathcal{K}} \mu_\alpha^{(n)}. \tag{776}$$

Let

$$[\widetilde{\mathcal{V}}_{\text{eff},J}^{(d)}(\tau)]_\alpha := \mathbf{1}_{\{m_d(\tau)<J\}} [\mathcal{V}_{\text{eff}}^{(d,m_d(\tau))}]_\alpha \tag{777}$$

and $m_d(\tau) := \left\lfloor \frac{\tau d}{\alpha_d} \right\rfloor$.

Lemma 21. *Under the conditions of Theorem 3, let J be a stopping time satisfying* (528)*, and fix $T>0$. Set $\tau_d := \frac{\alpha_d}{d}$ and $N_T := \left\lfloor \frac{Td}{\alpha_d} \right\rfloor$. Then, for every α, there exists $\mathcal{E}_\alpha^{(d)} : [0,T] \to \mathbb{R}$ such that*

$$\tilde{\mu}_{\alpha,J}^{(d)}(\tau) = \mu_\alpha^{(0)} + \int_0^\tau [\widetilde{\mathcal{V}}_{\text{eff},J}^{(d)}(s)]_\alpha \, ds + \mathcal{E}_\alpha^{(d)}(\tau), \qquad 0 \le \tau \le T, \tag{778}$$

and $\sup_{\tau\in[0,T]} |\mathcal{E}_\alpha^{(d)}(\tau)| \xrightarrow{\mathbb{P}} 0$.

Proof. By Lemma 19, for every fixed α,

$$\tilde{\mu}_{\alpha,J}^{(d)}(\tau) = \mu_\alpha^{(0)} + \int_0^\tau [\widetilde{\mathcal{V}}_{\text{raw},J}^{(d)}(s)]_\alpha \, ds + E_\alpha^{(d)}(\tau), \tag{779}$$

where $\sup_{\tau\in[0,T]} |E_\alpha^{(d)}(\tau)| \xrightarrow{\mathbb{P}} 0$. It therefore suffices to prove

$$\sup_{\tau\in[0,T]} \left| \int_0^\tau \left([\widetilde{\mathcal{V}}_{\text{raw},J}^{(d)}(s)]_\alpha - [\widetilde{\mathcal{V}}_{\text{eff},J}^{(d)}(s)]_\alpha \right) ds \right| \xrightarrow{\mathbb{P}} 0. \tag{780}$$

By the definitions (694) and (776), the two vector fields differ only in the bias term. Thus

$$[\mathcal{V}_{\mathrm{raw}}^{(d,n)}]_\alpha - [\mathcal{V}_{\mathrm{eff}}^{(d,n)}]_\alpha = -\sum_{\substack{i=1\\ k_i\in\mathcal{K}}}^{w} \left(\sqrt{d}\, b_{k_i}^{(n)} - \beta_{k_i}^{(d,n)}\right) \mathcal{T}_i^{\mathrm{bias}}(\mu^{(n)}). \tag{781}$$

Fix $\tau \in [0,T]$ and write $m := m_d(\tau) = \left\lfloor \frac{\tau}{\tau_d} \right\rfloor$. Since both vector fields are piecewise constant,

$$\int_0^\tau \left([\widetilde{\mathcal{V}}_{\mathrm{raw},J}^{(d)}(s)]_\alpha - [\widetilde{\mathcal{V}}_{\mathrm{eff},J}^{(d)}(s)]_\alpha\right) ds = -\tau_d \sum_{n=0}^{m-1} \mathbf{1}_{\{n<J\}} \sum_{\substack{i=1\\ k_i\in\mathcal{K}}}^{w} \left(\sqrt{d}\, b_{k_i}^{(n)} - \beta_{k_i}^{(d,n)}\right) \mathcal{T}_i^{\mathrm{bias}}(\mu^{(n)}) + R_{\mathrm{cell},\alpha}^{(d)}(\tau), \tag{782}$$

where the final term is

$$R_{\mathrm{cell},\alpha}^{(d)}(\tau) := -(\tau - m\tau_d)\mathbf{1}_{\{m<J\}} \sum_{\substack{i=1\\ k_i\in\mathcal{K}}}^{w} \left(\sqrt{d}\, b_{k_i}^{(m)} - \beta_{k_i}^{(d,m)}\right) \mathcal{T}_i^{\mathrm{bias}}(\mu^{(m)}). \tag{783}$$

The first term in (782) converges to zero uniformly in τ in probability by Lemma 20. Hence it remains only to control $R_{\mathrm{cell},\alpha}^{(d)}$.

Before the stopping time (528), all letters satisfy a uniform operator-norm bound. Therefore, for the fixed word α, $\left|\mathcal{T}_i^{\mathrm{bias}}(\mu^{(n)})\right| \le C_\alpha$ uniformly for $n < J$. Moreover, Lemma 13 gives $|b_k^{(n)}| \le C$ uniformly on $\{n < J\}$. We also have a uniform bound on $\beta_k^{(d,n)}$. Indeed, since $A_k, B_k \succeq 0$,

$$\left|\mu_{(M_l^{(r)},B_k)}^{(n)}\right| \le \|M_l^{(r,n)}\|_{\mathrm{op}}\, t_k^{B,(n)}, \tag{784}$$

$$\left|\mu_{(A_k,M_l^{(r)})}^{(n)}\right| \le \|M_l^{(r,n)}\|_{\mathrm{op}}\, t_k^{A,(n)}. \tag{785}$$

The coefficients $a_{kl}^{(r,d,n)}$ are uniformly bounded on (528), and hence

$$|\psi_k^{(d,n)}| \le C\left(t_k^{A,(n)} + t_k^{B,(n)}\right) = CD_k^{(n)}. \tag{786}$$

Consequently, $|\beta_k^{(d,n)}| \le C$ whenever $D_k^{(n)} > 0$, while the same bound is trivial under our convention $\beta_k^{(d,n)} = 0$ when $D_k^{(n)} = 0$.

Since $0 \le \tau - m\tau_d < \tau_d = \frac{\alpha_d}{d}$, the bounds above give

$$\sup_{\tau\in[0,T]} |R_{\mathrm{cell},\alpha}^{(d)}(\tau)| \le C_\alpha \frac{\alpha_d}{d}(\sqrt{d}+1) \le C_\alpha\left(\frac{\alpha_d}{\sqrt{d}} + \frac{\alpha_d}{d}\right) \longrightarrow 0. \tag{787}$$

Combining this with Lemma 20 proves (780).

Finally, define

$$\mathcal{E}_\alpha^{(d)}(\tau) := E_\alpha^{(d)}(\tau) + \int_0^\tau \left([\widetilde{\mathcal{V}}_{\mathrm{raw},J}^{(d)}(s)]_\alpha - [\widetilde{\mathcal{V}}_{\mathrm{eff},J}^{(d)}(s)]_\alpha\right) ds. \tag{788}$$

Then we finish the proof by (779) and (780). □

F.2.6 Limiting slow dynamics

We now pass from the finite-dimensional effective dynamics (778) to the limiting dynamics. We first establish compactness of the trajectories.

Step 1: Compactness Fix a compact set $\mathfrak{D} \Subset \mathcal{M} \times \mathcal{Q}$, and recall $J_{\mathrm{mac}} := \inf\left\{n \geq 0 : \left(m_{\mathcal{K}}^{(n)}, q^{(1,n)}, q^{(2,n)}\right) \notin \mathfrak{D}\right\}$. Let C_* be the constant in Lemma 17, and define $J_{\mathrm{sp}} := \inf\left\{n \geq 0 : \max_{1\leq k\leq K}\left\{\|U_k^{(n)}\|_{\mathrm{op}}, \|V_k^{(n)}\|_{\mathrm{op}}\right\} > C_*\right\}$. For $N_T := \left\lfloor \frac{Td}{\alpha_d} \right\rfloor$, set

$$\mathcal{E}_{\mathrm{sp}}^{(d)}(T) := \{J_{\mathrm{sp}} > J_{\mathrm{mac}} \wedge N_T\}. \tag{789}$$

By Lemma 17,

$$\mathbb{P}\left(\mathcal{E}_{\mathrm{sp}}^{(d)}(T)\right) \longrightarrow 1. \tag{790}$$

For $0 \leq n \leq N_T$, define the auxiliary weights

$$\check{U}_k^{(n)} := \begin{cases} U_k^{(n\wedge J_{\mathrm{mac}})}, & \text{on } \mathcal{E}_{\mathrm{sp}}^{(d)}(T), \\ U_k^{(0)}, & \text{on } \mathcal{E}_{\mathrm{sp}}^{(d)}(T)^c, \end{cases} \qquad \check{V}_k^{(n)} := \begin{cases} V_k^{(n\wedge J_{\mathrm{mac}})}, & \text{on } \mathcal{E}_{\mathrm{sp}}^{(d)}(T), \\ V_k^{(0)}, & \text{on } \mathcal{E}_{\mathrm{sp}}^{(d)}(T)^c. \end{cases} \tag{791}$$

Let $\check{M}_k^{(n)} = \check{U}_k^{(n)}(\check{V}_k^{(n)})^T$, $\check{A}_k^{(n)} = \check{U}_k^{(n)}(\check{U}_k^{(n)})^T$, $\check{B}_k^{(n)} = \check{V}_k^{(n)}(\check{V}_k^{(n)})^T$, and define the corresponding auxiliary moments $\check{\mu}_\alpha^{(n)} := \frac{1}{d}\mathrm{Tr}\left[\check{Z}_1^{(n)} \cdots \check{Z}_w^{(n)}\right]$, $\alpha = (Z_1, \ldots, Z_w) \in \mathscr{Z}^w$. Their interpolation is denoted by

$$\check{\mu}^{(d)}(\tau) = \check{\mu}^{(n)}, \qquad \tau \in \left[\frac{n\alpha_d}{d}, \frac{(n+1)\alpha_d}{d}\right). \tag{792}$$

Lemma 22. *Under the conditions of Theorem 3, for every fixed $T > 0$ and every $\rho > C_*^2$, the sequence $\left\{\check{\mu}^{(d)}(\cdot)\right\}_{d\geq 1}$ is C-tight in $D([0,T]; X_\rho^0)$. Moreover, for every word α of length w, $\sup_{\tau\in[0,T]} |\check{\mu}_\alpha^{(d)}(\tau)| \leq C_*^{2w}$ almost surely, and the auxiliary process agrees with the empirical process stopped at J_{mac} with probability tending to one.*

Proof. By construction and (789), almost surely

$$\sup_{0\leq n\leq N_T} \max_{1\leq k\leq K}\left\{\|\check{U}_k^{(n)}\|_{\mathrm{op}}, \|\check{V}_k^{(n)}\|_{\mathrm{op}}\right\} \leq C_*. \tag{793}$$

Hence every letter $\check{Z}_k^{(n)} \in \left\{\check{M}_k^{(n)}, (\check{M}_k^{(n)})^T, \check{A}_k^{(n)}, \check{B}_k^{(n)}\right\}$ satisfies $\|\check{Z}_k^{(n)}\|_{\mathrm{op}} \leq C_*^2$. Therefore, for every word $|\alpha| = w$, $|\check{\mu}_\alpha^{(d)}(\tau)| \leq C_*^{2w}$. In particular, $\check{\mu}^{(d)}(\tau) \in X_\rho^0$ for $0 \leq \tau \leq T$ and every $\rho > C_*^2$.

We next record a uniform bound on the finite-d effective vector field. Before the stopping time, Lemma 13 gives $\max_{k,l,r} |a_{kl}^{(r,d,n)}| \leq C$. Moreover, $|\psi_k^{(d,n)}| \leq C\left(t_k^{A,(n)} + t_k^{B,(n)}\right)$, and hence, under the convention $\beta_k^{(d,n)} = 0$ when $t_k^{A,(n)} + t_k^{B,(n)} = 0$, $|\beta_k^{(d,n)}| \leq C$. Consequently, there exists a constant $B < \infty$, independent of d, n, w, such that

$$\mathbf{1}_{\{n<J_{\mathrm{sp}}\wedge J_{\mathrm{mac}}\}} \sup_{|\alpha|=w} \left|[\mathcal{V}_{\mathrm{eff}}^{(d,n)}]_\alpha\right| \leq BwC_*^{2w}. \tag{794}$$

We now use Lemma 21. Set $J := J_{\mathrm{sp}} \wedge J_{\mathrm{mac}}$. On $\mathcal{E}_{\mathrm{sp}}^{(d)}(T)$, stopping at J agrees with stopping at J_{mac} throughout the time interval $[0,T]$. Therefore, for every fixed word α, the auxiliary process admits the representation

$$\check{\mu}_\alpha^{(d)}(\tau) = \check{\mu}_\alpha^{(d)}(0) + \mathbf{1}_{\mathcal{E}_{\mathrm{sp}}^{(d)}(T)} \int_0^\tau [\widetilde{\mathcal{V}}_{\mathrm{eff,J}}^{(d)}(s)]_\alpha \, ds + \check{\mathcal{E}}_\alpha^{(d)}(\tau), \tag{795}$$

where

$$\sup_{\tau\in[0,T]} |\check{\mathcal{E}}_\alpha^{(d)}(\tau)| \xrightarrow{\mathbb{P}} 0 \tag{796}$$

for every fixed α. On $\mathcal{E}_{\mathrm{sp}}^{(d)}(T)^c$, we set $\check{\mathcal{E}}_\alpha^{(d)} \equiv 0$.

Define the continuous approximation

$$\hat{\mu}^{(d)}(\tau) := \check{\mu}^{(d)}(0) + \mathbf{1}_{\mathcal{E}_{\mathrm{sp}}^{(d)}(T)} \int_0^\tau \widetilde{\mathcal{V}}_{\mathrm{eff,J}}^{(d)}(s) \, ds. \tag{797}$$

The convergence in (796) can be upgraded to convergence in X_ρ^0, exactly as in the tied case (Lemma 4):

$$\sup_{\tau\in[0,T]} \left\| \check{\mu}^{(d)}(\tau) - \hat{\mu}^{(d)}(\tau) \right\|_\rho \xrightarrow{\mathbb{P}} 0. \tag{798}$$

It remains to prove compactness of $\{\hat{\mu}^{(d)}\}$. By (794), for every $|\alpha| = w$,

$$\sup_{\tau\in[0,T]} |\hat{\mu}_\alpha^{(d)}(\tau)| \le (1 + TBw) C_*^{2w}. \tag{799}$$

Furthermore, for $0 \le s < t \le T$,

$$\left\| \hat{\mu}^{(d)}(t) - \hat{\mu}^{(d)}(s) \right\|_\rho \le B|t-s| \sup_{w\ge 1} w \left(\frac{C_*^2}{\rho}\right)^w =: L_\rho |t-s|, \tag{800}$$

where $L_\rho < \infty$ is independent of d. The remainder of the argument is identical to the proof of Lemma 5 in the tied case. Namely, (799) defines a compact subset of X_ρ^0, because its tail decays exponentially, while (800) gives uniform equicontinuity. The Arzelà–Ascoli theorem therefore implies that $\{\hat{\mu}^{(d)}\}_{d\ge1}$ is tight in $C([0,T]; X_\rho^0)$. Together with (798), this proves that $\{\check{\mu}^{(d)}\}_{d\ge1}$ is C-tight in $D([0,T]; X_\rho^0)$.

Finally, on $\mathcal{E}_{\mathrm{sp}}^{(d)}(T)$, the auxiliary process is precisely the empirical moment process stopped at J_{mac}. Since $\mathbb{P}(\mathcal{E}_{\mathrm{sp}}^{(d)}(T)) \to 1$, the two processes are asymptotically equivalent on $[0,T]$. □

Step 2: Convergence of the effective vector field

Lemma 23. *Under the conditions of Theorem 3, fix $0 < T_1 < T_2 \le T$, and let J be a stopping time satisfying* (528)*. Let C_* denote the corresponding spectral bound in Lemma 17 and fix $\rho > C_*^2$. Let $\mathcal{V}_{\mathrm{eff}}^{(d,n)}$ be the finite-d effective vector field defined in* (776)*, and let V_{ext} be the extended limiting vector field defined in* (505)*. Then,*

$$\sup_{\tau\in[T_1,T_2]} \mathbf{1}_{\{m_d(\tau)<J\}} \left\| \mathcal{V}_{\mathrm{eff}}^{(d,m_d(\tau))} - V_{\mathrm{ext}}(\mu^{(m_d(\tau))}) \right\|_\rho \xrightarrow{\mathbb{P}} 0. \tag{801}$$

Proof. Throughout the proof, all estimates are uniform on (528). Write $q^{(n)} := \left(q^{(1,n)}, q^{(2,n)}\right)$ and $m_\star^{(n)} := m^\star(q^{(n)})$, where the non-trainable components of $m_\star^{(n)}$ remain fixed at their initial values.

For notational convenience, define

$$A_{kl}^{(r)}(m,q) := \left[\nabla_{q^{(r)}} \Phi(m,q)\right]_{kl}, \qquad r = 1,2. \tag{802}$$

An empirical state $q^{(n)} \in \mathfrak{Q}$. This is because for arbitrary $c = (c_1, \dots, c_K)$, we can write $M_c := \sum_{k=1}^K c_k M_k$ and have

$$c^T(q^{(1)} + q^{(2)})c = \frac{1}{d}\left(\mathrm{Tr}[M_c M_c^T] + \mathrm{Tr}[M_c^2]\right) = \frac{2}{d}\|\operatorname{Sym} M_c\|_F^2 \ge 0, \tag{803}$$

Similarly

$$c^T(q^{(1)} - q^{(2)})c = \frac{2}{d}\|\operatorname{Skew} M_c\|_F^2 \ge 0, \tag{804}$$

where $\operatorname{Sym} M_c := \frac{M_c + M_c^T}{2}$ and $\operatorname{Skew} M_c := \frac{M_c - M_c^T}{2}$. Hence the projection in V_{ext} (see Appendix F.2.2) acts trivially along the empirical trajectory:

$$\widehat{q}(\mu^{(n)}) = q^{(n)}. \tag{805}$$

Recall

$$a_{kl}^{(r,d,n)} := \frac{1}{2} \sum_{i,j,i',j'=1}^{L} \mathcal{C}_{ij,i'j'}^{(r)} \mathbb{E}_n \left[\partial^2_{(ijk),(i'j'l)} \mathcal{L}(G^{(n)})\right]. \tag{806}$$

We first prove its convergence to $A_{kl}^{(r)}(m^{(n)}, q^{(n)})$.

Before the stopping time (528), $\max_k \|M_k^{(n)}\|_{\mathrm{op}} \le C_*^2$, while compactness of $\mathfrak{D}$ gives

$$\max_k |m_k^{(n)}| = \max_k \frac{1}{\sqrt{d}} |\mathrm{Tr} M_k^{(n)}| \le R_* \tag{807}$$

for a constant $R_* < \infty$. Therefore Lemma 10 applies uniformly in $n < J$.

For fixed indices $(ijk), (i'j'l)$, apply Lemma 10 to $\partial^2_{(ijk),(i'j'l)} \mathcal{L}(G)$. Assumption 9.1 guarantees the required regularity and polynomial growth. By Price's theorem,

$$A_{kl}^{(r)}(m^{(n)}, q^{(n)}) = \frac{1}{2} \sum_{i,j,i',j'} \mathcal{C}_{ij,i'j'}^{(r)} \mathbb{E}\left[\partial^2_{(ijk),(i'j'l)} \mathcal{L}(G_{\mathrm{G}}^{(n)})\right], \tag{808}$$

where $G_{\mathrm{G}}^{(n)}$ is Gaussian with the same mean and covariance as $G^{(n)}$. Consequently,

$$\sup_{n<J} \max_{k,l,r} \left| a_{kl}^{(r,d,n)} - A_{kl}^{(r)}(m^{(n)}, q^{(n)}) \right| \le \frac{C}{\sqrt{d}}. \tag{809}$$

We next replace $m^{(n)}$ by $m_\star^{(n)}$. By Lemma 13,

$$\left\| b_{\mathcal{K}}^{(n)} - \nabla_{m_{\mathcal{K}}} \Phi(m^{(n)}, q^{(n)}) \right\| \le \frac{C}{\sqrt{d}}. \tag{810}$$

Lemma 16 gives, with probability tending to one,

$$\|b_{\mathcal{K}}^{(n)}\| \le C_{\mathrm{tr}} e^{-c_{\mathrm{tr}} \alpha_d n} + \frac{C_b}{\sqrt{d}}, \qquad 0 \le n < N_T \wedge J. \tag{811}$$

Hence, uniformly for $n \ge \left\lceil \frac{T_1 d}{\alpha_d} \right\rceil$, we have $e^{-c_{\mathrm{tr}} \alpha_d n} \le e^{-c_{\mathrm{tr}} T_1 d}$, and therefore

$$\sup_{\substack{\lceil T_1 d/\alpha_d \rceil \le n \le \lfloor T_2 d/\alpha_d \rfloor \\ n<J}} \left\| \nabla_{m_{\mathcal{K}}} \Phi(m^{(n)}, q^{(n)}) \right\| = O_{\mathbb{P}}(d^{-1/2}). \tag{812}$$

Since $\nabla_{m_{\mathcal{K}}} \Phi(m_\star^{(n)}, q^{(n)}) = 0$, the uniform strong convexity in Assumption 10 implies

$$\lambda_0 \left\| m_{\mathcal{K}}^{(n)} - m_{\mathcal{K}}^\star(q^{(n)}) \right\| \le \left\| \nabla_{m_{\mathcal{K}}} \Phi(m^{(n)}, q^{(n)}) \right\|. \tag{813}$$

Thus

$$\sup_{\substack{\lceil T_1 d/\alpha_d \rceil \le n \le \lfloor T_2 d/\alpha_d \rfloor \\ n<J}} \left\| m^{(n)} - m_\star^{(n)} \right\| = O_{\mathbb{P}}(d^{-1/2}). \tag{814}$$

The map $(m, q) \longmapsto \nabla_{q^{(r)}} \Phi(m, q)$ is Lipschitz on a compact neighborhood of (528) and of $\{(m^\star(q), q) : q \in \mathrm{proj}_q \mathfrak{D}\}$. Therefore,

$$\sup_{\substack{\lceil T_1 d/\alpha_d \rceil \le n \le \lfloor T_2 d/\alpha_d \rfloor \\ n<J}} \max_{k,l,r} \left| A_{kl}^{(r)}(m^{(n)}, q^{(n)}) - A_{kl}^{(r)}(m_\star^{(n)}, q^{(n)}) \right| \xrightarrow{\mathbb{P}} 0. \tag{815}$$

Combining (809) and (815), we conclude that

$$\Delta_{a,d} := \sup_{\substack{\lceil T_1 d/\alpha_d \rceil \le n \le \lfloor T_2 d/\alpha_d \rfloor \\ n<J}} \max_{k,l,r} \left| a_{kl}^{(r,d,n)} - A_{kl}^{(r)}(m_\star^{(n)}, q^{(n)}) \right| \xrightarrow{\mathbb{P}} 0. \tag{816}$$

For $k \in \mathcal{K}$, recall $\psi_k^{(d,n)} = 2 \sum_{l=1}^K \sum_{r=1}^2 a_{kl}^{(r,d,n)} \left[\mu^{(n)}_{(M_l^{(r)}, B_k)} + \mu^{(n)}_{(A_k, M_l^{(r)})} \right]$ and $\beta_k^{(d,n)} = -\frac{\psi_k^{(d,n)}}{D_k^{(n)}}$, $D_k^{(n)} := t_k^{A,(n)} + t_k^{B,(n)}$. Now we prove the convergence of $\beta_k^{(d,n)}$ to β_k^{ext}

Along the empirical trajectory, (805) implies

$$\left| \psi_k^{(d,n)} - \psi_k^{\mathrm{ext}}(\mu^{(n)}) \right| \le 2\Delta_{a,d} \sum_{l=1}^K \sum_{r=1}^2 \left(\left| \mu^{(n)}_{(M_l^{(r)}, B_k)} \right| + \left| \mu^{(n)}_{(A_k, M_l^{(r)})} \right| \right). \tag{817}$$

The spectral bound (528) gives $\|M_l^{(r,n)}\|_{\mathrm{op}} \le C_*^2$. Since $A_k, B_k \succeq 0$, we have $\left|\mu^{(n)}_{(M_l^{(r)},B_k)}\right| \le C_*^2 t_k^{B,(n)}$ and $\left|\mu^{(n)}_{(A_k,M_l^{(r)})}\right| \le C_*^2 t_k^{A,(n)}$. Therefore

$$\left|\psi_k^{(d,n)} - \psi_k^{\mathrm{ext}}(\mu^{(n)})\right| \le C\Delta_{a,d} D_k^{(n)}. \tag{818}$$

Whenever $D_k^{(n)} > 0$, division by $D_k^{(n)}$ yields

$$\left|\beta_k^{(d,n)} - \beta_k^{\mathrm{ext}}(\mu^{(n)})\right| \le C\Delta_{a,d}. \tag{819}$$

By Lemma 12, $D_k^{(n)}$ is uniformly bounded away from zero before J, for all $k \in \mathcal{K}$, with probability tending to one. Hence (816) implies

$$\Delta_{\beta,d} := \sup_{\substack{\lceil T_1 d/\alpha_d\rceil \le n \le \lfloor T_2 d/\alpha_d\rfloor \\ n<J}} \max_{k\in\mathcal{K}} \left|\beta_k^{(d,n)} - \beta_k^{\mathrm{ext}}(\mu^{(n)})\right| \xrightarrow{\mathbb{P}} 0. \tag{820}$$

Finally, fix $n < J$ and a word $\alpha = (Z_1, \dots, Z_w)$. The weight-decay terms in $\mathcal{V}_{\mathrm{eff}}^{(d,n)}$ and $V_{\mathrm{ext}}(\mu^{(n)})$ coincide exactly. Thus their difference contains only $\mathcal{T}^{\mathrm{bias}}$ and $\mathcal{T}^{\mathrm{diff}}$ terms. Under the spectral bounds (528), for every trainable position i, $\left|\mathcal{T}_i^{\mathrm{bias}}(\mu^{(n)})\right| \le 2C_*^{2w}$ and $\left|\mathcal{T}_{i,l,r}^{\mathrm{diff}}(\mu^{(n)})\right| \le 2C_*^{2(w+1)}$. Consequently,

$$\left|[\mathcal{V}_{\mathrm{eff}}^{(d,n)} - V_{\mathrm{ext}}(\mu^{(n)})]_\alpha\right| \le Cw\left(\Delta_{\beta,d} C_*^{2w} + \Delta_{a,d} C_*^{2(w+1)}\right), \tag{821}$$

where C depends only on K.

Dividing by ρ^w and taking the supremum over all words of degree $w \ge 1$ gives

$$\left\|\mathcal{V}_{\mathrm{eff}}^{(d,n)} - V_{\mathrm{ext}}(\mu^{(n)})\right\|_\rho \le C\left(\Delta_{\beta,d} + C_*^2\Delta_{a,d}\right) \sup_{w\ge1} w\left(\frac{C_*^2}{\rho}\right)^w. \tag{822}$$

Because $\rho > C_*^2$ and $\sup_{w\ge1} w\left(\frac{C_*^2}{\rho}\right)^w < \infty$. Equations (816) and (820) therefore imply

$$\sup_{\substack{\lceil T_1 d/\alpha_d\rceil \le n \le \lfloor T_2 d/\alpha_d\rfloor \\ n<J}} \left\|\mathcal{V}_{\mathrm{eff}}^{(d,n)} - V_{\mathrm{ext}}(\mu^{(n)})\right\|_\rho \xrightarrow{\mathbb{P}} 0., \tag{823}$$

which proves (801). □

Step 3: Global well-posedness We call a pair $(\bar m, \bar\mu)$ a solution on $[0,T]$ if there exist $0 < s_- < s_+ < \infty$ such that $\bar\mu \in C([0,T]; X^0_{s_-})$,

$$\bar\mu(\tau) = \bar\mu(0) + \int_0^\tau V_{\mathrm{ext}}(\bar\mu(s))\, ds \tag{824}$$

holds as an identity in $X^0_{s_+}$, and, for every $\tau \in [0,T]$,

$$q(\bar\mu(\tau)) \in \mathfrak{Q}, \qquad \min_{k\in\mathcal{K}}\left(\bar t_k^A(\tau) + \bar t_k^B(\tau)\right) > 0, \tag{825}$$

with

$$\bar m_{\mathcal{K}}(\tau) = m^\star_{\mathcal{K}}(q(\bar\mu(\tau))), \qquad \bar m_{\setminus\mathcal{K}}(\tau) = \bar m_{\setminus\mathcal{K}}(0). \tag{826}$$

Lemma 24. *Under the conditions of Theorem 3, the extended slow system* (506)*, initialized at $\bar\mu(0)$, admits a unique solution $\bar\mu \in C([0,T]; X^0_{\rho_-})$ for every pair $C_*^2 < \rho_- < \rho_+$, where the integral equation* (824) *is understood in $X^0_{\rho_+}$. Moreover, $\bar\mu$ also satisfies the original ODE* (16)–(19)*.*

Proof. Fix $C_*^2 < \rho_- < \rho_+$. By Assumption 9.2, every initial letter satisfies $\|M_k^{(0)}\|_{\mathrm{op}}, \|A_k^{(0)}\|_{\mathrm{op}}, \|B_k^{(0)}\|_{\mathrm{op}} \le C_0^2$. Hence, for every word $|\alpha| = w$, $|\mu_\alpha^{(d)}(0)| \le C_0^{2w}$. Together with Assumption 9.3 and $\rho_- > C_0^2$, the same tail argument used in the tied case (Lemma 7) gives

$$\mu^{(d)}(0) \xrightarrow{\mathbb{P}} \bar{\mu}(0) \qquad \text{in } X^0_{\rho_-}. \tag{827}$$

By Lemma 22, the auxiliary processes $\{\check{\mu}^{(d)}\}$ are C-tight in $D([0,T]; X^0_{\rho_-})$. Hence, from every subsequence, we may extract a further subsequence, still indexed by d, such that

$$\check{\mu}^{(d)} \to \mu^* \qquad \text{in } D([0,T]; X^0_{\rho_-}), \tag{828}$$

where $\mu^* \in C([0,T]; X^0_{\rho_-})$ almost surely.

As in the proof of Lemma 7 for the tied model, by the Skorokhod representation theorem we may assume

$$\sup_{\tau\in[0,T]} \|\check{\mu}^{(d)}(\tau) - \mu^*(\tau)\|_{\rho_-} \longrightarrow 0 \qquad \text{almost surely.} \tag{829}$$

Fix $0 < \varepsilon < t \le T$ and first work before the stopping time (528). Lemma 21, together with Lemma 22, gives

$$\check{\mu}^{(d)}(t) - \check{\mu}^{(d)}(\varepsilon) = \int_\varepsilon^t \mathcal{V}_{\mathrm{eff}}^{(d)}(s)\, ds + R_d(\varepsilon, t), \tag{830}$$

where

$$\sup_{\varepsilon\le s\le t} \|R_d(\varepsilon, s)\|_{\rho_+} \xrightarrow{\mathbb{P}} 0. \tag{831}$$

By Lemma 23,

$$\sup_{s\in[\varepsilon,t]} \left\| \mathcal{V}_{\mathrm{eff}}^{(d)}(s) - V_{\mathrm{ext}}(\check{\mu}^{(d)}(s)) \right\|_{\rho_+} \xrightarrow{\mathbb{P}} 0. \tag{832}$$

The local Lipschitz estimate from Lemma 11 implies, on every bounded subset,

$$\|V_{\mathrm{ext}}(\mu) - V_{\mathrm{ext}}(\nu)\|_{\rho_+} \le \frac{L}{\rho_+ - \rho_-} \|\mu - \nu\|_{\rho_-}. \tag{833}$$

Hence (829) gives

$$\sup_{s\in[\varepsilon,t]} \left\| V_{\mathrm{ext}}(\check{\mu}^{(d)}(s)) - V_{\mathrm{ext}}(\mu^*(s)) \right\|_{\rho_+} \longrightarrow 0. \tag{834}$$

Passing to the limit in (830) yields

$$\mu^*(t) - \mu^*(\varepsilon) = \int_\varepsilon^t V_{\mathrm{ext}}(\mu^*(s))\, ds. \tag{835}$$

The effective vector fields $\mathcal{V}_{\mathrm{eff}}^{(d,n)}$ are uniformly bounded in $X^0_{\rho_+}$ before the stopping time, as established in the proof of Lemma 22. Therefore, letting $\varepsilon \downarrow 0$ in (835) and using the continuity of μ^* gives

$$\mu^*(t) = \bar{\mu}(0) + \int_0^t V_{\mathrm{ext}}(\mu^*(s))\, ds, \qquad 0 \le t \le T. \tag{836}$$

For every finite d, the empirical covariance belongs to $\mathfrak{Q}$ (see the proof of Lemma 23). Since $\mathfrak{Q}$ is closed and convergence in $X^0_{\rho_-}$ implies convergence of the second-order coordinates,

$$q(\mu^*(\tau)) \in \mathfrak{Q}, \qquad 0 \le \tau \le T. \tag{837}$$

Next, Lemma 12 gives a constant $c_T > 0$ such that, with probability tending to one, $t_k^{A,(n)} + t_k^{B,(n)} \ge c_T$ for every trainable k before the stopping time. Passing to the limit gives

$$t_k^A(\mu^*(\tau)) + t_k^B(\mu^*(\tau)) \ge c_T, \qquad k \in \mathcal{K}, \quad 0 \le \tau \le T. \tag{838}$$

Finally, Lemma 22 gives $|\check{\mu}_\alpha^{(d)}(\tau)| \le C_*^{2|\alpha|}$. Passing to the limit yields

$$|\mu_\alpha^*(\tau)| \le C_*^{2|\alpha|}. \tag{839}$$

Thus $\mathscr{P}(q(\mu^*(\tau))) = q(\mu^*(\tau))$, so along the limiting trajectory $\widetilde{m}^\star_{\mathcal{K}}(\mu^*(\tau)) = m^\star_{\mathcal{K}}(q(\mu^*(\tau)))$. Hence (836) is precisely the original system (16)–(19).

We now prove uniqueness. Suppose that μ_1 and μ_2 are two solutions with the same initial condition. Both solve the extended ODE. By Lemma 11, if $\mu_1(\tau_0) = \mu_2(\tau_0)$, at some $\tau_0 < T$, then the two trajectories agree on a nontrivial interval to the right of τ_0. The standard contradiction argument shows that $\mu_1(\tau) = \mu_2(\tau)$ for $0 \le \tau \le T$. Thus the solution is unique. We denote it by $\bar{\mu}$.

It remains to verify that the stopping time J_{mac} is asymptotically inactive. Recall that $\mathfrak{D} \Subset \mathcal{M} \times \mathcal{Q}$ is chosen to contain both the compact fast-time set $\mathcal{C}_{\mathrm{fast}} := \big\{(m_{\mathcal{K}}, q_0) : \Phi(m_{\mathcal{K}}, m_{\setminus\mathcal{K}}(0), q_0) \le \Phi(\widetilde{m}(0), q_0)\big\}$ and the compact slow trajectory $\mathcal{C}_{\mathrm{slow}} := \{(m^\star_{\mathcal{K}}(\bar{q}(\tau)), \bar{q}(\tau)) : 0 \le \tau \le T\}$ in its interior. Hence $\delta := \mathrm{dist}(\mathcal{C}_{\mathrm{fast}} \cup \mathcal{C}_{\mathrm{slow}}, \partial\mathfrak{D}) > 0$.

By C-tightness (Lemma 22) and uniqueness of the subsequential limit, the stopped moment trajectory converges uniformly in probability to $\bar{\mu}$. In particular,

$$\sup_{0 \le n \le N_T \wedge J_{\mathrm{mac}}} \|q^{(n)} - \bar{q}(n\alpha_d/d)\| \xrightarrow{\mathbb{P}} 0. \tag{840}$$

Moreover, for every fixed $T_1 > 0$, Lemma 16 and the strong convexity in Assumption 10 imply

$$\sup_{T_1 d/\alpha_d \le n \le N_T \wedge J_{\mathrm{mac}}} \|m^{(n)}_{\mathcal{K}} - m^\star_{\mathcal{K}}(q^{(n)})\| \xrightarrow{\mathbb{P}} 0. \tag{841}$$

Since $q \mapsto m^\star_{\mathcal{K}}(q)$ is Lipschitz on the compact subset of $\mathcal{Q}$, it follows that

$$\sup_{T_1 d/\alpha_d \le n \le N_T \wedge J_{\mathrm{mac}}} \Big\|(m^{(n)}_{\mathcal{K}}, q^{(n)}) - (m^\star_{\mathcal{K}}(\bar{q}(\tau_n)), \bar{q}(\tau_n))\Big\| \xrightarrow{\mathbb{P}} 0, \tag{842}$$

where $\tau_n = \frac{n\alpha_d}{d}$. Therefore, with probability tending to one, the trajectory stays at least $\delta/2$ away from $\partial\mathfrak{D}$ for $T_1 d/\alpha_d \le n \le N_T$.

It remains to control the initial regime and the intermediate regime. Choose a fixed fast-time horizon $t_{\mathrm{f}} > 0$ sufficiently large that $C_{\mathrm{loc}} e^{-c_{\mathrm{loc}} t_{\mathrm{f}}} < \delta/8$, where $C_{\mathrm{loc}}, c_{\mathrm{loc}}$ are the constants in Lemma 15. On $0 \le n \le t_{\mathrm{f}}/\alpha_d$, the fast-time convergence and the freezing of the slow coordinates proved in Appendix F.2.4 imply that the empirical trajectory remains uniformly close to the fast limiting trajectory. Since the latter stays in $\mathcal{C}_{\mathrm{fast}}$, this part of the trajectory remains at least a $\delta/4$ distance from $\partial\mathfrak{D}$ with probability tending to one.

For $t_{\mathrm{f}}/\alpha_d \le n \le T_1 d/\alpha_d$, Lemma 15 yields

$$\|m^{(n)}_{\mathcal{K}} - m^\star_{\mathcal{K}}(q^{(n)})\| \le C_{\mathrm{loc}} e^{-c_{\mathrm{loc}} t_{\mathrm{f}}} + o_{\mathbb{P}}(1) \tag{843}$$

uniformly before J_{mac}. Together with the uniform convergence of $q^{(n)}$ to $\bar{q}(n\alpha_d/d)$, this places the empirical state in a $\delta/4$-neighborhood of $\mathcal{C}_{\mathrm{slow}}$ throughout the intermediate regime. Combining the three time intervals gives

$$\mathbb{P}(J_{\mathrm{mac}} \le N_T) \longrightarrow 0. \tag{844}$$

This completes the proof. □

F.2.7 Final proof of Theorem 3

The fast-time convergence (20) was established in Appendix F.2.4. It remains to prove the convergence on the slow time scale.

Fix $0 < T_1 < T_2 \le T$. Choose $C_*^2 < \rho_- < \rho_+$, where C_* is the spectral constant appearing in Lemma 17.

By Lemma 22, the auxiliary moment trajectories $\{\check{\mu}^{(d)}\}_{d \ge 1}$ are C-tight in $D([0, T_2]; X^0_{\rho_-})$. Lemma 24 shows that every subsequential weak limit is almost surely equal to the same deterministic solution $\bar{\mu} \in C([0, T_2]; X^0_{\rho_-})$ of the slow system (16)–(19). Consequently,

$$\check{\mu}^{(d)} \to \bar{\mu} \qquad \text{in } D([0, T_2]; X^0_{\rho_-}). \tag{845}$$

Similarly to the tied case, this implies convergence in the uniform topology. Hence

$$\sup_{\tau \in [0, T_2]} \Big\|\check{\mu}^{(d)}(\tau) - \bar{\mu}(\tau)\Big\|_{\rho_-} \xrightarrow{\mathbb{P}} 0. \tag{846}$$

In particular, for every fixed word α,

$$\sup_{\tau\in[0,T_2]}\left|\check{\mu}_\alpha^{(d)}(\tau)-\bar{\mu}_\alpha(\tau)\right|\le \rho_-^{|\alpha|}\sup_{\tau\in[0,T_2]}\left\|\check{\mu}^{(d)}(\tau)-\bar{\mu}(\tau)\right\|_{\rho_-}, \tag{847}$$

and therefore

$$\sup_{\tau\in[0,T_2]}\left|\check{\mu}_\alpha^{(d)}(\tau)-\bar{\mu}_\alpha(\tau)\right|\xrightarrow{\mathbb{P}}0. \tag{848}$$

We now transfer this convergence to the SGD trajectory. Lemma 24 gives $\mathbb{P}(J_{\mathrm{mac}}\le N_{T_2})\longrightarrow 0$, where $N_{T_2}:=\left\lfloor\frac{T_2 d}{\alpha_d}\right\rfloor$, while Lemma 17 gives $\mathbb{P}\left(J_{\mathrm{sp}}\le J_{\mathrm{mac}}\wedge N_{T_2}\right)\longrightarrow 0$. Thus the event

$$\mathcal{G}_d:=\{J_{\mathrm{mac}}>N_{T_2},\ J_{\mathrm{sp}}>N_{T_2}\} \tag{849}$$

satisfies $\mathbb{P}(\mathcal{G}_d)\longrightarrow 1$. On $\mathcal{G}_d$, the auxiliary trajectory coincides with the original SGD trajectory throughout $[0,T_2]$. Hence, for every fixed α and every $\varepsilon>0$,

$$\mathbb{P}\left(\sup_{\tau\in[0,T_2]}\left|\tilde{\mu}_\alpha^{(d)}(\tau)-\bar{\mu}_\alpha(\tau)\right|>\varepsilon\right)\le\mathbb{P}(\mathcal{G}_d^c)+\mathbb{P}\left(\sup_{\tau\in[0,T_2]}\left|\check{\mu}_\alpha^{(d)}(\tau)-\bar{\mu}_\alpha(\tau)\right|>\varepsilon\right)\longrightarrow 0. \tag{850}$$

This proves (22).

It remains to establish the convergence of the mean variables on $[T_1,T_2]$. (22) implies

$$\sup_{\tau\in[0,T_2]}\left(\|q^{(1,d)}(\tau)-\bar{q}^{(1)}(\tau)\|_F+\|q^{(2,d)}(\tau)-\bar{q}^{(2)}(\tau)\|_F\right)\xrightarrow{\mathbb{P}}0. \tag{851}$$

On the other hand, recall that (see (841))

$$\sup_{\lceil T_1 d/\alpha_d\rceil\le n\le\lfloor T_2 d/\alpha_d\rfloor}\left\|m_{\mathcal{K}}^{(n)}-m_{\mathcal{K}}^\star(q^{(n)})\right\|\xrightarrow{\mathbb{P}}0. \tag{852}$$

By Assumption 10 and the implicit function theorem, the map $q\longmapsto m_{\mathcal{K}}^\star(q)$ is Lipschitz on a compact neighborhood of the limiting covariance trajectory. Combining this with (851) gives

$$\sup_{\tau\in[T_1,T_2]}\left\|m_{\mathcal{K}}^\star(q^{(d)}(\tau))-m_{\mathcal{K}}^\star(\bar{q}(\tau))\right\|\xrightarrow{\mathbb{P}}0. \tag{853}$$

Since $\bar{m}_{\mathcal{K}}(\tau)=m_{\mathcal{K}}^\star(\bar{q}(\tau))$, (852) and (853) imply

$$\sup_{\tau\in[T_1,T_2]}\left\|\tilde{m}^{(d)}(\tau d)-\bar{m}(\tau)\right\|_\infty\xrightarrow{\mathbb{P}}0. \tag{854}$$

This proves (21) and completes the proof of Theorem 3.

F.3 Truncation to finite orders

We now show that the infinite-dimensional dynamics can be approximated by a finite-order truncation. A truncated moment sequence need not correspond to actual matrices, and in particular its covariance parameters need not belong to the cone $\mathfrak{Q}$. For this reason, the truncated system is defined using the extended vector field V_{ext} introduced in (505).

For $M\ge 2$, define the projection

$$[\Pi_M\mu]_\alpha:=\begin{cases}\mu_\alpha, & |\alpha|\le M,\\ 0, & |\alpha|>M.\end{cases} \tag{855}$$

Let $Y_M:=\{\mu:\mu_\alpha=0 \text{ for every } |\alpha|>M\}$. Recall that V_{ext} is well defined on the open set

$$\mathcal{U}_s=\left\{\mu\in X_s:\mathscr{P}(q(\mu))\in\mathcal{Q},\quad\min_{k\in\mathcal{K}}\left(t_k^A(\mu)+t_k^B(\mu)\right)>0\right\}, \tag{856}$$

where $\mathscr{P}$ denotes the projection onto the covariance cone $\mathfrak{Q}$. We now define the degree-M truncated slow dynamics by

$$\frac{d}{d\tau}\mu_\alpha^{(M)}(\tau)=[V_{\mathrm{ext}}(\mu^{(M)}(\tau))]_\alpha,\qquad |\alpha|\le M, \tag{857}$$

with the boundary condition

$$\mu_\alpha^{(M)}(\tau) \equiv 0, \qquad |\alpha| > M, \tag{858}$$

and initial condition

$$\mu_\alpha^{(M)}(0) = \begin{cases} \bar{\mu}_\alpha(0), & |\alpha| \le M, \\ 0, & |\alpha| > M. \end{cases} \tag{859}$$

Notice that Y_M is invariant under V_{ext}. Hence (857)–(859) is exactly the restriction of the extended infinite-dimensional equation to the invariant finite-dimensional subspace Y_M. The following corollary proves that the truncated system can approximate the original system arbitrarily well.

Corollary 8. *Under the conditions of Theorem 3, fix $T > 0$, and let $R := C_*^2$, where C_* is the constant in Lemma 17. There exist constants $s_0 > R$, $C_T < \infty$ and $M_0(T) < \infty$, independent of M, such that for every $M \ge M_0(T)$, the truncated system* (857)–(859) *has a unique solution on $[0, T]$. Moreover, for every finite word α,*

$$\sup_{0 \le \tau \le T} \left| \mu_\alpha^{(M)}(\tau) - \bar{\mu}_\alpha(\tau) \right| \le C_T \left(\frac{R}{s_0} \right)^{M+1} s_0^{|\alpha|}. \tag{860}$$

Its proof follows the same general strategy as in the tied case (Corollary 2). The main additional difficulty in the untied setting is that the coefficients are themselves nonlinear functions of the moment sequence. In particular, the effective bias $\beta_k = -\psi_k/(t_k^A + t_k^B)$ contains a potentially small denominator. We therefore first establish uniform control of these coefficients in a neighborhood of the trajectory $\bar{\mu}$. Once these estimates are obtained, the remainder of the argument is parallel to the tied truncation proof.

Proof. By Lemma 24,

$$|\bar{\mu}_\alpha(\tau)| \le R^{|\alpha|}, \qquad 0 \le \tau \le T. \tag{861}$$

For $k \in \mathcal{K}$, write $D_k(\mu) := t_k^A(\mu) + t_k^B(\mu)$. Thus,

$$d_T := \min_{\substack{0 \le \tau \le T \\ k \in \mathcal{K}}} D_k(\bar{\mu}(\tau)) > 0. \tag{862}$$

Moreover, $\mathcal{K}_q := \{\mathscr{P}(q(\bar{\mu}(\tau))) : 0 \le \tau \le T\}$ is a compact subset of the open set $\mathcal{Q}$. Hence $\eta_T := \operatorname{dist}(\mathcal{K}_q, \mathcal{Q}^c) > 0$.

Fix $s_0 > R$. Since the covariance consists of second-order moments, there exists $C_q < \infty$, depending only on K, such that

$$\|q(\mu) - q(\nu)\|_F \le C_q s_0^2 \|\mu - \nu\|_{s_0}. \tag{863}$$

Since $\mathscr{P}$ is non-expansive,

$$\|\mathscr{P}(q(\mu)) - \mathscr{P}(q(\bar{\mu}(\tau)))\|_F \le C_q s_0^2 \|\mu - \bar{\mu}(\tau)\|_{s_0}. \tag{864}$$

Likewise, since D_k is the sum of two first-order moments,

$$|D_k(\mu) - D_k(\bar{\mu}(\tau))| \le 2 s_0 \|\mu - \bar{\mu}(\tau)\|_{s_0}. \tag{865}$$

Choose

$$0 < \delta_T \le \min\left\{ \frac{\eta_T}{2 C_q s_0^2}, \frac{d_T}{4 s_0} \right\}. \tag{866}$$

Then

$$\|\mu - \bar{\mu}(\tau)\|_{s_0} \le \delta_T \tag{867}$$

implies

$$\mathscr{P}(q(\mu)) \in \mathcal{Q}, \qquad D_k(\mu) \ge \frac{1}{2} D_k(\bar{\mu}(\tau)) \ge \frac{d_T}{2}, \qquad k \in \mathcal{K}. \tag{868}$$

Define $[\mathcal{W}(\mu)]_\alpha := [V_{\mathrm{ext}}(\mu)]_\alpha + 2\gamma |\alpha|_{\mathcal{K}} \mu_\alpha$. Following the proof of Lemma 11, (868) implies the following Lipchitz condition

$$|[\mathcal{W}(\mu) - \mathcal{W}(\bar{\mu}(\tau))]_\alpha| \le L_T |\alpha|_{\mathcal{K}} s_0^{|\alpha|} \|\mu - \bar{\mu}(\tau)\|_{s_0}, \tag{869}$$

for some $L_T < \infty$ independent of α and M.

By Assumption 9.5, we may assume that $\kappa_T := \frac{L_T}{2\gamma} < 1$

Since $V_{\mathrm{ext}}|_{Y_M}$ is a locally Lipschitz finite-dimensional vector field on this domain, the truncated system has a unique solution $\mu^{(M)}(\tau)$. Define

$$e^{(M)}(\tau) := \mu^{(M)}(\tau) - \bar{\mu}(\tau), \qquad E_M(\tau) := \|e^{(M)}(\tau)\|_{s_0}. \tag{870}$$

Let $\tau_M \le T$ be the first time at which either the solution $\mu^{(M)}(\tau)$ ceases to exist or $E_M(\tau) = \delta_T$. For sufficiently large M, $E_M(0) < \delta_T$, so $\tau_M > 0$.

Because Y_M is invariant under V_{ext}, the error satisfies, for every word α and every $0 \le \tau < \tau_M$,

$$\frac{d}{d\tau} e_\alpha^{(M)} = -2\gamma|\alpha|_{\mathcal{K}} e_\alpha^{(M)} + \mathcal{R}_\alpha^{(M)}, \tag{871}$$

where

$$|\mathcal{R}_\alpha^{(M)}(\tau)| \le L_T |\alpha|_{\mathcal{K}} s_0^{|\alpha|} E_M(\tau) \tag{872}$$

by (869).

If $|\alpha|_{\mathcal{K}} = 0$, then $\mathcal{R}_\alpha^{(M)} = 0$, and hence $e_\alpha^{(M)}(\tau) = e_\alpha^{(M)}(0)$. If $|\alpha|_{\mathcal{K}} \ge 1$, variation of constants gives

$$\frac{|e_\alpha^{(M)}(\tau)|}{s_0^{|\alpha|}} \le \frac{|e_\alpha^{(M)}(0)|}{s_0^{|\alpha|}} + L_T|\alpha|_{\mathcal{K}} \int_0^\tau e^{-2\gamma|\alpha|_{\mathcal{K}}(\tau-u)} E_M(u)\, du \le \frac{|e_\alpha^{(M)}(0)|}{s_0^{|\alpha|}} + \kappa_T \sup_{0\le u\le\tau} E_M(u), \tag{873}$$

since

$$|\alpha|_{\mathcal{K}} \int_0^\tau e^{-2\gamma|\alpha|_{\mathcal{K}}(\tau-u)}\, du \le \frac{1}{2\gamma}. \tag{874}$$

The same inequality also holds when $|\alpha|_{\mathcal{K}} = 0$. Taking the supremum over α and over $0 \le u \le \tau$ gives

$$\sup_{0\le u\le\tau} E_M(u) \le E_M(0) + \kappa_T \sup_{0\le u\le\tau} E_M(u). \tag{875}$$

Therefore,

$$\sup_{0\le u\le\tau} E_M(u) \le \frac{E_M(0)}{1-\kappa_T}, \qquad 0 \le \tau < \tau_M. \tag{876}$$

It remains to estimate the initial truncation error. By (861),

$$E_M(0) = \sup_{w>M} \sup_{|\alpha|=w} \frac{|\bar{\mu}_\alpha(0)|}{s_0^w} \le \sup_{w>M} \left(\frac{R}{s_0}\right)^w = \left(\frac{R}{s_0}\right)^{M+1}. \tag{877}$$

Choose $M_0(T)$ sufficiently large that, for every $M \ge M_0(T)$, $\frac{1}{1-\kappa_T}\left(\frac{R}{s_0}\right)^{M+1} < \frac{\delta_T}{2}$. Then (876) implies $\sup_{0\le\tau<\tau_M} E_M(\tau) < \frac{\delta_T}{2}$. Hence the truncated trajectory cannot reach the boundary $E_M = \delta_T$.

Moreover, throughout this interval (868) gives a positive distance from the boundary of $\mathcal{U}_s$, while for every $|\alpha| \le M$,

$$|\mu_\alpha^{(M)}(\tau)| \le R^{|\alpha|} + \delta_T s_0^{|\alpha|}. \tag{878}$$

Thus all coordinates of the truncated system remain bounded and $D_k(\mu^{(m)})$ remain bounded away from zero. The solution $\mu^{(m)}$ therefore cannot blow up before time T. Consequently, $\tau_M = T$, and the truncated system has a unique solution throughout $[0, T]$.

Combining (876) and (877), we obtain

$$\sup_{0\le\tau\le T} \|\mu^{(M)}(\tau) - \bar{\mu}(\tau)\|_{s_0} \le \frac{1}{1-\kappa_T}\left(\frac{R}{s_0}\right)^{M+1}. \tag{879}$$

Finally,

$$\left|\mu_\alpha^{(M)}(\tau) - \bar{\mu}_\alpha(\tau)\right| \le s_0^{|\alpha|} \|\mu^{(M)}(\tau) - \bar{\mu}(\tau)\|_{s_0}, \tag{880}$$

which proves (860). □

F.4 Proof of Corollary 4

We first prove part (i). By Theorem 3, the fast dynamics move the mean variable to $m_1^\star(\bar{q}^{(1)}(0), \bar{q}^{(2)}(0))$, while the normalized moments remain asymptotically frozen. The slow trajectory therefore starts from $\bar{\mathbf{q}}_{12}(0) = 0$ with the mean variables $\bar{m}^\star$.

Set $D_1(0) := \bar{t}_1^A(0) + \bar{t}_1^B(0) > 0$ and write

$$a_{12}^{(r)} := \left[\nabla_{q^{(r)}} \Phi\left(\bar{m}^\star, \bar{q}^{(1)}(0), \bar{q}^{(2)}(0)\right)\right]_{12}, \qquad r = 1, 2. \tag{881}$$

Thus $g_{12}^\star = (a_{12}^{(1)}, a_{12}^{(2)})^T$.

Consider first $\bar{q}_{12}^{(1)} = \bar{\mu}_{(M_1, M_2^T)}$. Since only the first index is trainable, the slow dynamics (19) gives

$$\frac{d}{d\tau}\bar{q}_{12}^{(1)} = -\bar{\beta}_1 \bar{\mu}_{(A_1+B_1, M_2^T)} - 2\sum_{l=1}^{2}\sum_{r=1}^{2} a_{1l}^{(r)} \bar{\mu}_{\left(M_l^{(r)} B_1 + A_1 M_l^{(r)}, M_2^T\right)} - 2\gamma \bar{q}_{12}^{(1)}. \tag{882}$$

At $\tau = 0$, Assumption 4 implies freeness of the student and teacher algebras. Moreover, $\lim_{d\to\infty} \frac{1}{d}\mathrm{Tr} M_2 = 0$. Hence

$$\bar{\mu}_{(A_1+B_1, M_2^T)}(0) = 0, \tag{883}$$

and all the $l = 1$ terms in (882) vanish.

For the $l = 2$ terms, asymptotic freeness gives

$$\bar{\mu}_{(M_2 B_1 + A_1 M_2, M_2^T)}(0) = D_1(0)\bar{q}_{22}^{(1)}, \tag{884}$$

$$\bar{\mu}_{(M_2^T B_1 + A_1 M_2^T, M_2^T)}(0) = D_1(0)\bar{q}_{22}^{(2)}. \tag{885}$$

Therefore

$$\left.\frac{d}{d\tau}\bar{q}_{12}^{(1)}\right|_{\tau=0} = -2D_1(0)\left(a_{12}^{(1)}\bar{q}_{22}^{(1)} + a_{12}^{(2)}\bar{q}_{22}^{(2)}\right). \tag{886}$$

Similarly, applying (19) to $\bar{q}_{12}^{(2)} = \bar{\mu}_{(M_1, M_2)}$ gives

$$\left.\frac{d}{d\tau}\bar{q}_{12}^{(2)}\right|_{\tau=0} = -2D_1(0)\left(a_{12}^{(1)}\bar{q}_{22}^{(2)} + a_{12}^{(2)}\bar{q}_{22}^{(1)}\right). \tag{887}$$

Combining (886) and (887),

$$\left.\frac{d}{d\tau}\bar{\mathbf{q}}_{12}(\tau)\right|_{\tau=0} = -2D_1(0)\mathsf{Q}_2 g_{12}^\star. \tag{888}$$

Since $\mathsf{Q}_2 g_{12}^\star \neq 0$, the right-hand side is nonzero. Define

$$v_0 := \left\| \left.\frac{d}{d\tau}\bar{\mathbf{q}}_{12}(\tau)\right|_{\tau=0} \right\|_\infty > 0. \tag{889}$$

The limiting slow trajectory is continuously differentiable. Hence, there exists an index $r_0 \in \{1, 2\}$ and $T_0 > 0$ such that

$$\left|\frac{d}{d\tau}\bar{q}_{12}^{(r_0)}(\tau)\right| \geq \frac{v_0}{2}, \qquad 0 \leq \tau \leq T_0. \tag{890}$$

In particular, by continuity $\frac{d}{d\tau}\bar{q}_{12}^{(r_0)}(\tau)$ has the same sign as $\frac{d}{d\tau}\bar{q}_{12}^{(r_0)}(0)$ throughout $[0, T_0]$. Since $\bar{q}_{12}^{(r_0)}(0) = 0$,

$$\|\bar{\mathbf{q}}_{12}(T_0)\|_\infty \geq \frac{v_0 T_0}{2}. \tag{891}$$

Set $c := \frac{v_0 T_0}{4}$ and $T_+ := T_0$. Then

$$\|\bar{\mathbf{q}}_{12}(T_+)\|_\infty \geq 2c. \tag{892}$$

On the other hand, continuity and $\bar{\mathbf{q}}_{12}(0) = 0$ imply that there exists $T_- \in (0, T_+)$ such that

$$\sup_{0\leq\tau\leq T_-} \|\bar{\mathbf{q}}_{12}(\tau)\|_\infty < \frac{c}{2}. \tag{893}$$

Let $\widetilde{\mathbf{q}}_{12}^{(d)}(\tau)$ denote the interpolation of the empirical overlap. Theorem 3 gives

$$\sup_{\tau\in[0,T_+]} \left\|\widetilde{\mathbf{q}}_{12}^{(d)}(\tau) - \bar{\mathbf{q}}_{12}(\tau)\right\|_\infty \xrightarrow{\mathbb{P}} 0. \tag{894}$$

Using (892) and (893), with probability tending to one,

$$\left\lfloor \frac{T_- d}{\alpha_d} \right\rfloor < N_{\mathrm{wr}}^{(d)}(c) \le \left\lceil \frac{T_+ d}{\alpha_d} \right\rceil. \tag{895}$$

This proves part (i).

We next prove part (ii). Let $\bar{\mu}(0) \in \mathcal{M}_{\mathrm{free}}$ be the limiting initial moment sequence. Consider any state $\mu \in \mathcal{M}_{\mathrm{free}}$. In the slow vector field (19), the only substitutions which explicitly insert a teacher letter into the evolution of a trainable student letter are $\mathcal{T}_{i,2,r}^{\mathrm{diff}}$. Their coefficients are $\left[\nabla_{q^{(r)}}\Phi\left(m^\star(q), q^{(1)}, q^{(2)}\right)\right]_{12}$, which vanish on $\mathcal{M}_{\mathrm{free}}$. The same observation applies to the $l=2$ contribution to ψ_1, and hence to the effective bias β_1. Consequently, when the state belongs to $\mathcal{M}_{\mathrm{free}}$, the limiting slow dynamics do not generate mixed student–teacher moments. Thus, $\mathcal{M}_{\mathrm{free}}$ is invariant under the limiting slow flow. By uniqueness of the solution in Theorem 3, $\bar{\mu}(\tau) \in \mathcal{M}_{\mathrm{free}}$ for $0 \le \tau \le T$. In particular, $\bar{q}_{12}^{(1)}(\tau) = \bar{q}_{12}^{(2)}(\tau) = 0$ for $0 \le \tau \le T$.

Applying again the uniform convergence in Theorem 3, for every fixed $T > 0$,

$$\sup_{\tau\in[0,T]} \left\|\widetilde{\mathbf{q}}_{12}^{(d)}(\tau)\right\|_\infty \xrightarrow{\mathbb{P}} 0. \tag{896}$$

Hence, for every fixed $c > 0$,

$$\mathbb{P}\left(N_{\mathrm{wr}}^{(d)}(c) \le \left\lfloor \frac{Td}{\alpha_d} \right\rfloor\right) \longrightarrow 0. \tag{897}$$

This finishes the proof.

F.5 Convergence rate of strong recovery

We next study the convergence rate after weak recovery. Recall the first-order moments $t_k^A := \frac{1}{d}\mathrm{Tr}[A_k]$, $t_k^B := \frac{1}{d}\mathrm{Tr}[B_k]$ and the covariance parameters $q_{kl}^{(1)} := \frac{1}{d}\mathrm{Tr}[M_k M_l^T]$, $q_{kl}^{(2)} := \frac{1}{d}\mathrm{Tr}[M_k M_l]$. The limiting regularized population loss is

$$\Psi(m, t^A, t^B, q^{(1)}, q^{(2)}) := \Phi(m, q^{(1)}, q^{(2)}) + \frac{\gamma}{2}\sum_{k\in\mathcal{K}}\left(t_k^A + t_k^B\right). \tag{898}$$

Along the slow trajectory of Theorem 3, the mean variables are slaved to the covariance,

$$\bar{m}_{\mathcal{K}}(\tau) = m_{\mathcal{K}}^\star\left(\bar{q}^{(1)}(\tau), \bar{q}^{(2)}(\tau)\right). \tag{899}$$

Let $\mathcal{Q}_{\mathrm{untied}}$ denote the asymptotically realizable domain of $\left(m, t^A, t^B, q^{(1)}, q^{(2)}\right)$, namely the closure of all limiting order parameters generated by sequences of matrices $M_k = U_k V_k^T, A_k = U_k U_k^T, B_k = V_k V_k^T$ subject to the fixed teacher components for $k \notin \mathcal{K}$. Define

$$\Psi^\star := \inf_{(m,t^A,t^B,q^{(1)},q^{(2)})\in\mathcal{Q}_{\mathrm{untied}}} \Psi(m, t^A, t^B, q^{(1)}, q^{(2)}). \tag{900}$$

We refer to convergence to the global minimizers of Ψ as strong recovery.

Let $\bar{\vartheta}(\tau) := \left(\bar{t}^A(\tau), \bar{t}^B(\tau), \bar{q}^{(1)}(\tau), \bar{q}^{(2)}(\tau)\right)$ denote the corresponding components of the limiting slow trajectory. As in the tied setting, we impose an effective Polyak–Łojasiewicz condition along this trajectory.

Assumption 11. *There exist constants $T_0 > 0$ and $c > 0$ such that, for every $\tau \ge T_0$,*

$$-\frac{d}{d\tau}\Psi\left(\bar{m}(\tau), \bar{t}^A(\tau), \bar{t}^B(\tau), \bar{q}^{(1)}(\tau), \bar{q}^{(2)}(\tau)\right) \ge 2c\left[\Psi\left(\bar{m}(\tau), \bar{t}^A(\tau), \bar{t}^B(\tau), \bar{q}^{(1)}(\tau), \bar{q}^{(2)}(\tau)\right) - \Psi^\star\right]. \tag{901}$$

For finite d, define the regularized population loss

$$\Psi_d(U,V) := \mathbb{E}_x\left[\mathcal{L}\left(\left\{\frac{1}{\sqrt{d}}x_i^T U_k V_k^T x_j\right\}_{i,j,k=1}^{L,L,K}\right)\right] + \frac{\gamma}{2d}\sum_{k\in\mathcal{K}}\left(\|U_k\|_F^2 + \|V_k\|_F^2\right). \tag{902}$$

Corollary 9. *Suppose that the conditions of Theorem 3 and Assumption 11 hold. Let* $n_0 := \left\lceil \frac{T_0 d}{\alpha_d} \right\rceil$. *Then, for every fixed target accuracy* $\epsilon > 0$*, there exists*

$$n_\epsilon = n_0 + O\left(\frac{d}{\alpha_d}\log_+\frac{1}{\epsilon}\right) \tag{903}$$

such that

$$\lim_{d\to\infty}\mathbb{P}\left(\Psi_d(U^{(n_\epsilon)}, V^{(n_\epsilon)}) - \Psi^\star > \epsilon\right) = 0, \tag{904}$$

where $\log_+(x) := \max\{0, \log x\}$*. In particular, if* $T_0 = \Theta(1)$*, reaching an* ϵ*-suboptimal population loss requires at most* $O\left(\frac{d}{\alpha_d}\log\frac{1}{\epsilon}\right)$ *additional samples after entering the strong-recovery regime. For* $\alpha_d = \Theta((d\log d)^{-1})$*, this becomes* $O\left(d^2\log d\,\log\frac{1}{\epsilon}\right)$.

Proof. Define the limiting excess regularized population loss

$$\Delta(\tau) := \Psi\left(\bar{m}(\tau), \bar{t}^A(\tau), \bar{t}^B(\tau), \bar{q}^{(1)}(\tau), \bar{q}^{(2)}(\tau)\right) - \Psi^\star \geq 0. \tag{905}$$

By Assumption 11, for every $\tau \geq T_0$,

$$\frac{d}{d\tau}\Delta(\tau) \leq -2c\Delta(\tau). \tag{906}$$

Hence Grönwall's inequality gives

$$\Delta(\tau) \leq \Delta(T_0)e^{-2c(\tau-T_0)}, \qquad \tau \geq T_0. \tag{907}$$

Set $\Delta_0 := \Delta(T_0)$. If $\Delta_0 = 0$, there is nothing to prove. Otherwise, for a fixed $\epsilon > 0$, define

$$T_\epsilon := T_0 + \frac{1}{2c}\log_+\left(\frac{2\Delta_0}{\epsilon}\right). \tag{908}$$

Then

$$\Psi\left(\bar{m}(T_\epsilon), \bar{t}^A(T_\epsilon), \bar{t}^B(T_\epsilon), \bar{q}^{(1)}(T_\epsilon), \bar{q}^{(2)}(T_\epsilon)\right) - \Psi^\star \leq \frac{\epsilon}{2}. \tag{909}$$

We next transfer this bound to the finite-dimensional population loss. For $\tau \geq 0$, define $U^{(d)}(\tau) := U^{(\lfloor \tau d/\alpha_d \rfloor)}$ and $V^{(d)}(\tau) := V^{(\lfloor \tau d/\alpha_d \rfloor)}$. For every fixed $0 < T_1 < T_2 < \infty$, Theorem 3, together with the uniform spectral bounds in Lemma 17, implies

$$\sup_{\tau\in[T_1,T_2]}\left|\Psi_d\left(U^{(d)}(\tau), V^{(d)}(\tau)\right) - \Psi\left(\bar{m}(\tau), \bar{t}^A(\tau), \bar{t}^B(\tau), \bar{q}^{(1)}(\tau), \bar{q}^{(2)}(\tau)\right)\right| \xrightarrow{\mathbb{P}} 0, \tag{910}$$

where we use Theorem 1.

Since $T_0 > 0$ and $T_\epsilon < \infty$, we may apply (910) on $[T_0, T_\epsilon]$. Consequently,

$$\mathbb{P}\left(\left|\Psi_d\left(U^{(d)}(T_\epsilon), V^{(d)}(T_\epsilon)\right) - \Psi\left(\bar{m}(T_\epsilon), \bar{t}^A(T_\epsilon), \bar{t}^B(T_\epsilon), \bar{q}^{(1)}(T_\epsilon), \bar{q}^{(2)}(T_\epsilon)\right)\right| > \frac{\epsilon}{2}\right) \longrightarrow 0. \tag{911}$$

Set $n_\epsilon := \left\lceil \frac{T_\epsilon d}{\alpha_d} \right\rceil$. Using (909) and (911), we obtain

$$\mathbb{P}\left(\Psi_d(U^{(n_\epsilon)}, V^{(n_\epsilon)}) - \Psi^\star > \epsilon\right) \longrightarrow 0. \tag{912}$$

Finally, by (908),

$$n_\epsilon - n_0 = O\left(\frac{d}{\alpha_d}\log_+\frac{1}{\epsilon}\right), \tag{913}$$

which proves the claimed sample complexity. □

F.6 Proof of Lemma 10

We concatenate the indices (i, j, k) into a single index $a \in \{1, \dots, P\}$, where $P = L^2K = \Theta(1)$, and write $G^{(d)} = (G_a^{(d)})_{a=1}^P$.

As in the proof of Theorem 1, concatenate the input vectors into $X = [x_1^T, \dots, x_L^T]^T \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = \mathcal{C} \otimes I_d$. For each $a = (i, j, k)$, there exists a symmetric $Ld \times Ld$ matrix A_a such that $G_a^{(d)} = X^T A_a X$. Writing $X = \Sigma^{1/2} z$, with $z \sim \mathcal{N}(0, I_{Ld})$, and setting $\widetilde{A}_a := \Sigma^{1/2} A_a \Sigma^{1/2}$, we obtain a Gaussian quadratic form

$$G_a^{(d)} = z^T \widetilde{A}_a z. \tag{914}$$

By the same norm estimates as in the proof of Theorem 1, the assumptions $\|M_k\|_{\mathrm{op}} \le C_*$ imply, uniformly over a,

$$\|\widetilde{A}_a\|_{\mathrm{op}} \le \frac{C}{\sqrt{d}}, \qquad \|\widetilde{A}_a\|_F \le C, \tag{915}$$

where C depends only on $C_*, \mathcal{C}, L, K$. Moreover, $\max_k \frac{1}{\sqrt{d}} |\mathrm{Tr}(M_k)| \le R_*$ implies that the means $\mu_a^{(d)} := \mathbb{E} G_a^{(d)} = \mathrm{Tr}(\widetilde{A}_a)$ are uniformly bounded.

Let

$$W_a^{(d)} := G_a^{(d)} - \mu_a^{(d)} = z^T \widetilde{A}_a z - \mathrm{Tr}(\widetilde{A}_a), \tag{916}$$

and let $Z^{(d)}$ be a centered Gaussian vector having the same covariance as $W^{(d)}$. Thus $G_{\mathrm{G}}^{(d)} = \mu^{(d)} + Z^{(d)}$.

As in Theorem 1, Wick's formula gives

$$\Xi_{ab}^{(d)} := \mathrm{Cov}(W_a^{(d)}, W_b^{(d)}) = 2\mathrm{Tr}(\widetilde{A}_a \widetilde{A}_b). \tag{917}$$

In particular, $\Xi^{(d)}$ is uniformly bounded.

We now compare $W^{(d)}$ with $Z^{(d)}$ through Gaussian interpolation. Let $Z^{(d)}$ be independent of z, and define

$$Y_t := \mu^{(d)} + \sqrt{t}\, W^{(d)} + \sqrt{1-t}\, Z^{(d)}, \qquad t \in [0, 1], \tag{918}$$

and $F(t) := \mathbb{E}[f(Y_t)]$. Then $F(1) = \mathbb{E}[f(G^{(d)})]$ and $F(0) = \mathbb{E}[f(G_{\mathrm{G}}^{(d)})]$.

For the centered quadratic forms $W_a^{(d)}$, define $\Gamma_{ab} := 2 z^T \widetilde{A}_a \widetilde{A}_b z$. The Gaussian integration by parts gives

$$\mathbb{E}\left[W_a^{(d)} \partial_a f(Y_t) \right] = \sqrt{t} \sum_{b=1}^P \mathbb{E}\left[\Gamma_{ab}\, \partial_{ab}^2 f(Y_t) \right]. \tag{919}$$

On the other hand, Gaussian integration by parts for $Z^{(d)}$ gives

$$\mathbb{E}\left[Z_a^{(d)} \partial_a f(Y_t) \right] = \sqrt{1-t} \sum_{b=1}^P \Xi_{ab}^{(d)} \mathbb{E}\left[\partial_{ab}^2 f(Y_t) \right]. \tag{920}$$

Differentiating $F(t)$ and combining (919)–(920) yields

$$F'(t) = \frac{1}{2} \sum_{a,b=1}^P \mathbb{E}\left[\left(\Gamma_{ab} - \Xi_{ab}^{(d)} \right) \partial_{ab}^2 f(Y_t) \right]. \tag{921}$$

It remains to control the fluctuation of Γ_{ab}. By (917), $\mathbb{E}[\Gamma_{ab}] = \Xi_{ab}^{(d)}$. Since $\widetilde{A}_a$ and $\widetilde{A}_b$ are symmetric, we have

$$\mathbb{E}\left| \Gamma_{ab} - \Xi_{ab}^{(d)} \right|^2 \le C \|\widetilde{A}_a \widetilde{A}_b\|_F^2 \le C \|\widetilde{A}_a\|_{\mathrm{op}}^2 \|\widetilde{A}_b\|_F^2 \le \frac{C}{d}, \tag{922}$$

where we used (915) in the last step.

As in the final part of the proof of Theorem 1, for every fixed $p < \infty$,

$$\sup_d \mathbb{E}\|W^{(d)}\|^p < \infty. \tag{923}$$

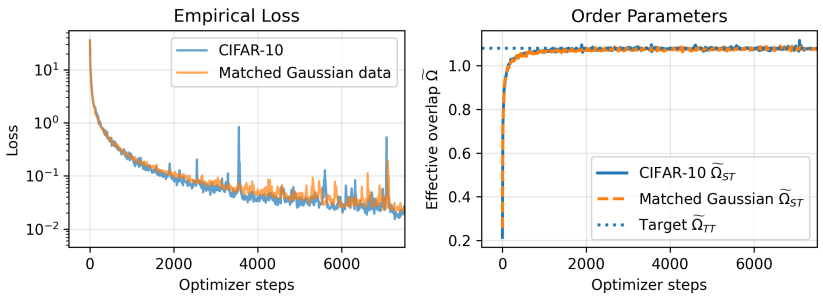


Figure 5: Comparison of empirical training loss (**Left**) and effective student–teacher overlap $\widetilde{\Omega}_{ST}$ (**Right**) for single-layer attention models trained on CIFAR-10 patches and Gaussian inputs with matched covariances. Both models exhibit nearly identical loss and overlap trajectories, with $\widetilde{\Omega}_{ST}$ approaching the target $\widetilde{\Omega}_{TT}$. We use $L = 16$, $d = 192$, Adam with learning rate 10^{-3}, and batch size 128.

The Gaussian vector $Z^{(d)}$ has uniformly bounded covariance, so its moments are uniformly bounded as well. Since the means $\mu^{(d)}$ are uniformly bounded, it follows that

$$\sup_d \sup_{t\in[0,1]} \mathbb{E}\|Y_t\|^p < \infty \tag{924}$$

for every fixed p. Since the second derivatives of f grow polynomially, (924) therefore implies

$$\sup_d \sup_{t\in[0,1]} \mathbb{E}\left|\partial^2_{ab} f(Y_t)\right|^2 \le C_f. \tag{925}$$

Applying Cauchy–Schwarz to (921), and using (922) and (925), gives

$$|F'(t)| \le \frac{C}{\sqrt{d}}, \qquad t \in [0,1], \tag{926}$$

uniformly over all matrix collections.

Integrating from 0 to 1, we conclude that

$$\left|\mathbb{E}[f(G^{(d)})] - \mathbb{E}[f(G^{(d)}_{\mathrm{G}})]\right| = |F(1) - F(0)| \le \frac{C}{\sqrt{d}}. \tag{927}$$

This proves the lemma.

G Additional Experiments

To investigate the robustness of the Gaussian approximation beyond Gaussian inputs, we conduct a teacher–student experiment using CIFAR-10 images. Each 32×32 RGB image is divided into $L = 16$ non-overlapping 8×8 patches, yielding input tokens $X = (x_1, \dots, x_L)^\top \in \mathbb{R}^{L\times d}$ with $d = 192$. Each patch is normalized using layer normalization and subsequently centered by subtracting its empirical mean across images.

To construct a covariance-matched Gaussian surrogate, we estimate the feature covariance $\widehat{\Sigma} \in \mathbb{R}^{d\times d}$ and token covariance $\widehat{C} \in \mathbb{R}^{L\times L}$ from the CIFAR-10 patches. Writing $x_i^{(n)}$ for the centered token at position i in image n, these matrices are defined by

$$\widehat{\Sigma} = \frac{1}{NL}\sum_{n=1}^{N}\sum_{i=1}^{L} x_i^{(n)}(x_i^{(n)})^\top, \qquad \widehat{C}_{ij} = \frac{1}{N\,\mathrm{Tr}(\widehat{\Sigma})}\sum_{n=1}^{N}(x_i^{(n)})^\top x_j^{(n)}, \tag{928}$$

where N denotes the number of images used for covariance estimation. We then generate synthetic Gaussian inputs according to

$$X_{\mathrm{G}} = \widehat{C}^{1/2} Z \widehat{\Sigma}^{1/2}, \qquad Z_{ia} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \tag{929}$$

so that $\mathbb{E}[(X_{\mathrm{G}})_{ia}(X_{\mathrm{G}})_{jb}] = \widehat{C}_{ij}\widehat{\Sigma}_{ab}$. This construction matches the token and feature covariance statistics of CIFAR-10. Both the teacher and student use a single-layer attention model

$$f_S(X) = X + \mathrm{softmax}\left(\frac{XSX^\top}{\sqrt{d}}\right)X, \tag{930}$$

where $S \in \mathbb{R}^{d\times d}$ is the attention matrix. We fix a randomly initialized teacher S_T and train two students, initialized identically, on CIFAR-10 and covariance-matched Gaussian inputs with the mean square loss, respectively. To account for the anisotropic feature covariance, we define the effective attention matrices $\widetilde{S} = \widehat{\Sigma}^{1/2}S\widehat{\Sigma}^{1/2}$ and $\widetilde{S}_T = \widehat{\Sigma}^{1/2}S_T\widehat{\Sigma}^{1/2}$. The effective student–teacher overlap is then

$$\widetilde{\Omega}_{ST} = \frac{1}{d}\,\mathrm{Tr}(\widetilde{S}\widetilde{S}_T^\top) = \frac{1}{d}\,\mathrm{Tr}(S\widehat{\Sigma}S_T^\top\widehat{\Sigma}), \tag{931}$$

with the teacher self-overlap $\widetilde{\Omega}_{TT} = \frac{1}{d}\,\mathrm{Tr}(\widetilde{S}_T\widetilde{S}_T^\top)$ serving as a reference.

As shown in Figure 5, the two input distributions exhibit remarkably similar learning behavior. Their empirical losses follow nearly identical trajectories, while the effective overlaps $\widetilde{\Omega}_{ST}$ rapidly increase and approach the target $\widetilde{\Omega}_{TT}$, remaining closely aligned throughout training. Despite the non-Gaussian and spatially structured nature of CIFAR-10 patches, the covariance-matched Gaussian surrogate reproduces the observed loss and overlap trajectories with high accuracy. These results provide numerical evidence that our characterization of attention-indexed models in Theorem 1 and Corollary 1 remains informative beyond the Gaussian input setting considered in our theoretical analysis.

H Declaration of LLM usage

We used LLMs in the following stages of the preparation of this work: editing (e.g., grammar, spelling, word choice), drafting sections of the paper, assisting in the writing of proofs, facilitating or running experiments, visualizing results for submission. All LLM-generated material used in this manuscript was manually verified by the authors to ensure that it performs as intended, in particular with regard to the proof and the numerical experiments.

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 main claims made in the abstract and introduction accurately reflect the paper's contributions and scope: characterizing the landscape of the population dynamics of attention-indexed models.

   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: The limitations are discussed in the conclusion.

   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: We provide all the assumptions and proofs.

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: We provide all the experimental settings and hyperparameters.

   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: [No]

   Justification: The experiments are only for simple synthetic settings and can be easily reproduced.

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: We provide all the experimental settings and hyperparameters.

   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: The experiments are for demonstrative purposes. For different random seeds the results are similar.

   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: The experiments are for toy models and can be executed on personal computers within minutes.

   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: Our work is theoretical.

   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: [No]

    Justification: Our work has many potential positive societal impact, none of which we feel we should emphasize due to its theoretical nature.

    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: We do not release datasets or models.

    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: We do not use existing assets.

    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: [N/A]

    Justification: We do not release new assets.

    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: This paper does not involve crowdsourcing nor research with 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: This paper does not involve crowdsourcing nor research with 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: [N/A]

    Justification: We only use LLMs for writing and faciliating experiments.

    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.