CLUSTERING THRESHOLDS FOR HOMOGENIZED TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the evolution of spherical token representations under randomly resampled attention and MLP layers, in a continuous-depth model where all tokens are driven by the same random fields. Our main result is that the long-time behavior is decided by the sign of a single explicit quantity. When this quantity is negative, tokens started close together collapse to a single cluster with probability tending to one as their initial diameter tends to zero. When it is positive, distinct tokens almost surely do not converge. We also give a sufficient condition under which the mean deficit decays exponentially and synchronization holds almost surely from every initial configuration, and we show that the same rate governs the Jacobian of the token flow. We connect the thresholds separating different clustering behavior with (quenched and annealed) Lyapunov exponents, and identify a regime where token representations can cluster even as small perturbations of their shared representation are amplified.

1 INTRODUCTION

The evolution of token representations through a transformer can be studied as a system of interacting particles, with network depth playing the role of time (Geshkovski et al., 2023; 2025). In models where the tokens are constrained to the sphere, this formulation allows us to describe how their relative positions change under repeated application of attention and MLP layers. Of particular interest is the possibility that all token representations become arbitrarily close as depth increases, since such a collapse removes the differences between their representations. Although clustering has been established for several pure-attention models (Geshkovski et al., 2023), the addition of MLP layers raises the question of how clustering depends on their activation functions and initialization scales.

In this paper we consider this question for randomly initialized layers whose parameters are resampled across depth. The homogenized model introduced by Koubbi et al. (2026) describes this setting through a continuous-depth limit in which the centered attention field becomes a common noise, meaning that all tokens evolve under the same realization of a random field. For transformers with post-layer normalization, Agazzi et al. (2026) derive an MLP common noise in the infinite-width and infinite-depth limits, with fixed attention parameters contributing a deterministic drift. We combine their MLP noise with the homogenized attention of Koubbi et al. (2026), allowing us to study how the relative strength of the two random fields affects token separation.

Our main question is how the interaction between attention and MLP layers at initialization affects representation collapse and vanishing gradients in deep transformers. To answer it, we study the exponential rate at which an infinitesimal distance between tokens changes along a realization of the dynamics. We demonstrate that attention contributes an explicit quantity to the rate produced by the MLP, and that the sign of the resulting rate determines a local clustering criterion.

1.1 CONTRIBUTIONS

We show that two explicit numbers govern the model at initialization:

$$\lambda_{\mathrm{MLP}} = \tfrac{1}{2}\big[(d-3)\kappa'(1) - (d-1)\kappa(1)\big], \qquad \lambda_{\mathrm{ATT}} = -\frac{\sigma_V^2(d-1)}{2d},$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

where d is the embedding dimension, κ is the covariance kernel of the MLP noise, set by the activation and the initialization variances, and σ_V is the scale of the value matrices. Our main results are the following.

- The sign of $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}$ decides local clustering (Theorem 1). If it is negative, tokens that start close together collapse to a single point with probability tending to one as their initial diameter vanishes. If it is positive, two distinct tokens almost surely never converge, from any initial configuration. The threshold does not depend on the number of tokens.
- We give an explicit sufficient condition for synchronization from every initial configuration (Theorem 4). Under this condition, the mean squared distance between tokens decays exponentially in expectation, and each pairwise distance decays almost surely at rate $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}$ (Corollary 6). Some additional condition is necessary: with an odd activation and no biases, antipodal tokens remain antipodal, so synchronization cannot hold from every initial configuration.
- We make explicit the link between token synchronization and Lyapunov exponents, which is classical in the theory of random dynamical systems (Baxendale, 1989; Flandoli et al., 2017). At coincident tokens, $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}$ is the pathwise exponential growth rate (Lyapunov exponent) of the input-output Jacobian, and of its transpose, in the directions that separate tokens (Theorem 9). We also compute the corresponding rate for directions that move all tokens together, and show that when it is negative the parameter gradients of early blocks acting on coincident tokens vanish exponentially with depth (Proposition 32). More generally, our analysis illustrates how techniques from stochastic flows can lead to insights on the residual-stream dynamics of deep transformers.

Finally, we present numerical simulations of transformer models that confirm the theoretical predictions (Section 5).

1.2 Related work

Continuous-depth self-attention has been studied through interacting-particle and mean-field models. Early works established clustering and concentration phenomena for deterministic attention dynamics (Geshkovski et al., 2023; 2025), followed by quantitative results on convergence and metastability (Chen et al., 2025; Bruno et al., 2025a;b; Castin et al., 2026). Sander et al. (2022) introduce a Sinkhorn-normalized attention and show that it admits a Wasserstein gradient-flow interpretation. Burger et al. (2025) identify a related gradient-flow structure for normalized self-attention and study its stationary states and interaction energy. Low-temperature concentration and metastable time scales are studied by Alcalde et al. (2026). These works keep the attention dynamics deterministic across depth. In our setting the parameters are resampled, so the continuous-depth dynamics is instead a stochastic flow driven by noise shared by the tokens.

Random initialization leads to a second line of work, closer to ours. Fedorov et al. (2026) show that resampling value matrices can qualitatively change the deterministic clustering picture, including through attraction toward antipodal configurations. Koubbi et al. (2026) derive homogenized limits for randomly resampled multi-head attention and study representation collapse in several asymptotic regimes. Depending on the scaling, their limits are deterministic or driven by common noise. Agazzi et al. (2026) obtain common noise from randomly initialized MLP blocks, with fixed attention contributing a deterministic drift, and prove synchronization under a coercivity condition on the noise. Our model combines these two stochastic mechanisms while keeping the number of tokens, dimension and temperature fixed. The sign of an explicit pathwise Lyapunov exponent then separates local synchronization from non-synchronization.

This places the problem in the broader theory of synchronization by common noise and stochastic flows. General criteria for synchronization and singleton random attractors were developed in Flandoli et al. (2017), with closely related results for isotropic stochastic flows and their Lyapunov exponents (Cranston et al., 2016; Baxendale, 1991; 1989). Closer to transformer models, Engel & Shalova (2026) study a spherical random flow generated by random linear maps in which an antipodal symmetry remains, while Shalova (2026) shows that additional noise can break this symmetry and restore synchronization. The Lyapunov viewpoint also connects our results to work on signal propagation in transformers. Noci et al. (2022) relate rank collapse to gradient propagation, and

Cowsik et al. (2024) derive forward and backward growth exponents at random initialization. Their criteria concern quantities averaged over initialization. Our analysis instead separates pathwise and second-moment growth and isolates the transverse Jacobian directions that control relative token positions.

2 SETUP AND NOTATION

We consider $N \geq 2$ initial tokens $X_0^i \in \mathbb{S}^{d-1} \subset \mathbb{R}^d$ for $i \in [N]$, with $d \geq 3$.

Given matrices $\theta = (V, H)$, we define attention for a configuration $X = (X^1, \dots, X^N)$ and a query $x \in \mathbb{S}^{d-1}$ by

$$\mathrm{Attn}_\theta[X](x) := \frac{\sum_{k=1}^N e^{\langle Hx, X^k\rangle} V X^k}{\sum_{k=1}^N e^{\langle Hx, X^k\rangle}}. \tag{1}$$

This is the usual softmax attention, with $H = K^\top Q$ for query and key matrices Q, K. We impose the initialization law directly on (V, H): their entries are independent, with $V_{ab} \sim \mathcal{N}(0, \sigma_V^2/d)$ and $H_{ab} \sim \mathcal{N}(0, \sigma_H^2/d)$, where $\sigma_V, \sigma_H \geq 0$. The scale σ_V controls the attention amplitude, while σ_H controls the scale of the logits. Denote this law by ρ_* on $\Theta := (\mathbb{R}^{d\times d})^2$. At layer ℓ, we sample $\theta_\ell \sim \rho_*$ and write $\mathrm{Attn}_\ell := \mathrm{Attn}_{\theta_\ell}$.

For the MLP block, let $\varphi : \mathbb{R} \to \mathbb{R}$ be a Lipschitz activation, applied coordinatewise, and let $m \geq 1$ be the hidden width. We use matrices $U_\ell \in \mathbb{R}^{m\times d}$ and $A_\ell \in \mathbb{R}^{d\times m}$, with biases $b_\ell \in \mathbb{R}^m$ and $c_\ell \in \mathbb{R}^d$, and define

$$G_\ell^m(x) := \frac{1}{\sqrt{m}} A_\ell \varphi\left(\frac{1}{\sqrt{d}} U_\ell x + b_\ell\right) + c_\ell. \tag{2}$$

The entries of U_ℓ and A_ℓ are independent standard Gaussians, and the bias entries are independent centered Gaussians with variances σ_b^2 and σ_c^2, respectively, for fixed $\sigma_b, \sigma_c \geq 0$. All parameters are mutually independent, shared by all tokens, and sampled independently across layers and between the attention and MLP blocks. Both maps are therefore centered.

For $x, y \in \mathbb{S}^{d-1}$, the MLP covariance is independent of m and satisfies $\mathbb{E}[G_\ell^m(x) G_\ell^m(y)^\top] = \kappa(\langle x, y\rangle) I_d$. For independent $u \sim \mathcal{N}(0, I_d)$ and $b \sim \mathcal{N}(0, \sigma_b^2)$, it is given by

$$\kappa(\langle x, y\rangle) = \mathbb{E}\left[\varphi\left(\frac{\langle u, x\rangle}{\sqrt{d}} + b\right) \varphi\left(\frac{\langle u, y\rangle}{\sqrt{d}} + b\right)\right] + \sigma_c^2.$$

We use RMS normalization $d^{-1/2}$ and no numerical offset, so $\mathsf{RMSNorm}(z) := z/\|z\|$ for $z \neq 0$. This convention keeps the tokens on the unit sphere. Hence, for a network of depth L, setting a timestep $\Delta t := T/L$ for a $T > 0$, the updates of each block are:

$$\begin{cases} Y_\ell^i = \mathsf{RMSNorm}\Big(X_\ell^i + \sqrt{\Delta t}\ \mathrm{Attn}_\ell[X_\ell](X_\ell^i)\Big), \\ X_{\ell+1}^i = \mathsf{RMSNorm}\Big(Y_\ell^i + \sqrt{\Delta t}\, G_\ell^m(Y_\ell^i)\Big), \end{cases} \qquad \ell = 0, \dots, L-1. \tag{3}$$

We now pass to continuous depth and infinite MLP width, keeping N and d fixed. The scaling limit is studied by Koubbi et al. (2026) for the attention block and by Agazzi et al. (2026) for the MLP construction above, including its infinite-width limit. We combine the two limiting noise terms in the system below. The MLP contribution is described by a centered Gaussian Brownian field $B_t(x)$ with covariance

$$\mathbb{E}[B_t(x) B_s(y)^\top] = (t \wedge s)\kappa(\langle x, y\rangle) I_d. \tag{4}$$

For the continuous-time analysis we keep the kernel κ of the MLP above and we assume in addition that $\kappa \in C^2([-1, 1])$.

For the attention contribution, we introduce a cylindrical Wiener process W on $L^2(\rho_*)$, independent of B. Let $P_x := I_d - xx^\top$ be the tangent projection at $x \in \mathbb{S}^{d-1}$.

With this notation, the continuous-depth model studied in this paper is the Itô system described by Koubbi et al. (2026)

$$\begin{aligned} dX_t^i = P_{X_t^i} \int_\Theta \mathrm{Attn}_\theta[X_t](X_t^i) W(d\theta, dt) + P_{X_t^i}\, dB_t(X_t^i) \\ - \frac{1}{2} X_t^i \left[\int_\Theta \|P_{X_t^i} \mathrm{Attn}_\theta[X_t](X_t^i)\|^2 \rho_*(d\theta) + (d-1)\kappa \right] dt, \qquad X_{t=0} = X_0. \end{aligned} \tag{5}$$

Both noises are common to all tokens, as are the parameters in each discrete layer. In other words, all tokens are driven by the same realization of the random field that arises from sampling the parameters. The radial drift comes from normalization and keeps the tokens on the sphere.

To measure the separation of tokens on the configuration space $\mathcal{M} := (\mathbb{S}^{d-1})^N$, define their overlaps and deficits by

$$\rho_{ij}(X) := \langle X^i, X^j \rangle, \qquad u_{ij}(X) := 1 - \rho_{ij}(X) = \tfrac{1}{2}\|X^i - X^j\|^2.$$

Furthermore, defining the largest deficit $M(X) := \max_{i \neq j} u_{ij}(X)$, synchronization means that $M(X_t) \to 0$, or equivalently that the configuration approaches $\Delta_N := \{X \in \mathcal{M} : X^1 = \cdots = X^N\}$. The drift of a scalar function of the configuration along equation 5 is given by the generator $\mathcal{L}$ of the process, and the size of its fluctuations by the carré du champ Γ. Appendix A recalls both and computes them for the system at hand.

3 Main results

Before stating the main result, we present a heuristic argument to find the exponent that describes the asymptotic dynamics near a synchronized state. At initialization the same random maps are applied to every token, so two tokens that coincide at some depth remain equal at every later depth, and the synchronized configurations Δ_N are invariant.

Let $z_0 := X_0^i - X_0^j$ denote the difference between two tokens started an infinitesimal distance apart, with $z_0 \neq 0$. We write z_t for its evolution under the linearization of equation 5 along Δ_N, with the realization of the noises held fixed.

Two constants govern the evolution of z_t and hence the main result. The first is produced by the MLP alone, that is at $\sigma_V = 0$, and depends on the activation function and the initialization variances only through the covariance kernel κ, as computed by Agazzi et al. (2026),

$$\lambda_{\mathrm{MLP}} := \tfrac{1}{2}\big[(d-3)\kappa'(1) - (d-1)\kappa\big]. \tag{6}$$

Negative values of λ_{MLP} correspond to local contraction and positive values to local expansion. The second is the nonpositive contribution of attention to this exponent,

$$\lambda_{\mathrm{ATT}} := -\frac{\sigma_V^2 (d-1)}{2d}. \tag{7}$$

To gain some intuition, writing $\sigma^2 := \kappa + \kappa'(1) + \sigma_V^2/d$, consider the case where the length $\|z_t\|$ is a geometric Brownian motion,

$$d\|z_t\| = \big(\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}} + \tfrac{1}{2}\sigma^2\big)\|z_t\|\, dt + \sigma \|z_t\|\, dB_t, \qquad \text{so} \qquad \|z_t\| = \|z_0\|\, e^{(\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}})t + \sigma B_t},$$

where B is a standard real Brownian motion. Since $B_t/t \to 0$ almost surely, the separation has the exponential growth rate

$$\lim_{t \to \infty} \frac{1}{t} \log \frac{\|z_t\|}{\|z_0\|} = \lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}} \qquad \text{almost surely.} \tag{8}$$

So, in this linearized regime, the two blocks compete through the sign of $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}$.

The next theorem shows that the sign of this exponent also controls the nonlinear dynamics, for configurations that start close to Δ_N.

Theorem 1. *Let $d \geq 3$, and let λ_{MLP} and λ_{ATT} be as in equation 6 and equation 7, respectively.*

(i) Suppose $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}} > 0$*. For every initial configuration* X_0 *and every pair* $i \neq j$ *with* $u_{ij}(X_0) > 0$*,*

$$\mathbb{P}_{X_0}\big(u_{ij}(X_t) \to 0 \text{ as } t \to \infty\big) = 0.$$

(ii) Suppose $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}} < 0$*. For every* $\varepsilon \in (0,1)$*, there exists* $\delta > 0$ *such that, for every initial configuration* X_0 *satisfying* $\max_{i,j} u_{ij}(X_0) < \delta$*,*

$$\mathbb{P}_{X_0}\big(\lim_{t\to\infty} \max_{i,j} u_{ij}(X_t) = 0\big) \geq 1 - \varepsilon.$$

Remark 2. The radius δ in part (ii) may depend on ε and the fixed model parameters, including N.

Remark 3. As a direct consequence of Theorem 1, if $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}} > 0$ a configuration of distinct tokens that start sufficiently close to one another have zero probability of synchronizing. On the other hand, if $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}} < 0$ the probability of synchronization is positive and tends to one as the initial token diameter tends to zero.

Proof sketch. We measure the token separations through powers of their deficits: for $p \in \mathbb{R} \setminus \{0\}$, set

$$V_p(X) := \sum_{i<j} u_{ij}(X)^p. \tag{9}$$

For $p > 0$, this functional vanishes exactly on Δ_N and satisfies $M^p \leq V_p$, so a small sublevel set of V_p consists of configurations whose tokens are all close to one another. When the whole configuration is close to Δ_N, applying Itô's formula to V_p with sufficiently small $p > 0$ gives $\mathcal{L}V_p \leq -c_0 V_p$ when $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}} < 0$, from which a stopped supermartingale argument bounds the probability of leaving the sublevel set and proves convergence on the event of no exit. When $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}} > 0$, the exact overlap formulas give a positive logarithmic drift for the deficit of any sufficiently close pair, uniformly in the other token positions. A calculation with a small negative power of this deficit bounds the probability that it becomes arbitrarily small before the pair leaves this neighbourhood. Since distinct tokens cannot collide at finite times, the Markov property then rules out convergence of any initially distinct pair, from every initial configuration. The full proof is given in Appendix C.

Theorem 1 gives no control once the tokens are far apart. A global counterpart holds under a stronger explicit condition (see Theorem 4 below). Before stating this result, we introduce the quantity

$$\bar{\lambda}_{\mathrm{MLP}} := (d-1)\kappa'(1) - (d-3)\kappa \tag{10}$$

which is the growth coefficient of the expected deficit under the MLP alone, as detailed in Agazzi et al. (2026). The following relation between $\bar{\lambda}_{\mathrm{MLP}}$ and λ_{MLP} holds

$$2\lambda_{\mathrm{MLP}} = \bar{\lambda}_{\mathrm{MLP}} - 2\big(\kappa + \kappa'(1)\big). \tag{11}$$

Moreover, let $\rho(X) := N^{-2}\sum_{i,j} \rho_{ij}(X)$ denote the mean overlap. The mean deficit $1 - \rho(X) = N^{-2}\sum_{i,j} u_{ij}(X)$ is nonnegative and vanishes exactly on Δ_N.

Theorem 4. *Let* $d \geq 3$ *and suppose*

$$c := -\bar{\lambda}_{\mathrm{MLP}} + 2\lambda_{\mathrm{ATT}}\mathbb{E}_H\|H\|_{\mathrm{op}}^2 > 0.$$

Then for every $X_0 \in \mathcal{M}$*,*

$$\mathbb{E}[1 - \rho(X_t)] \leq e^{-ct}\big(1 - \rho(X_0)\big),$$

so $\rho(X_t) \to 1$ *in* L^1*, in probability, and almost surely.*

Remark 5. For the Gaussian initialization in Section 2, $\mathbb{E}_H\|H\|_{\mathrm{op}}^2 \leq d\sigma_H^2$, so it suffices to assume $-\bar{\lambda}_{\mathrm{MLP}} > -2d\lambda_{\mathrm{ATT}}\sigma_H^2 = (d-1)\sigma_V^2\sigma_H^2$.

In addition, the following result establishes that under the same hypothesis of Theorem 4 the almost sure convergence has an exponential rate.

Corollary 6. *Under the assumptions of Theorem 4, for every pair* $i \neq j$ *with* $X_0^i \neq X_0^j$*,* $0 < \gamma < -2(\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}})$*,*

$$e^{\gamma t}\big(1 - \rho_{ij}(X_t)\big) \longrightarrow 0 \qquad \textit{almost surely}.$$

Both Theorem 4 and Corollary 6 are proved in Appendix D

Remark 7. The sufficient condition $c > 0$ in Theorem 4 is not necessary for global synchronization. For example, without attention ($\sigma_V = 0$), a linear MLP with kernel $\kappa(r) = ar + b$, $a, b > 0$, synchronizes almost surely from every initial configuration (Shalova, 2026, Theorem 3.4), although $c = (d-3)b - 2a$ can be nonpositive. Nevertheless, the existence of a gap between these necessary and sufficient thresholds is natural, as local synchronization established by Theorem 1(ii) need not extend to arbitrary initial configurations. Indeed, removing the bias ($b = 0$) gives $\lambda_{\mathrm{MLP}} = -a < 0$, but tokens initially near two antipodal points approach two distinct clusters with high probability (Lemma 25 in Appendix D; see also Engel & Shalova, 2026; Shalova, 2026). Such systems satisfy neither the global sufficient condition of Theorem 4 nor the positive-exponent hypothesis of Theorem 1(i), illustrating that the asymptotic dynamical landscape between these two thresholds can be quite rich.

4 Thresholds as Lyapunov exponents

The parameter $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}$ in Section 3 is not only a synchronization threshold, but it controls the local exponential rate of the distance between two trajectories of the same dynamics with perturbed initial condition. In the language of Dynamical Systems, we refer to these exponents as Lyapunov exponents. We explain what this means, compute the exponents of the model, use them to compare our criteria with those in the literature, and to draw a connection with training. Baxendale (1989) gives the general theory.

Sampling the parameters once produces a flow Φ_t of the configuration space, sending an input X_0 to its representation $X_t = \Phi_t(X_0)$ at depth t. Since equation 5 has sufficiently regular coefficients on a compact space, each Φ_t is a C^1 map (Appendix A). An infinitesimal perturbation w_0 of the input changes the output, to first order, by $D\Phi_t(X_0)w_0$, where $D\Phi_t(X_0)$ is the input-output Jacobian, the object that backpropagation transports.

Definition 8. Fix a realization of the parameters. For an initial configuration X_0 and a nonzero initial perturbation w_0, the Lyapunov exponent in the direction w_0 is $\lambda(X_0, w_0) := \lim_{t\to\infty} t^{-1} \log\big(\|D\Phi_t(X_0)w_0\|/\|w_0\|\big)$, whenever the limit exists.

A perturbation is thus amplified by $e^{\lambda \Delta t}$ per layer. Growth under repeated random maps is multiplicative, so the natural average is $t^{-1}\log$. Surprisingly, although the flow is random the limit is not: by the multiplicative ergodic theorem, in the form proved by Baxendale (1989) for stochastic flows on compact manifolds, the limit exists almost surely under an ergodic measure and takes at most $\dim \mathcal{M}$ deterministic values, the direction w_0 selecting one of them. Two draws of the parameters therefore give different flows with the same exponents, which is what allows us to speak of the rate of a random model.

Because the logarithm is taken before averaging over the parameters, λ describes a single network. Following the terminology of disordered systems, we call it quenched. The signal-propagation literature on transformers averages over the parameters first. For a family of linear maps A_t we therefore compare

$$\lambda(A) := \lim_{t\to\infty} \frac{1}{t} \log \|A_t\|, \qquad \bar{\lambda}(A) := \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}\|A_t\|_F^2, \tag{12}$$

whenever the limits exist, where $\|\cdot\|$ is the operator norm and $\|\cdot\|_F$ the Frobenius norm. The first is the top Lyapunov exponent. The second, the annealed rate, is the growth rate of a second moment. Thus $\bar{\lambda}$ describes an average over initializations, and λ the dynamics of the tokens after a specific realization of the random sampling of the parameters of the network.

The rates also depend on the direction of the perturbation. At $X = (x, \dots, x) \in \Delta_N$, a tangent perturbation $w = (w^1, \dots, w^N)$ splits as

$$w = w^{\parallel} + w^{\perp}, \qquad w^{\parallel} := (\bar{w}, \dots, \bar{w}), \qquad w^{\perp} := (w^1 - \bar{w}, \dots, w^N - \bar{w}), \qquad \bar{w} := \frac{1}{N}\sum_{i=1}^{N} w^i.$$

The longitudinal part $w^{\parallel}$ moves all tokens together. The transverse part $w^{\perp}$ has zero sum and therefore corresponds to perturbations that leave the barycenter/center of mass of the particles fixed.

Along Δ_N, the Jacobian $J_t := D\Phi_t(X_0)$ preserves the longitudinal and transverse tangent spaces, mapping each at X_0 into its counterpart at X_t. We denote the corresponding restrictions by $J_t^{\parallel}$ and $J_t^{\perp}$.

Theorem 9. *Let $X_0 \in \Delta_N$. The quenched limits in equation 12 exist almost surely, are deterministic, and are given by*

$$\lambda(J^{\perp}) = \lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}, \qquad \lambda(J) = \lambda(J^{\parallel}) = \lambda_{\mathrm{MLP}} + \frac{2}{d-1}\lambda_{\mathrm{ATT}}.$$

The annealed limits exist and are given by

$$\bar{\lambda}(J^{\perp}) = \bar{\lambda}_{\mathrm{MLP}} + \frac{2(d-3)}{d-1}\lambda_{\mathrm{ATT}}, \qquad \bar{\lambda}(J) = \bar{\lambda}(J^{\parallel}) = \bar{\lambda}_{\mathrm{MLP}} - \frac{4}{d-1}\lambda_{\mathrm{ATT}}.$$

All four rates are unchanged when each Jacobian is replaced by its transpose.

The proof is given in Appendix E.2. The quenched transverse exponent is the threshold of Theorem 1, namely $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}$. The two quenched exponents differ by

$$\lambda(J^{\parallel}) - \lambda(J^{\perp}) = -\frac{d-3}{d-1}\,\lambda_{\mathrm{ATT}} \;\geq 0, \tag{13}$$

so attention contracts token differences more than their common motion, and the full Jacobian inherits the longitudinal rate: a quantity built from the whole Jacobian does not in general detect the contraction that decides clustering.

We note that, since $\kappa(1) + \kappa'(1) > 0$, the formulas in Theorem 9 imply

$$\bar{\lambda}(J^{\perp}) = 2\lambda(J^{\perp}) + \tfrac{1}{2}\Sigma, \qquad \text{where} \qquad \Sigma := 4\big(\kappa + \kappa'(1)\big) + \frac{4\sigma_V^2}{d} > 0, \tag{14}$$

with an analogous formula holding for $J^{\parallel}$. The gap $\bar{\lambda}(J^{\perp}) - 2\lambda(J^{\perp}) = \Sigma/2 > 0$ in equation 14 can be interpreted as a variance term (or, in the language of stochastic calculus, an Itô correction term) arising from the common noise and the second moment in the definition of the annealed exponent. As a consequence of this inequality[1] the negativity of the annealed exponent $\bar{\lambda}(J^{\perp})$ (annealed decay) implies pathwise decay, and therefore the conclusions of Theorem 1(ii) hold. We further note that this stronger condition is closely connected with Theorem 4: whenever $\sigma_V^2 = 0$ (and hence $\lambda_{\mathrm{ATT}} = 0$), its global synchronization condition coincides with negativity of $\bar{\lambda}_{\mathrm{MLP}} = \bar{\lambda}(J^{\perp})$, as in the L^1 synchronization result of Agazzi et al. (2026).

The distinction between annealed and quenched exponents also separates our results from those of Cowsik et al. (2024, Sections 3.3–3.4), whose angle and gradient exponents correspond in our model to $\bar{\lambda}(J^{\perp})$ and $\bar{\lambda}(J)$: both are annealed, and the second is built from the full Jacobian. Our results quantify the gap between annealed and quenched exponents in the model at hand (equation 14) and show that the pathwise clustering threshold differs, in general, from the annealed threshold they identify.

The longitudinal exponent also connects our results to training. Indeed, consider a residual block at an early depth s with parameters $\vartheta \in \mathbb{R}^q$ and a synchronized state X_s of the residual stream. Let $B := D_\vartheta X_s(\vartheta)$ be the derivative of its output. Writing $J_{t,s}$ for the Jacobian from depth s to depth t, the chain rule gives, for the gradient of the training loss $\mathcal{R} \in C^1(\mathcal{M})$ (depending on the output X_t),

$$\nabla_\vartheta \mathcal{R}(X_t) = B^\top \big(J_{t,s}^{\parallel}\big)^\top \nabla \mathcal{R}(X_t)^{\parallel}, \qquad \|\nabla_\vartheta \mathcal{R}(X_t)\| \leq \|B\|\,\|J_{t,s}^{\parallel}\|\,\|\nabla \mathcal{R}(X_t)\|.$$

As a consequence of the above bound, the parameter gradient of this block vanishes exponentially, almost surely as $t \to \infty$ with s fixed, whenever $\lambda(J^{\parallel}) < 0$, leading to a vanishing gradient problem. This result is formalized in Proposition 32, in Appendix E. We finally note that by equation 13, for $d > 3$ there exists a hyperparameter range where $\lambda(J^{\perp}) < 0 < \lambda(J^{\parallel})$, i.e., where token differences locally contract while common perturbations grow. We expect this regime to be beneficial for training, as it is compatible with both the development of simple, clustered representations of the input context through depth while retaining the sensitivity of these clusters to parameter changes.

[1] whose non-strict form also follows from Jensen's inequality, though without an explicit quantification of the gap

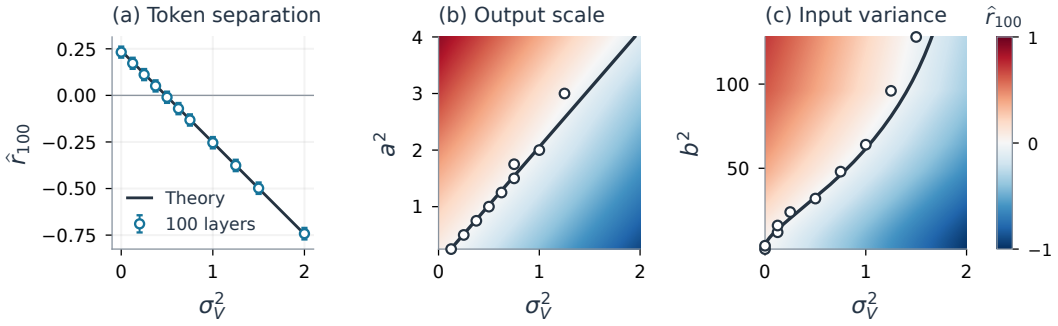


Figure 1: Local separation after 100 layers: (a) $\widehat{r}_L$ at $a^2 = 1$, $b^2 = 32$ (points) and $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}$ (line). (b,c) $\widehat{r}_L$ on grids of (σ_V^2, a^2) and (σ_V^2, b^2). Black curves: $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}} = 0$. Circles: simulated settings whose 95% interval contains zero. Each setting uses 8192 networks.

5 Numerical Experiments

We simulate the discrete transformer equation 3 with $L = 100$ layers and $\Delta t = 0.002$, so that $T = L\Delta t = 0.2$, and with the finite MLP equation 2. Unless stated otherwise, $L = 100$, $\Delta t = 0.002$, $T = L\Delta t = 0.2$, and $\varphi(z) = a\,\mathrm{GELU}(bz)$ with both biases zero. The default hyperparameters are $N = 4$, $d = 64$, $m = 128$, $a^2 = 1$, $b^2 = 32$, $\sigma_V^2 = 0.5$ and $\sigma_H^2 = 1$. Expectations are estimated by averages over independent networks.

Local separation. Let $e_1, \dots, e_d$ denote the standard basis of $\mathbb{R}^d$. We start from $X_0^i = e_1 + 10^{-5} e_{i+1}$, $i = 1, \dots, 4$, and record for each network

$$\frac{1}{2T} \log \frac{u_{12}(X_L)}{u_{12}(X_0)} = \frac{1}{T} \log \frac{\|X_L^1 - X_L^2\|}{\|X_0^1 - X_0^2\|}. \tag{15}$$

Its average over 8192 networks is denoted $\widehat{r}_{100}$.

Figure 1 tests Theorem 1. Panel (a) varies σ_V^2 at the default MLP scales and compares $\widehat{r}_L$ with the separation rate $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}$. Panels (b,c) vary σ_V^2 jointly with a^2 or b^2. They compare the sign of $\widehat{r}_L$ with the threshold $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}} = 0$ of Theorem 1, where λ_{MLP} and λ_{ATT} are given by equation 6 and equation 7. The measured rates agree well with the prediction and reproduce the predicted contraction and expansion as the attention and MLP scales vary.

Global condition. We start from $N = 24$ tokens sampled independently and uniformly on $\mathbb{S}^{d-1}$ and run $L = 1000$ layers with $T = 8/d$ and $\Delta t = T/L$. We use $\varphi = \mathrm{ReLU}$, $m = 1024$, $\sigma_V = 0.5$, $\sigma_H = 1$, and vary $\sigma_b = \sigma_c \in \{0.6, 0.7, \dots, 1.2\}$. Figure 2 compares the mean deficit $\mathbb{E}[1 - \rho(X_L)]$ with the upper bound $e^{-cT}\mathbb{E}[1 - \rho(X_0)]$ from Theorem 4 and Remark 5, at $d = 7, 31, 150$. Expectations are estimated over 256 independent networks.

Jacobian. Start instead from the synchronized configuration $X_0 = (e_1, e_1, e_1, e_1)$. Let J_T be the Jacobian of the 100-layer transformer at X_0, the discrete counterpart of J_t in Section 4. Consider the transverse and longitudinal perturbations

$$w_0^{\perp} = \tfrac{1}{\sqrt{2}}(e_2, -e_2, 0, 0), \qquad w_0^{\parallel} = \tfrac{1}{2}(e_2, e_2, e_2, e_2).$$

Figure 3(a,b) tests Theorem 9. It compares $T^{-1}\mathbb{E}\log\|J_T w_0\|$ with the quenched exponents $\lambda(J^{\perp})$ and $\lambda(J^{\parallel})$, and $T^{-1}\log\mathbb{E}\|J_T w_0\|^2$ with the annealed exponents $\bar{\lambda}(J^{\perp})$ and $\bar{\lambda}(J^{\parallel})$, estimated from 8192 networks.

Backpropagation. With the same input, consider the squared loss $\mathcal{R}(X) = \frac{1}{2}\left\|\frac{X^1 - X^2}{\sqrt{2}} - e_2\right\|^2$, whose gradient at X_ℓ is transverse. Theorem 9 states that the rates are unchanged under transposition, so $\mathbb{E}\log(\|\nabla_{X_0}\mathcal{R}(X_\ell)\|/\|\nabla\mathcal{R}(X_\ell)\|)$ is predicted to be $\lambda(J^{\perp})\ell\Delta t$. We compute $\nabla_{X_0}\mathcal{R}(X_\ell)$ by backpropagation for $\ell = 10, 20, \dots, 100$. Figure 3(c) shows the result for $\sigma_V^2 = 0$ and $\sigma_V^2 = 0.8$, which lie on either side of the threshold.

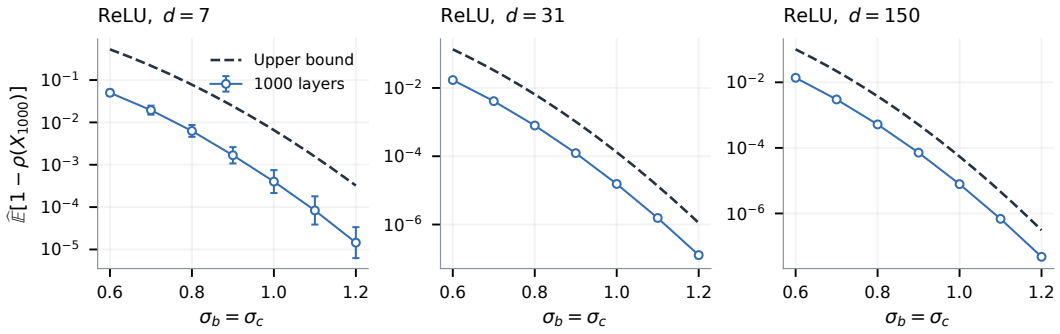


Figure 2: Global contraction with ReLU from 24 random tokens. Left to right: $d = 7, 31, 150$, with $T = 8/d$. Points are measured mean deficit $\mathbb{E}[1 - \rho(X_L)]$ as $\sigma_b = \sigma_c$ varies. Dashed curves are the theoretical upper bound.

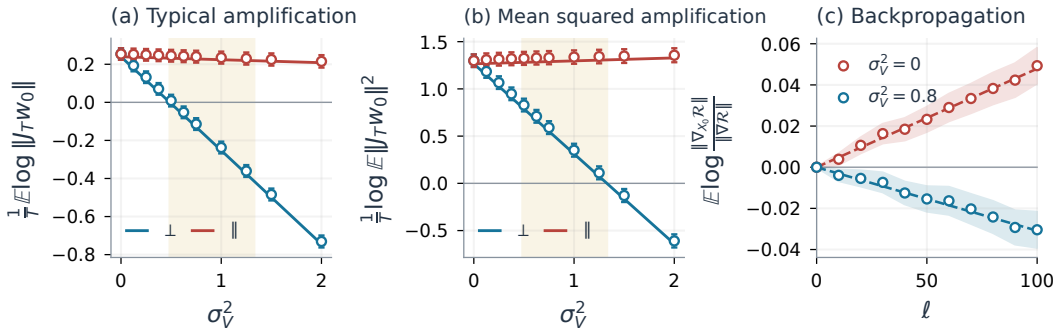


Figure 3: Quenched, annealed and backpropagated rates: (a,b) $T^{-1}\mathbb{E}\log\|J_T w_0\|$ and $T^{-1}\log\mathbb{E}\|J_T w_0\|^2$ for $w_0 = w_0^{\perp}$ and $w_0 = w_0^{\parallel}$, over 8192 networks. Lines are the predicted quenched (a) and annealed (b) exponents, while the shaded band is $\lambda(J^{\perp}) < 0 < \bar{\lambda}(J^{\perp})$. (c) evolution over layers of gradient norms of $\mathcal{R}$, over 4096 networks. Dashed lines are $\lambda(J^{\perp})\ell\Delta t$.

6 Conclusion

In this work we studied clustering and information propagation in deep transformers at initialization. Our analysis establishes a sharp threshold for local clustering with high probability, together with sufficient conditions for global clustering. It also computes the typical and averaged growth rates of two kinds of sensitivity: to transverse perturbations, which separate tokens within a cluster, and to longitudinal perturbations, which move them collectively.

These results refine earlier propagation diagnostics. They identify hyperparameter regimes in which typical and averaged sensitivities behave in qualitatively different ways, and likewise for transverse and longitudinal perturbations. Through the chain rule, longitudinal contraction further yields a vanishing-gradient criterion for the parameters of deep architectures. Taken together, these distinctions can guide the choice of initialization hyperparameters, since they separate three effects: representation collapse, collective sensitivity, and the suppression of parameter learning.

Limitations and future work. Our analysis relies on Gaussian initialization and on a continuous-depth, infinite-width model. Adapting quantitative convergence estimates offers a natural route toward finite-depth guarantees. Our results also point to several exciting directions: sharp criteria for global clustering, the evolution of the exponents during training, the computation of Lyapunov exponents for non-synchronized invariant measures, and identifying when the resulting initialization guidance improves learning in practice. More generally, we hope this work helps build a bridge between the theory of random dynamical systems and the design of deep architectures.

AI USE STATEMENT

Generative AI tools were used in this work to review mathematical proofs for potential errors or gaps, to assist with drafting and editing the manuscript, and to assist with writing code used in the numerical experiments. All AI-assisted outputs were reviewed by the authors before being incorporated into the work. The authors take full responsibility for the correctness of the mathematical arguments, code, experimental results, and final content of the paper.

REPRODUCIBILITY STATEMENT

The theoretical results are proved in the appendices. Section 2 specifies the discrete updates and the limiting stochastic model. The local conditions can be evaluated from the coefficients in equation 25 and equation 7, while the global condition uses the constant in equation 29. The appendices provide the overlap calculations, the estimates of their remainders, and the complete stopping and linearization arguments used in the main results. The assumptions under which these calculations apply are stated in Section 2, and the cases they leave open in Section 6.

REFERENCES

Andrea Agazzi, Giuseppe Bruno, Eloy Mosig García, Samuele Saviozzi, and Marco Romito. Stochastic scaling limits and synchronization by noise in deep transformer models, 2026. URL https://arxiv.org/abs/2604.26898.

Albert Alcalde, Leon Bungert, Konstantin Riedl, and Tim Roith. Quantifying concentration phenomena of mean-field transformers in the low-temperature regime. *arXiv preprint arXiv:2605.10931*, 2026.

P. H. Baxendale. Statistical equilibrium and two-point motion for a stochastic flow of diffeomorphisms. In *Spatial Stochastic Processes*, volume 19 of *Progr. Probab.*, pp. 189–218. Birkhäuser Boston, Boston, MA, 1991. doi: 10.1007/978-1-4612-0451-0_9.

Peter H. Baxendale. Lyapunov exponents and relative entropy for a stochastic flow of diffeomorphisms. *Probability Theory and Related Fields*, 81:521–554, 1989. doi: 10.1007/BF00367301. URL https://doi.org/10.1007/BF00367301.

Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. Emergence of meta-stable clustering in mean-field transformer models. In *International Conference on Learning Representations*, 2025a.

Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. A multiscale analysis of mean-field transformers in the moderate interaction regime. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-4443.

Martin Burger, Samira Kabri, Yury Korolev, Tim Roith, and Lukas Weigand. Analysis of mean-field models arising from self-attention dynamics in transformer architectures with layer normalization. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 383(2298):20240233, 2025. doi: 10.1098/rsta.2024.0233.

Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers. *Foundations of Computational Mathematics*, 2026. Accepted for publication.

Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Quantitative clustering in mean-field transformer models. *arXiv preprint arXiv:2504.14697*, 2025.

Aditya Cowsik, Tamra Nebabu, Xiao-Liang Qi, and Surya Ganguli. Geometric dynamics of signal propagation predict trainability of transformers, 2024. URL https://arxiv.org/abs/2403.02579.

Michael Cranston, Benjamin Gess, and Michael Scheutzow. Weak synchronization for isotropic flows. *Discrete and Continuous Dynamical Systems - B*, 21(9):3003–3014, 2016. doi: 10.3934/dcdsb.2016084.

Maximilian Engel and Anna Shalova. Random quadratic form on a sphere: Synchronization by common noise. *arXiv preprint arXiv:2603.06187*, 2026. URL https://arxiv.org/abs/2603.06187.

Lev Fedorov, Michaël E. Sander, Romuald Elie, Pierre Marion, and Mathieu Laurière. Clustering in deep stochastic transformers. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://openreview.net/forum?id=opfSL15Uc4.

Franco Flandoli, Benjamin Gess, and Michael Scheutzow. Synchronization by noise. *Probability Theory and Related Fields*, 168(3):511–556, 2017. doi: 10.1007/s00440-016-0716-2. URL https://doi.org/10.1007/s00440-016-0716-2.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36:57026–57037, 2023.

Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025.

H. Koubbi, B. Geshkovski, and P. Rigollet. Homogenized transformers. *arXiv preprint arXiv:2604.01978*, 2026.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurélien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27198–27211, 2022.

Michael E. Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3515–3530. PMLR, 2022.

Anna Shalova. Random quadratic form with random forcing: Metastable synchronization by noise. *arXiv preprint arXiv:2608.16664*, 2026. URL https://arxiv.org/abs/2608.16664.

A THE ACTION OF THE GENERATOR ON THE OVERLAPS

In this appendix we recall the generator and the carré du champ, justify the regularity of the stochastic flow used in Section 4, and identify the MLP covariance of Section 2.

Generator and carré du champ. For a smooth $f : \mathcal{M} \to \mathbb{R}$, the generator of equation 5 is

$$\mathcal{L}f(X) := \lim_{h \downarrow 0} \frac{1}{h}\, \mathbb{E}\big[f(X_h) - f(X) \mid X_0 = X\big],$$

so that $f(X_t) - \int_0^t \mathcal{L}f(X_s)\, ds$ is a martingale and $\mathcal{L}f$ is the drift of $f(X_t)$. The carré du champ $\Gamma(f,g) := \mathcal{L}(fg) - f\mathcal{L}g - g\mathcal{L}f$ measures the fluctuation around that drift: the martingale part of $f(X_t)$ has quadratic variation $\int_0^t \Gamma(f,f)(X_s)\, ds$. For a finite-dimensional diffusion $dX = b\, dt + \sigma\, dW$ these are $\mathcal{L}f = \langle b, \nabla f\rangle + \frac{1}{2}\mathrm{Tr}(\sigma\sigma^\top \nabla^2 f)$ and $\Gamma(f,f) = \|\sigma^\top \nabla f\|^2$. The attention noise in equation 5 is indexed by θ rather than by finitely many coordinates, but it enters $\mathcal{L}$ and Γ only through its covariance, the sum over noise coordinates being replaced by an integral against ρ_*. Lemma 11 computes that covariance. We write $\mathcal{L}_{\mathrm{MLP}}, \Gamma_{\mathrm{MLP}}$ and $\mathcal{L}_{\mathrm{hom}}, \Gamma_{\mathrm{hom}}$ for the contributions of the two noises.

Regularity of the flow. Let $h(x)$ be a Hilbert-space representation of the scalar covariance, so that $\langle h(x), h(y)\rangle = \kappa(\langle x, y\rangle)$. For $u \in T_x\mathbb{S}^{d-1}$ and $v \in T_y\mathbb{S}^{d-1}$, differentiating this identity gives

$$\langle D_u h(x), D_v h(y)\rangle = \kappa'(\langle x, y\rangle)\langle u, v\rangle + \kappa''(\langle x, y\rangle)\langle u, y\rangle\langle x, v\rangle.$$

In a smooth local tangent frame, the relation $1 - \langle x, y\rangle = O(\|x - y\|^2)$ and the boundedness of κ'' show that the squared Hilbert-norm increment of these derivatives is $O(\|x-y\|^2)$. Thus the projected MLP coefficient has locally Lipschitz first derivatives. For the attention coefficient, differentiating the softmax weights gives the corresponding bounds in $L^2(\rho_*)$, since the Gaussian parameter law has the required moments. The usual differentiable-flow theorem for Itô equations therefore gives a C^1 flow, which is defined for all times because the state space is compact.

The MLP covariance. The following lemma justifies the covariance formula of Section 2 and records the structure of κ used in Lemma 18 and Remark 23.

Lemma 10. *For the block equation 2, $\mathbb{E}[G_\ell^m(x)G_\ell^m(y)^\top] = \kappa(\langle x, y\rangle)I_d$ with κ as in Section 2, independently of m. Moreover $\kappa(r) = \sum_{n\geq 0} k_n r^n$ with $k_n \geq 0$ and $\sum_n k_n < \infty$. If φ is odd and $\sigma_b = \sigma_c = 0$, then κ is odd. If $\sigma_b > 0$ or $\sigma_c > 0$, and φ is not almost everywhere zero, then $\kappa(1) + \kappa(-1) > 0$. Moreover $\sum_n nk_n = d^{-1}\mathbb{E}[\varphi'(\sqrt{v}\,Z)^2] \leq \mathrm{Lip}(\varphi)^2/d < \infty$, where $Z \sim \mathcal{N}(0,1)$ and $v = 1/d + \sigma_b^2$. In particular $\kappa \in C^1([-1,1])$, its derivative is given by the termwise differentiated series, and $\kappa'(1) = d^{-1}\mathbb{E}[\varphi'(\sqrt{v}\,Z)^2]$.*

Proof. Write $G_\ell^m(x) = m^{-1/2}A\,\varphi(g(x)) + c$ with $g(x) := Ux/\sqrt{d} + b$. The entries of A are independent standard Gaussians, independent of g, so $\mathbb{E}[(Aa)(Aa')^\top \mid g] = \langle a, a'\rangle I_d$, while the cross terms with c vanish. Hence $\mathbb{E}[G_\ell^m(x)G_\ell^m(y)^\top] = m^{-1}\sum_{k=1}^m \mathbb{E}[\varphi(g_k(x))\varphi(g_k(y))]I_d + \sigma_c^2 I_d$, and the m summands are equal, which gives the stated formula. Each $(g_k(x), g_k(y))$ is a centered Gaussian pair with variances $v := 1/d + \sigma_b^2$ and covariance $r/d + \sigma_b^2$, $r = \langle x, y\rangle$. Expanding $z \mapsto \varphi(\sqrt{v}z)$ in normalized Hermite polynomials with coefficients a_n, which is possible because a Lipschitz φ is square integrable under the Gaussian, Mehler's formula gives

$$\kappa(r) = \sigma_c^2 + \sum_{n\geq 0} a_n^2 \Big(\frac{r/d + \sigma_b^2}{v}\Big)^n.$$

Expanding each power binomially produces only nonnegative coefficients in r, and the series converges at $r = 1$, where it equals $\sigma_c^2 + \sum_n a_n^2 < \infty$. This proves the expansion. If φ is odd, $a_n = 0$ for even n, and with $\sigma_b = \sigma_c = 0$ every term is an odd power of r. Finally $\kappa(1) + \kappa(-1) = 2\sigma_c^2 + \sum_n a_n^2(1 + \varrho^n)$ with $\varrho := (\sigma_b^2 - 1/d)/v \in (-1, 1]$. Every term is nonnegative, the first is positive if $\sigma_c > 0$, and if $\sigma_b > 0$ then $|\varrho| < 1$, so every n with $a_n \neq 0$ contributes positively.

For the derivative, let $f(z) := \varphi(\sqrt{v}\,z)$ and let h_n be the normalized Hermite polynomials, so that $a_n = \mathbb{E}[f(Z)h_n(Z)]$. Since φ is Lipschitz, f is absolutely continuous, with $|f'| \leq \sqrt{v}\,\mathrm{Lip}(\varphi)$ almost everywhere and at most linear growth. The polynomials satisfy $h_n' = \sqrt{n}\,h_{n-1}$ and $zh_{n-1}(z) - h_{n-1}'(z) = \sqrt{n}\,h_n(z)$. Gaussian integration by parts, $\mathbb{E}[g'(Z)] = \mathbb{E}[Zg(Z)]$ for absolutely continuous g with $\mathbb{E}|g'(Z)| < \infty$, applied to $g = fh_{n-1}$, gives

$$\mathbb{E}[f'(Z)h_{n-1}(Z)] = \mathbb{E}\big[f(Z)\big(Zh_{n-1}(Z) - h_{n-1}'(Z)\big)\big] = \sqrt{n}\,a_n, \qquad n \geq 1.$$

Since $f' \in L^2(\gamma)$ and (h_n) is an orthonormal basis of $L^2(\gamma)$, Parseval's identity yields

$$\sum_{n\geq 1} na_n^2 = \mathbb{E}[f'(Z)^2] = v\,\mathbb{E}[\varphi'(\sqrt{v}\,Z)^2] \leq v\,\mathrm{Lip}(\varphi)^2.$$

Now write $(r/d + \sigma_b^2)/v = \alpha r + \beta$ with $\alpha := 1/(dv)$ and $\beta := \sigma_b^2/v$, so that $\alpha, \beta \geq 0$ and $\alpha + \beta = 1$. The binomial expansion gives $k_m = \sum_{n\geq m} a_n^2 \binom{n}{m}\alpha^m\beta^{n-m}$ for $m \geq 1$, and, all terms being nonnegative, Tonelli's theorem yields

$$\sum_{m\geq 1} mk_m = \sum_{n\geq 1} a_n^2 \sum_{m=1}^n m\binom{n}{m}\alpha^m\beta^{n-m} = \sum_{n\geq 1} a_n^2\, n\alpha(\alpha+\beta)^{n-1} = \alpha\sum_{n\geq 1} na_n^2 = \frac{1}{d}\,\mathbb{E}[\varphi'(\sqrt{v}\,Z)^2].$$

Finally, $|nk_n r^{n-1}| \leq nk_n$ on $[-1,1]$, so the differentiated series converges uniformly there by the Weierstrass test, which gives $\kappa \in C^1([-1,1])$ with $\kappa'(r) = \sum_n nk_n r^{n-1}$. □

To compute the overlap generator, we write the attention field as the value matrix applied to a weighted average of the tokens. For $\theta = (V, H)$ and a configuration X, set

$$m_i(\theta, X) := \mathrm{Attn}_\theta[X](X^i) = V\tilde{m}_i, \qquad \tilde{m}_i := \sum_k w_k^i X^k, \qquad w_k^i \propto e^{\langle HX^i, X^k\rangle}, \tag{16}$$

where $\sum_k w_k^i = 1$. We introduce the following quantities

$$s_{ij} = \sigma_V^2\,\mathbb{E}_H\big[\langle \tilde{m}_i, \tilde{m}_j\rangle\big], \qquad s_i := s_{ii}, \tag{17}$$

$$\delta_{ij} := s_i + s_j - 2s_{ij} = \sigma_V^2\mathbb{E}_H\|\tilde{m}_i - \tilde{m}_j\|^2. \tag{18}$$

The following identities express the second moments in terms of these weights and the pairwise deficits.

Lemma 11. *The following identities hold:*

$$\mathbb{E}_\theta[m_i m_j^\top] = \frac{s_{ij}}{d} I_d,$$

$$\|\tilde{m}_i\|^2 = 1 - \sum_{k,l} w_k^i w_l^i \, u_{kl}, \qquad \langle \tilde{m}_i, \tilde{m}_j \rangle = 1 - \sum_{k,l} w_k^i w_l^j \, u_{kl}. \tag{19}$$

Proof. We first condition on H, so that the weighted averages are fixed when taking the expectation over V. Since V has i.i.d. $\mathcal{N}(0, \sigma_V^2/d)$ entries and is independent of H, for fixed p, q we have $\mathbb{E}_V[Vpq^\top V^\top] = \frac{\sigma_V^2}{d}\langle p, q\rangle I_d$. Applying this identity to the two weighted averages and then taking the expectation over H gives the stated covariance. To obtain equation 19, we expand $\langle \tilde{m}_i, \tilde{m}_j \rangle = \sum_{k,l} w_k^i w_l^j \rho_{kl} = \sum_{k,l} w_k^i w_l^j (1 - u_{kl})$ and use $\sum_k w_k^i = \sum_l w_l^j = 1$. The identity for the squared norm is the case in which the two weighted averages coincide. □

We define the following coefficients

$$b_\kappa(r) = -\kappa(d-1)r + \kappa(r)(d - 2 + r^2), \qquad q_\kappa(r) = 2(1 - r^2)\big(\kappa - r\kappa(r)\big). \tag{20}$$

Lemma 12. *For $i \neq j$, the MLP contribution satisfies*

$$\mathcal{L}_{\mathrm{MLP}} \rho_{ij} = b_\kappa(\rho_{ij}), \qquad \Gamma_{\mathrm{MLP}}(\rho_{ij}, \rho_{ij}) = q_\kappa(\rho_{ij}),$$

where b_κ, q_κ are defined in equation 20. The attention contribution is

$$\mathcal{L}_{\mathrm{hom}} \rho_{ij} = \frac{1}{2d}\Big[(1 - \rho_{ij})(d - 2 - \rho_{ij})(s_i + s_j) - (d - 2 + \rho_{ij}^2)\, \delta_{ij}\Big], \tag{21}$$

$$\Gamma_{\mathrm{hom}}(\rho_{ij}, \rho_{ij}) = \frac{1 - \rho_{ij}^2}{d}\big(\delta_{ij} + 2(1 - \rho_{ij}) s_{ij}\big). \tag{22}$$

Since the MLP and attention noises have no cross-variation, their contributions add to give $\mathcal{L}$ and Γ.

Proof. For the MLP contribution, we apply Itô's formula to $\langle X^i, X^j \rangle$, using the covariance $\kappa(\rho_{ij}) I_d$ between the two Gaussian increments and κI_d for each individual increment. The projection identities $\mathrm{Tr}(P_{X^i} P_{X^j}) = d - 2 + \rho_{ij}^2$ and $\mathrm{Tr}(P_{X^i}) = d - 1$ then give b_κ and q_κ in equation 20. For attention, the same calculation uses the second moments from Lemma 11 and gives

$$\mathcal{L}_{\mathrm{hom}} \rho_{ij} = \frac{1}{d}(d - 2 + \rho_{ij}^2) s_{ij} - \frac{d-1}{2d} \rho_{ij}(s_i + s_j),$$

where the two coefficients follow from the same projection identities. Substituting $s_{ij} = \frac{1}{2}(s_i + s_j - \delta_{ij})$ and factoring $d - 2 + r^2 - (d-1)r = (1 - r)(d - 2 - r)$ then yields equation 21.

To compute the quadratic variation, we write the martingale coefficient of the overlap as $\Psi_{ij} = \langle P_{X^i} X^j, m_i \rangle + \langle P_{X^j} X^i, m_j \rangle$, whose squared expectation contributes $\mathbb{E}_\theta[\Psi_{ij}^2]$. The Gaussian identity $\mathbb{E}_V[\langle p, Vq\rangle \langle p', Vq'\rangle] = \frac{\sigma_V^2}{d}\langle p, p'\rangle\langle q, q'\rangle$ reduces this expectation to inner products of the projected tokens. Using $\|P_{X^i} X^j\|^2 = \|P_{X^j} X^i\|^2 = 1 - \rho_{ij}^2$ and $\langle P_{X^i} X^j, P_{X^j} X^i \rangle = -\rho_{ij}(1 - \rho_{ij}^2)$, we obtain

$$\Gamma_{\mathrm{hom}}(\rho_{ij}, \rho_{ij}) = \frac{1 - \rho_{ij}^2}{d}\big(s_i + s_j - 2\rho_{ij} s_{ij}\big).$$

The identity $s_i + s_j - 2\rho_{ij} s_{ij} = \delta_{ij} + 2(1 - \rho_{ij}) s_{ij}$ now gives equation 22. □

B Proof of the pairwise boundary expansion

In this Appendix, we will prove Lemma 14, which gives an expansion near synchronization of the action of both the generator $\mathcal{L}$ and the carré-du-champ Γ on the distance between tokens.

Lemma 13. *Recall equation 17 and equation 18. For every $X \in \mathcal{M}$ and every $i \neq j$,*

$$\big|s_i - \sigma_V^2\big| \leq \sigma_V^2 M(X), \qquad \big|s_{ij} - \sigma_V^2\big| \leq \sigma_V^2 M(X), \qquad \delta_{ij} \leq 16\, L_{\mathrm{att}}^2 \, u_{ij}(X)\, M(X),$$

with

$$L_{\mathrm{att}}^2 := \sigma_V^2 \mathbb{E}_H \|H\|_{\mathrm{op}}^2. \tag{23}$$

Proof. The first two estimates follow by averaging equation 19: the weights are nonnegative and sum to one, while every deficit satisfies $u_{kl} \le M$. For the third estimate, we use the cancellation $\sum_k (w_k^i - w_k^j) = 0$ to write $\tilde{m}_i - \tilde{m}_j = \sum_k (w_k^i - w_k^j)(X^k - X^i)$. This centers the token positions at X^i and gives

$$\|\tilde{m}_i - \tilde{m}_j\| \le \Big(\max_k \|X^k - X^i\|\Big) \sum_k \big|w_k^i - w_k^j\big| \le \sqrt{2M} \sum_k \big|w_k^i - w_k^j\big|.$$

It remains to bound the change in the weights by the distance between the two queries. The weights are softmaxes of the logit vectors $a_k = \langle HX^i, X^k\rangle$ and $b_k = \langle HX^j, X^k\rangle$. Along the interpolation $w(\theta) := \mathrm{softmax}(a + \theta(b-a))$, set $c := b - a$. Differentiating yields $\frac{d}{d\theta} w_k = w_k(c_k - \sum_l w_l c_l)$, from which $\sum_k |\frac{d}{d\theta} w_k| \le 2\|c\|_\infty$ and hence $\sum_k |w_k^i - w_k^j| \le 2\|c\|_\infty$. Since $c_k = \langle H(X^j - X^i), X^k\rangle$ and $\|X^k\| = 1$, we have

$$\|c\|_\infty \le \|H\|_{\mathrm{op}} \|X^i - X^j\| = \|H\|_{\mathrm{op}} \sqrt{2u_{ij}}.$$

Combining the weight estimate with the centered representation gives $\|\tilde{m}_i - \tilde{m}_j\| \le 4\|H\|_{\mathrm{op}} \sqrt{M u_{ij}}$. Squaring and taking $\sigma_V^2 \mathbb{E}_H$ proves the remaining bound. □

Lemma 14. *Let $d \ge 3$. There exist $M_* \in (0,1)$ and $C \ge 1$, depending only on d, σ_V, σ_H and the kernel κ, such that for every $X \in \mathcal{M}$ with $M(X) \le M_*$ and every $i \ne j$,*

$$\begin{aligned} \big|\mathcal{L} u_{ij}(X) + \Lambda^*\, u_{ij}(X)\big| &\le C\, u_{ij}(X)\, M(X), \\ \big|\Gamma(u_{ij}, u_{ij})(X) - \Sigma\, u_{ij}(X)^2\big| &\le C\, u_{ij}(X)^2\, M(X), \end{aligned} \tag{24}$$

where

$$\Lambda^* := \kappa(d-3) - \kappa'(1)(d-1) + \frac{\sigma_V^2}{d}(d-3), \qquad \Sigma := 4\big(\kappa + \kappa'(1)\big) + \frac{4\sigma_V^2}{d}. \tag{25}$$

In particular $\Sigma \ge 4\kappa > 0$.

Proof. Fix a pair $i \ne j$ and write $u := u_{ij}$ and $M := M(X)$, so that $u \le M$. Throughout the proof, C denotes a constant depending only on d, σ_V, σ_H and the kernel κ, and its value may change from line to line. We first expand the overlap drift, using the preceding lemma to replace the attention moments by their values at synchronization.

For the MLP part, $b_\kappa(1) = 0$ and $b_\kappa'(1) = \bar{\lambda}_{\mathrm{MLP}}$ give $b_\kappa(1-u) = -\bar{\lambda}_{\mathrm{MLP}} u + O(u^2) = a_\kappa u + O(u^2)$, where $a_\kappa := \kappa(d-3) - \kappa'(1)(d-1)$. The remainder has the required form because $u^2 \le uM$. For attention, substituting $1 - \rho_{ij} = u$ and $d - 2 - \rho_{ij} = d - 3 + u$ into equation 21 gives

$$\mathcal{L}_{\mathrm{hom}} \rho_{ij} = \frac{1}{2d}\big[u(d-3+u)(s_i + s_j) - (d - 2 + \rho_{ij}^2)\delta_{ij}\big].$$

Lemma 13 bounds the error in replacing $s_i + s_j$ by its synchronized value through $|s_i + s_j - 2\sigma_V^2| \le 2\sigma_V^2 M$. Using this estimate together with the bound on δ_{ij} and $d - 2 + \rho_{ij}^2 \le d - 1$, we obtain

$$\Big|\mathcal{L}_{\mathrm{hom}} \rho_{ij} - \frac{\sigma_V^2(d-3)}{d} u\Big| \le \frac{1}{2d}\Big[u(d-3)\, 2\sigma_V^2 M + u^2(s_i + s_j) + (d-1)\, 16 L_{\mathrm{att}}^2 u M\Big] \le C\, uM.$$

Adding the MLP contribution and using $\mathcal{L} u_{ij} = -\mathcal{L}\rho_{ij}$ gives the first bound in equation 24, since $\Lambda^* = a_\kappa + \sigma_V^2(d-3)/d$.

We next turn to the quadratic variation, where the leading term is quadratic in the deficit. Expanding $\kappa(1-u) = \kappa - \kappa'(1)u + O(u^2)$ in equation 20 gives $q_\kappa(1-u) = 4(\kappa + \kappa'(1))u^2 + O(u^3)$ for the MLP part, with $u^3 \le u^2 M$. For attention, the identity $1 - \rho_{ij}^2 = u(2-u)$ rewrites equation 22 as

$$\Gamma_{\mathrm{hom}}(\rho_{ij}, \rho_{ij}) = \frac{1}{d} u(2-u)\big(\delta_{ij} + 2u\, s_{ij}\big).$$

Replacing s_{ij} by its synchronized value and estimating δ_{ij} by Lemma 13, we find

$$\Big|\Gamma_{\mathrm{hom}}(\rho_{ij}, \rho_{ij}) - \frac{4\sigma_V^2}{d} u^2\Big| \le \frac{1}{d}\Big[2u\, 16 L_{\mathrm{att}}^2 u M + 4u^2 \sigma_V^2 M + 2u^3(\sigma_V^2 + \sigma_V^2 M)\Big] \le C\, u^2 M.$$

The two noises have no cross-variation, and $\Gamma(u_{ij}, u_{ij}) = \Gamma(\rho_{ij}, \rho_{ij})$, so summing the two expansions gives the second bound in equation 24 with $\Sigma = 4(\kappa + \kappa'(1)) + 4\sigma_V^2/d$. The assumptions $\kappa'(1) \ge 0$ and $\kappa = \kappa(1) > 0$ also give $\Sigma \ge 4\kappa > 0$.

The constants are independent of N, since the bounds on the softmax weights use only their nonnegativity and unit sum. □

C PROOFS FOR THE LOCAL DICHOTOMY

The proof of Theorem 1(ii) uses the pairwise expansion near Δ_N, while part (i) uses the exact overlap formulas for a selected pair, uniformly in the other token positions. We use positive powers to obtain decay near synchronization and negative powers to prevent a positive deficit from tending to zero. To keep the sign determining this choice explicit, set

$$h_0 := \Lambda^* + \tfrac{1}{2}\Sigma = -2(\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}),$$

where the identities follow from Proposition 15. Let C, M_* be the constants in Lemma 14, and for $\delta \in (0, M_*]$ write $\mathcal{U}_\delta := \{X \in \mathcal{M} : M(X) < \delta\}$.

For $p \in \mathbb{R} \setminus \{0\}$ and $u_{ij} > 0$, Itô's formula gives the identity used in both arguments:

$$\mathcal{L}\big(u_{ij}^p\big) = p\, u_{ij}^p \Big(\frac{\mathcal{L}u_{ij}}{u_{ij}} - \frac{1-p}{2}\, \frac{\Gamma(u_{ij}, u_{ij})}{u_{ij}^2} \Big). \tag{26}$$

For $p > 0$, initially coincident pairs contribute zero, since pathwise uniqueness and the common noises keep these pairs coincident. Lemma 20 ensures that every other deficit stays positive at finite times. We may therefore apply the generator identities after localization away from zero deficits and then remove the localization. At configurations with coincident tokens, the inequalities are understood on the corresponding invariant collision stratum.

Proposition 15. *With λ_{MLP} and λ_{ATT} as in equation 6 and equation 7,*

$$-\tfrac{1}{2}\Big(\Lambda^* + \tfrac{1}{2}\Sigma\Big) = \lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}. \tag{27}$$

Proof. Substituting the coefficients from equation 25, we obtain

$$2\Lambda^* + \Sigma = 2(d-1)\kappa - 2(d-3)\kappa'(1) + \frac{2\sigma_V^2(d-1)}{d} = -4(\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}).$$

Dividing by -4 gives equation 27. □

C.1 CLUSTERING

When $h_0 > 0$, taking p small in equation 26 leaves a strictly negative leading coefficient. We first show that this coefficient dominates the remainders throughout a sufficiently small neighbourhood, with a choice of p that is the same for all pairs.

Lemma 16. *Suppose $h_0 > 0$. Then there exist $p \in (0,1)$, $\delta \in (0, M_*]$ and $c_0 > 0$ such that*

$$\mathcal{L}V_p(X) \le -c_0\, V_p(X) \qquad \textit{for every } X \in \mathcal{U}_\delta.$$

Proof. Fix $X \in \mathcal{U}_\delta$ and a pair $i \neq j$ with $u_{ij} > 0$. Since the quadratic-variation term in equation 26 has a negative coefficient for $p \in (0,1)$, we use the upper drift bound and the lower quadratic-variation bound from Lemma 14:

$$\frac{\mathcal{L}u_{ij}}{u_{ij}} \le -\Lambda^* + CM, \qquad \frac{\Gamma(u_{ij}, u_{ij})}{u_{ij}^2} \ge \Sigma - CM.$$

Substituting these bounds into the power identity, and using $1 - p > 0$, gives

$$\mathcal{L}\big(u_{ij}^p\big) \le p\, u_{ij}^p \Big(-\Lambda^* - \tfrac{1}{2}\Sigma + \tfrac{p}{2}\Sigma + CM\big(1 + \tfrac{1-p}{2}\big)\Big) \le p\, u_{ij}^p \Big(-h_0 + \tfrac{p}{2}\Sigma + 2CM \Big).$$

We first choose $p \in (0,1)$ so that $\frac{p}{2}\Sigma \le \frac{h_0}{4}$, and then $\delta \in (0, M_*]$ so that $2C\delta \le \frac{h_0}{4}$. The bracket is consequently at most $-\frac{h_0}{2}$ on $\mathcal{U}_\delta$, which gives $\mathcal{L}(u_{ij}^p) \le -\frac{ph_0}{2} u_{ij}^p$. Pairs with $u_{ij} = 0$ remain coalesced and contribute nothing, so summing over all pairs proves the lemma with $c_0 := ph_0/2$. □

As $p \downarrow 0$, the condition used in this proof approaches $\Lambda^* + \Sigma/2 > 0$, retaining the quadratic-variation contribution. For comparison, setting $p = 1$ removes this contribution from equation 26, and the same calculation then requires $\Lambda^* > 0$.

Proof of Theorem 1(ii). By Proposition 15, the assumption $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}} < 0$ is equivalent to $h_0 > 0$. Let p, δ, c_0 be given by Lemma 16, and set $\eta_0 := \delta^p$. For $0 < \eta \le \eta_0$ and X_0 satisfying $V_p(X_0) < \eta$, define $\tau := \inf\{t \ge 0 : V_p(X_t) \ge \eta\}$. Since $M(X)^p \le V_p(X)$, the inequality $V_p(X) < \eta \le \delta^p$ implies $M(X) < \delta$. Thus $\{V_p < \eta\} \subseteq \mathcal{U}_\delta$, and the drift bound of Lemma 16 applies throughout $[0, \tau)$.

We apply Itô's formula up to τ, with an additional localization that keeps every initially positive deficit away from zero. Initially coincident pairs remain coincident and contribute zero. By Lemma 20, the additional stopping times tend to infinity, leaving $V_p(X_{t\wedge\tau})$ as a nonnegative local supermartingale and hence a supermartingale. Doob's maximal inequality gives

$$\mathbb{P}_{X_0}\Big(\sup_{t\ge 0} V_p(X_{t\wedge\tau}) \ge \eta\Big) \le \frac{V_p(X_0)}{\eta}.$$

By continuity, $V_p(X_\tau) = \eta$ on $\{\tau < \infty\}$, so

$$\mathbb{P}_{X_0}(\tau = \infty) \ge 1 - \frac{V_p(X_0)}{\eta}.$$

To prove convergence on the non-exit event, we retain the strictly negative drift in the same localized calculation. Fatou's lemma gives

$$\mathbb{E}_{X_0}\left[\int_0^\tau V_p(X_s)\, ds\right] \le c_0^{-1} V_p(X_0) < \infty,$$

and therefore $\int_0^\tau V_p(X_s)\, ds < \infty$ almost surely. The stopped supermartingale also converges almost surely to a finite limit. On $\{\tau = \infty\}$, this limit must be zero, since a positive limit would contradict the finiteness of the time integral. Since $u_{ij}(X_t)^p \le V_p(X_t)$ for every pair, all deficits tend to zero on the same event.

Given $\varepsilon \in (0, 1)$, fix $\eta = \eta_0$ and define

$$\delta_\varepsilon := \left(\frac{\varepsilon\eta_0}{\binom{N}{2}}\right)^{1/p}.$$

If $u_{ij}(X_0) < \delta_\varepsilon$ for every $i \ne j$, then

$$V_p(X_0) = \sum_{i<j} u_{ij}(X_0)^p < \binom{N}{2}\delta_\varepsilon^p = \varepsilon\eta_0 < \eta_0.$$

The preceding stopping argument therefore applies and yields

$$\mathbb{P}_{X_0}\big(\lim_{t\to\infty} u_{ij}(X_t) = 0 \text{ for all } i \ne j\big) \ge \mathbb{P}_{X_0}(\tau = \infty) \ge 1 - \frac{V_p(X_0)}{\eta_0} \ge 1 - \varepsilon.$$

This proves part (ii), with δ_ε as the radius in its statement. □

C.2 Repulsion

When $h_0 < 0$, a small negative power gives a supermartingale that becomes large as a pairwise deficit tends to zero. We first establish the generator inequality in a neighbourhood of synchronization, then use it to exclude convergence even for trajectories that leave this neighbourhood and later return.

Lemma 17. *Suppose $h_0 < 0$ and set $h := -h_0 > 0$. Then there exist $p < 0$ and $\delta \in (0, M_*]$ such that*

$$\mathcal{L}V_p(X) \le 0 \qquad \textit{for every } X \in \mathcal{U}_\delta \textit{ with pairwise distinct tokens.}$$

Proof. Write $p = -q$ with $0 < q \le 1$. For a negative power, the upper bound in equation 26 uses the lower bound on the relative drift and the upper bound on the relative quadratic variation. Lemma 14 gives

$$\frac{\mathcal{L}u_{ij}}{u_{ij}} \ge -\Lambda^* - CM, \qquad \frac{\Gamma(u_{ij}, u_{ij})}{u_{ij}^2} \le \Sigma + CM,$$

which we substitute into the power identity to obtain

$$\mathcal{L}\big(u_{ij}^{-q}\big) = q\,u_{ij}^{-q}\Big(-\frac{\mathcal{L}u_{ij}}{u_{ij}} + \frac{1+q}{2}\frac{\Gamma(u_{ij},u_{ij})}{u_{ij}^2}\Big) \le q\,u_{ij}^{-q}\Big(\Lambda^* + \tfrac{1}{2}\Sigma + \tfrac{q}{2}\Sigma + 2CM\Big).$$

The last bracket equals $-h+\frac{q}{2}\Sigma+2CM$. Choosing $q \in (0,1]$ so that $\frac{q}{2}\Sigma \le \frac{h}{4}$ and then $\delta \in (0, M_*]$ so that $2C\delta \le \frac{h}{4}$ makes it at most $-h/2 < 0$ throughout $\mathcal{U}_\delta$. Summing this bound over pairs proves the required inequality. □

Proof of Theorem 1(i). Fix a pair $i \neq j$ with $u_{ij}(X_0) > 0$, write $u := u_{ij}$, and set $h := 2(\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}) > 0$. We first establish a negative-power estimate when this pair is close, without restricting the other tokens.

Using $\delta_{ij} = s_i + s_j - 2s_{ij}$ in equation 21–equation 22 gives

$$\mathcal{L}_{\mathrm{hom}} \log u = -\frac{d-1}{2d}(s_i+s_j) + \frac{d-3+u}{2du}\delta_{ij} \ge -\frac{\sigma_V^2(d-1)}{d} = 2\lambda_{\mathrm{ATT}},$$

since $d \ge 3$, $\delta_{ij} \ge 0$ and $s_i, s_j \le \sigma_V^2$. For the MLP contribution, Taylor expansion in equation 20 gives

$$\mathcal{L}_{\mathrm{MLP}} \log u = -\frac{b_\kappa(1-u)}{u} - \frac{q_\kappa(1-u)}{2u^2} = 2\lambda_{\mathrm{MLP}} + O(u).$$

Consequently, for sufficiently small $u > 0$, $\mathcal{L}\log u \ge h - Cu$, uniformly in the other token positions. Moreover, Lemmas 12 and 19 give a constant $K < \infty$ such that $\Gamma(u,u)/u^2 \le K$ whenever $u > 0$. Choose $q > 0$ and $\delta > 0$ sufficiently small that $C\delta + qK/2 \le h/2$. Then

$$\mathcal{L}(u^{-q}) = q\,u^{-q}\left(-\mathcal{L}\log u + \frac{q}{2}\frac{\Gamma(u,u)}{u^2}\right) \le -\frac{qh}{2}u^{-q} \qquad \text{on } \{0 < u < \delta\}.$$

Suppose first that $0 < u(X_0) < \delta$. For $0 < \varepsilon < u(X_0)$, define

$$\tau_\varepsilon := \inf\{t \ge 0 : u(X_t) \le \varepsilon\}, \qquad \tau_\delta := \inf\{t \ge 0 : u(X_t) \ge \delta\}.$$

The preceding generator inequality makes $u(X_{t\wedge\tau_\varepsilon\wedge\tau_\delta})^{-q}$ a nonnegative supermartingale. By continuity and optional stopping,

$$\mathbb{P}_{X_0}(\tau_\varepsilon < \tau_\delta) \le \varepsilon^q u(X_0)^{-q}.$$

Letting $\varepsilon \downarrow 0$ therefore gives

$$\mathbb{P}_{X_0}\big(u(X_t) \to 0 \text{ and } \tau_\delta = \infty\big) = 0. \tag{28}$$

For an arbitrary initial configuration with $u(X_0) > 0$, convergence $u(X_t) \to 0$ would imply the existence of a rational $r \ge 0$ such that $u(X_s) < \delta$ for every $s \ge r$. Lemma 20 ensures that $u(X_r) > 0$ almost surely. Conditional on X_r, the Markov property and equation 28 give probability zero to convergence while remaining in $\{0 < u < \delta\}$. Taking the countable union over rational $r \ge 0$ proves the claim. □

D PROOFS FOR THE GLOBAL CONDITION

This appendix proves the global results stated in Theorem 4 and Corollary 6.

We begin by finding a bound to the action of each part of the generator on the mean deficit $1-\rho$. The MLP contribution is controlled by the constant $\bar{\lambda}_{\mathrm{MLP}}$ as shown by the following result.

Lemma 18. *Let κ be as in Section 2. Then*

$$\bar{\lambda}_{\mathrm{MLP}} = -\inf_{r\in[-1,1)} \frac{b_\kappa(r)}{1-r}. \tag{29}$$

Moreover $\mathcal{L}_{\mathrm{MLP}}(1-\rho) \le \bar{\lambda}_{\mathrm{MLP}}(1-\rho)$ on $\mathcal{M}$.

Proof. Recall by the definition $\bar{\lambda}_{\mathrm{MLP}} := (d-1)\kappa'(1) - (d-3)\kappa$. We begin by proving that

$$\bar{\lambda}_{\mathrm{MLP}} = -\inf_{r\in[-1,1)} \frac{b_\kappa(r)}{1-r}.$$

From equation 20, $b'_\kappa(s) = -(d-1)\kappa + \kappa'(s)(d-2+s^2) + 2s\kappa(s)$, and $b'_\kappa(1) = (d-1)\kappa'(1) - (d-3)\kappa = \bar{\lambda}_{\mathrm{MLP}}$. Differentiating the series expansion of κ term wise as in Lemma 10,

$$b'_\kappa(s) = -(d-1)\kappa + \sum_{n\geq 0} k_n \big[n s^{n-1}(d-2+s^2) + 2s^{n+1}\big].$$

For $|s| \leq 1$, each bracket is at most its value $n(d-1)+2$ at $s=1$, because $0 < d-2+s^2 \leq d-1$ and $|s|^{n\pm 1} \leq 1$. Since $k_n \geq 0$, this gives $b'_\kappa(s) \leq b'_\kappa(1) = \bar{\lambda}_{\mathrm{MLP}}$ on $[-1,1]$. As $b_\kappa(1) = 0$,

$$\frac{b_\kappa(r)}{1-r} = -\frac{1}{1-r}\int_r^1 b'_\kappa(s)\, ds \geq -\bar{\lambda}_{\mathrm{MLP}} \qquad \text{for } r \in [-1,1),$$

with equality in the limit $r \uparrow 1$. The infimum in equation 29 is therefore $-\bar{\lambda}_{\mathrm{MLP}}$.
Lemma 12 gives $\mathcal{L}_{\mathrm{MLP}}\rho_{ij} = b_\kappa(\rho_{ij})$. By equation 29, $b_\kappa(r) \geq -\bar{\lambda}_{\mathrm{MLP}}(1-r)$ for every $r \in [-1,1)$, and the same inequality holds at $r=1$ because $b_\kappa(1) = 0$. Since $1-\rho = N^{-2}\sum_{i,j}(1-\rho_{ij})$ and the generator is linear,

$$\mathcal{L}_{\mathrm{MLP}}(1-\rho) = -\frac{1}{N^2}\sum_{i,j} b_\kappa(\rho_{ij}) \leq \frac{\bar{\lambda}_{\mathrm{MLP}}}{N^2}\sum_{i,j}(1-\rho_{ij}) = \bar{\lambda}_{\mathrm{MLP}}(1-\rho). \qquad \square$$

The attention estimate requires control of δ_{ij} in equation 18

Lemma 19. *For every $X \in \mathcal{M}$ and every $i \neq j$,*

$$\delta_{ij} \leq L_{\mathrm{att}}^2 \|X^i - X^j\|^2,$$

with L_{att} defined as in equation 23. If $H_{ab} \sim \mathcal{N}(0, \sigma_H^2/d)$, then $L_{\mathrm{att}}^2 \leq \sigma_V^2 d \sigma_H^2$.

Proof. Fix the token configuration and regard the weighted average $\tilde{m}$ in equation 16 as a function of the query x. Differentiating its softmax weights shows that $D_h\tilde{m}(x)$ is the covariance, under $w(x)$, between X^k and $\langle Hh, X^k\rangle$. The Cauchy–Schwarz inequality and $\|X^k\| = 1$ consequently give $\mathrm{Lip}(\tilde{m}) \leq \|H\|_{\mathrm{op}}$. Since $\delta_{ij} = \sigma_V^2 \mathbb{E}_H \|\tilde{m}_i - \tilde{m}_j\|^2$, taking the squared Lipschitz bound and averaging over H proves the first assertion. For the stated Gaussian law, the further estimate follows from $\mathbb{E}_H\|H\|_{\mathrm{op}}^2 \leq \mathbb{E}_H\|H\|_F^2 = d\sigma_H^2$. $\square$

We also use this global Lipschitz estimate to exclude collisions at finite times. This permits the logarithmic and negative-power calculations below, even on trajectories that leave the neighbourhood of synchronization.

Lemma 20. *If $X_0^i \neq X_0^j$, then almost surely $X_t^i \neq X_t^j$ for every finite $t \geq 0$.*

Proof. We show that the logarithm of a positive deficit cannot diverge to minus infinity in finite time. The exact overlap formulas and Lemma 19 give a constant C such that, on $\mathcal{M}$,

$$|\mathcal{L}u_{ij}| \leq C u_{ij}, \qquad \Gamma(u_{ij}, u_{ij}) \leq C u_{ij}^2.$$

For the MLP contribution, these bounds follow from $b_\kappa(1) = 0$ and $\kappa - r\kappa(r) = O(1-r)$. For attention, we substitute $\delta_{ij} \leq 2L_{\mathrm{att}}^2 u_{ij}$ and $|s_i|, |s_{ij}| \leq \sigma_V^2$ into equation 21–equation 22. Itô's formula, applied before the deficit reaches zero, therefore writes $\log u_{ij}(X_t)$ as its initial value plus a drift with bounded rate and a continuous local martingale with bounded quadratic-variation rate. After localization, both terms have finite limits at any finite stopping time, which excludes the divergence of $\log u_{ij}$ at a finite collision time. $\square$

Proposition 21. *Let $d \geq 3$. Then $\mathcal{L}_{\mathrm{hom}}(1-\rho) \leq (d-1)d^{-1}L_{\mathrm{att}}^2(1-\rho)$ on $\mathcal{M}$.*

Proof. We bound the negative part of the attention overlap drift. In equation 21, the term $(1-\rho_{ij})(d-2-\rho_{ij})(s_i+s_j)$ is nonnegative because $d \geq 3$ and $\rho_{ij} \leq 1$. Discarding this term and using $d-2+\rho_{ij}^2 \leq d-1$, together with Lemma 19 and $\|X^i - X^j\|^2 = 2(1-\rho_{ij})$, gives

$$-\mathcal{L}_{\mathrm{hom}}\rho_{ij} \leq \frac{d-1}{2d}\delta_{ij} \leq \frac{d-1}{d}L_{\mathrm{att}}^2(1-\rho_{ij}).$$

Since $1-\rho$ is linear in the overlaps, no quadratic-variation term enters, and averaging this bound over the pairs gives the claim. □

Proof of Theorem 4. By equation 7 and the definition of L_{att} in Lemma 23, $(d-1)d^{-1}L_{\mathrm{att}}^2 = -2\lambda_{\mathrm{ATT}}\mathbb{E}_H\|H\|_{\mathrm{op}}^2$. Combining Lemma 18 with Proposition 21 gives $\mathcal{L}(1-\rho) \leq -c(1-\rho)$ on $\mathcal{M}$. Since $1-\rho$ is smooth and bounded on $\mathcal{M}$, Dynkin's formula and Grönwall's inequality yield $\mathbb{E}[1-\rho(X_t)] \leq e^{-ct}(1-\rho(X_0))$, which proves convergence in L^1 and, by Markov's inequality, in probability.

The same drift inequality makes $1-\rho(X_t)$ a nonnegative supermartingale, so it converges almost surely to a finite limit. Fatou's lemma and the expectation bound imply that this limit has expectation zero. It therefore vanishes almost surely, which is almost sure synchronization. □

Proof of Corollary 6. The global hypothesis gives $0 \geq \bar{\lambda}_{\mathrm{MLP}}$. By equation 11, $\lambda_{\mathrm{MLP}} < 0$, and $\lambda_{\mathrm{ATT}} \leq 0$ by equation 7, so their sum is negative. Fix a pair with $u_{ij}(X_0) > 0$. The convergence just proved for Theorem 4 gives $M(X_t) \to 0$ almost surely, since $M(X) \leq N^2(1-\rho(X))$. Lemma 20 keeps $u_{ij}(X_t) > 0$ at every finite time, so Itô's formula gives

$$\log u_{ij}(X_t) = \log u_{ij}(X_0) + \int_0^t \Big[-\Lambda^* - \tfrac{1}{2}\Sigma + \varepsilon_s\Big]ds + n_t,$$

where n is a continuous local martingale starting at zero. The global bounds in the proof of Lemma 20 give $\langle n\rangle_t \leq Ct$, for a constant C depending only on the fixed model parameters. Lemma 14 gives $|\varepsilon_s| \leq CM(X_s)$ whenever $M(X_s) \leq M_*$, so $\varepsilon_s \to 0$ almost surely. The strong law for continuous local martingales gives $n_t/t \to 0$, and Cesàro convergence gives $t^{-1}\int_0^t \varepsilon_s\,ds \to 0$ almost surely. Dividing by t and using Proposition 15, we obtain

$$\lim_{t\to\infty}\frac{1}{t}\log u_{ij}(X_t) = -\Lambda^* - \tfrac{1}{2}\Sigma = 2(\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}).$$

For the stated range of γ, the logarithmic rate of $e^{\gamma t}u_{ij}(X_t)$ is strictly negative, proving its convergence to zero. Initially coincident pairs remain coincident because their tokens solve the same equation with the same driving noises. □

Before concluding the appendix, it is worth stating the following.

Proposition 22. *Under the same assumption of Theorem 4, every stationary distribution of the dynamics is the law of $(Z,\dots,Z)$, where Z has a stationary distribution for the one-token dynamics (equation 5 with $N=1$). If this one-token stationary distribution is unique, then so is the stationary distribution of the full system.*

Proof. To identify the support of an invariant probability measure π, we integrate the drift inequality and use invariance to obtain

$$0 = \int_{\mathcal{M}} \mathcal{L}(1-\rho)\,d\pi \leq -c\int_{\mathcal{M}}(1-\rho)\,d\pi.$$

As $1-\rho$ is nonnegative and vanishes exactly on Δ_N, its zero integral implies $\pi(\Delta_N) = 1$. On this invariant set, the process is the diagonal image of the one-token diffusion. If the latter has a unique invariant law, its diagonal image is consequently the unique invariant law of the full system. □

Lemma 23. *Let $N = 2$ and let the activation be odd with $\sigma_b = \sigma_c = 0$, so that κ is odd by Lemma 10. Then $X_0^2 = -X_0^1$ implies $X_t^2 = -X_t^1$ for all $t \geq 0$.*

Proof. First, for every $x \in \mathbb{S}^{d-1}$,

$$\mathbb{E}\|B_t(x) + B_t(-x)\|^2 = 2td\big(\kappa(1) + \kappa(-1)\big) = 0,$$

so $B_t(-x) = -B_t(x)$ almost surely. By continuity of the field, this holds simultaneously for all x and t on an event of probability one. Second, let $X = (y, -y)$. For the query $-y$ the logits are $\langle H(-y), y\rangle = -\langle Hy, y\rangle$ and $\langle H(-y), -y\rangle = \langle Hy, y\rangle$, so the two softmax weights are exchanged, and $\mathrm{Attn}_\theta[X](-y) = -\mathrm{Attn}_\theta[X](y)$ for every θ. Since also $P_{-y} = P_y$, the radial drifts of the two tokens coincide. Now let y_t solve the equation for the first token of equation 5 in which the second token is replaced by $-y_t$:

$$\begin{aligned} dy_t =& P_{y_t} \int_\Theta \mathrm{Attn}_\theta[(y_t, -y_t)](y_t)\, W(d\theta, dt) + P_{y_t}\, dB_t(y_t) \\ & -\tfrac{1}{2} y_t \Big[\int_\Theta \big\| P_{y_t}\, \mathrm{Attn}_\theta[(y_t, -y_t)](y_t) \big\|^2 \rho_*(d\theta) + (d-1)\kappa \Big] dt, \end{aligned}$$

with $y_0 = X_0^1$. Its coefficients are those of equation 5 composed with the smooth map $y \mapsto (y, -y)$, so it has a unique strong solution. By the two symmetries above, $(y_t, -y_t)$ solves equation 5, and pathwise uniqueness for equation 5 gives $X_t = (y_t, -y_t)$ for all $t \geq 0$. Hence $\rho_{12}(X_t) = -1$ for all times, consistently with the vanishing at $\rho_{12} = -1$ of the MLP drift $b_\kappa(-1)$, of $q_\kappa(-1)$, of equation 22 through its factor $1 - \rho_{12}^2$, and of the bracket in equation 21, which becomes $2(d-1)(s_1 + s_2) - (d-1)\delta_{12} = 0$ since $\delta_{12} = 2(s_1 + s_2)$. Homogenized attention thus preserves the antipodal symmetry rather than breaking it, and by Lemma 10 any bias, $\sigma_b > 0$ or $\sigma_c > 0$, makes $\kappa(1) + \kappa(-1)$ strictly positive. □

Lemma 24. *If $-\bar{\lambda}_{\mathrm{MLP}} > 0$, and in particular under the assumption of Theorem 4, then $\kappa(1) + \kappa(-1) > 0$. Consequently, the assumption of Theorem 4 fails for every odd activation with $\sigma_b = \sigma_c = 0$.*

Proof. Since $\lambda_{\mathrm{ATT}} \leq 0$, the assumption $c > 0$ of Theorem 4 implies $-\bar{\lambda}_{\mathrm{MLP}} > 0$. By Lemma 18, $-\bar{\lambda}_{\mathrm{MLP}}$ is the infimum of $b_\kappa(r)/(1-r)$ over $r \in [-1, 1)$, so evaluating at $r = -1$ and using equation 20,

$$0 < -\bar{\lambda}_{\mathrm{MLP}} \leq \frac{b_\kappa(-1)}{2} = \frac{(d-1)\big(\kappa(1) + \kappa(-1)\big)}{2}.$$

For an odd activation with $\sigma_b = \sigma_c = 0$, κ is odd by Lemma 10, so $\kappa(1) + \kappa(-1) = 0$. □

For $S \subseteq \{1, \dots, N\}$, let $F_S : \mathcal{M} \to \mathcal{M}$ change the sign of the tokens in S, that is, $F_S(X)^i := -X^i$ for $i \in S$ and $F_S(X)^i := X^i$ otherwise. Then $\rho_{ij}(F_S(X)) = -\rho_{ij}(X)$ when exactly one of i, j lies in S, and $\rho_{ij}(F_S(X)) = \rho_{ij}(X)$ otherwise.

Lemma 25. *Let $\sigma_V = 0$, and let the activation be odd with $\sigma_b = \sigma_c = 0$.*

(i) For every $S \subseteq \{1, \dots, N\}$, if X_t solves equation 5, then $F_S(X_t)$ solves equation 5, driven by the same noise, with initial condition $F_S(X_0)$.

(ii) Suppose $\lambda_{\mathrm{MLP}} < 0$. For every $\varepsilon \in (0, 1)$ there exists $\delta > 0$ such that, for every nonempty proper subset S and every X_0 with $\max_{i,j} u_{ij}(F_S(X_0)) < \delta$, with probability at least $1 - \varepsilon$,

$$u_{ij}(X_t) \to 0 \ \text{ if } i, j \in S \text{ or } i, j \notin S, \qquad u_{ij}(X_t) \to 2 \ \text{ if exactly one of } i, j \text{ lies in } S.$$

In particular $\mathbb{P}_{X_0}\big(\max_{i,j} u_{ij}(X_t) \to 0\big) \leq \varepsilon$, and the set of such X_0 is open in $\mathcal{M}$.

Proof. (i) As in the proof of Lemma 23, κ is odd and $B_t(-x) = -B_t(x)$ for all x and t on an event of probability one. For $\sigma_V = 0$, equation 5 decouples into $dX_t^i = P_{X_t^i}\, dB_t(X_t^i) - \frac{1}{2}(d-1)\kappa X_t^i\, dt$, one equation for each token, all driven by the same field. If X_t^i solves it, then

$$d(-X_t^i) = -P_{X_t^i}\, dB_t(X_t^i) + \tfrac{1}{2}(d-1)\kappa X_t^i\, dt = P_{-X_t^i}\, dB_t(-X_t^i) - \tfrac{1}{2}(d-1)\kappa(-X_t^i)\, dt,$$

because $P_{-x} = P_x$ and $B_t(-x) = -B_t(x)$. Hence $-X_t^i$ solves the same equation, and so does $F_S(X_t)$.

(ii) Let δ be the radius given by Theorem 1(ii) for ε, which applies since $\lambda_{\mathrm{MLP}}+\lambda_{\mathrm{ATT}} = \lambda_{\mathrm{MLP}} < 0$. By (i) and pathwise uniqueness, $Y_t := F_S(X_t)$ is the solution of equation 5 started at $F_S(X_0)$, so $\max_{i,j} u_{ij}(Y_t) \to 0$ with probability at least $1-\varepsilon$. On this event, $u_{ij}(X_t) = u_{ij}(Y_t) \to 0$ when i and j lie on the same side of S, and $u_{ij}(X_t) = 1 + \rho_{ij}(Y_t) = 2 - u_{ij}(Y_t) \to 2$ otherwise. Since S is a nonempty proper subset, at least one pair lies across S, so $\max_{i,j} u_{ij}(X_t) \to 2$ on this event. The set of admissible X_0 is open because $X \mapsto \max_{i,j} u_{ij}(F_S(X))$ is continuous. □

E Proofs for the Jacobian

We now apply the pairwise calculation to the linearized equation along Δ_N. The exponent of a single nonzero pairwise variation will first give the exponent of a transverse vector, since its squared norm is a finite sum of squared pairwise differences. We then use the same relation to compute the second moment. Throughout this appendix, $\mathcal{L}$ and Γ also denote the generator and quadratic covariation of the lifted process on the tangent bundle, so their action on a variation is computed from the linearized equation.

Lemma 26. *Let $a > 0$ and $b \geq 0$. Let $x_t \in \mathbb{S}^{d-1}$ and $z_t \in \mathbb{R}^d$ be continuous semimartingales with $\langle z_t, x_t\rangle = 0$ for all t, and suppose that*

$$dz_t = -\big(z_t x_t^\top + x_t z_t^\top\big)\, dG_t + P_{x_t}\, dM_t - \tfrac{1}{2}(d-1)a\, z_t\, dt + x_t\, \alpha_t\, dt, \tag{30}$$

where α is a real progressively measurable process with locally integrable paths, and G, M are continuous local martingales in $\mathbb{R}^d$ with

$$d\langle G^k, G^l\rangle_t = a\,\delta_{kl}\, dt, \qquad d\langle M^k, M^l\rangle_t = b\,\|z_t\|^2\,\delta_{kl}\, dt, \qquad d\langle G^k, M^l\rangle_t = 0.$$

Then $Y_t := \|z_t\|^2$ satisfies

$$dY_t = \big((d-1)b - (d-3)a\big)Y_t\, dt + d\mathcal{N}_t, \qquad d\langle \mathcal{N}\rangle_t = 4(a+b)\, Y_t^2\, dt,$$

where $\mathcal{N}$ is a continuous local martingale starting at zero.

Proof. Itô's formula gives $dY_t = 2\langle z_t, dz_t\rangle + \sum_k d\langle z^k\rangle_t$. Since $z_t \perp x_t$ and $\langle z_t, P_{x_t} v\rangle = \langle z_t, v\rangle$, the first term equals

$$2\langle z_t, dz_t\rangle = -(d-1)a\, Y_t\, dt + d\mathcal{N}_t, \qquad d\mathcal{N}_t := -2Y_t\langle x_t, dG_t\rangle + 2\langle z_t, dM_t\rangle.$$

The martingale part of dz_t is $-z_t\langle x_t, dG_t\rangle - x_t\langle z_t, dG_t\rangle + P_{x_t} dM_t$. Its three terms have pairwise zero covariation: the first two because $\langle z_t, x_t\rangle = 0$, the last with the first two because G and M have zero covariation. Using $\|x_t\| = 1$ and $\operatorname{Tr} P_{x_t} = d-1$,

$$\sum_k d\langle z^k\rangle_t = \big(a\, Y_t + a\, Y_t + (d-1)b\, Y_t\big)dt.$$

Adding the two contributions gives the drift $\big((d-1)b - (d-3)a\big)Y_t$. Finally, $d\langle\langle x, G\rangle\rangle_t = a\, dt$, $d\langle\langle z, M\rangle\rangle_t = b\, Y_t^2\, dt$ and the cross covariation vanishes, so $d\langle\mathcal{N}\rangle_t = 4\big(aY_t^2 + bY_t^2\big)dt$. □

We state the scalar evolution used in Section 3. For $X_0 \in \Delta_N$ and a tangent variation w_0, set $w_t := D\Phi_t(X_0)w_0$, $z_{ij,t} := w_t^i - w_t^j$ and $Y_{ij,t} := \|z_{ij,t}\|^2$. The coefficients Λ^* and Σ are given in equation 25.

Proposition 27. *Under the assumptions of Section 2, fix $X_0 \in \Delta_N$ and a nonzero $w_0 \in T_{X_0}\mathcal{M}$, both independent of the noise. Let w_t, $z_{ij,t}$ and $Y_{ij,t}$ be as above. For each pair with $Y_{ij,0} > 0$, there is a real-valued standard Brownian motion B^{ij}, adapted to the driving filtration, such that*

$$dY_{ij,t} = -\Lambda^* Y_{ij,t}\, dt + \sqrt{\Sigma}\, Y_{ij,t}\, dB_t^{ij}, \qquad Y_{ij,t=0} = \|w_0^i - w_0^j\|^2. \tag{31}$$

Equivalently, by Itô's formula the norm itself satisfies

$$d\|z_{ij,t}\| = \Big[\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}} + \tfrac{1}{2}\Big(\kappa + \kappa'(1) + \frac{\sigma_V^2}{d}\Big)\Big]\|z_{ij,t}\|\, dt + \sqrt{\kappa + \kappa'(1) + \frac{\sigma_V^2}{d}}\, \|z_{ij,t}\|\, dB_t^{ij},$$

which is the form used in Section 3. These Brownian motions need not be independent across pairs. If $Y_{ij,0} = 0$, then $Y_{ij,t} = 0$ for all $t \geq 0$ almost surely. For every pair with $Y_{ij,0} > 0$, almost surely,

$$\lim_{t\to\infty} \frac{1}{t} \log \frac{\|z_{ij,t}\|}{\|z_{ij,0}\|} = -\tfrac{1}{2}\big(\Lambda^* + \tfrac{1}{2}\Sigma\big). \tag{32}$$

If $w_0 \in (T_{X_0}\Delta_N)^\perp$, the same limit holds for $t^{-1}\log(\|w_t\|/\|w_0\|)$ almost surely.

Proof of Proposition 27. Permutation equivariance along Δ_N makes the variational equation preserve both the transverse space and each subspace $w^i = w^j$. In particular, a pair with $Y_{ij,0} = 0$ remains zero.

Write $X_t = \Phi_t(X_0) = (x_t, \dots, x_t)$. Differentiating equation 5 along X_t gives an Itô equation for each w_t^i, with no correction term. We compute the three contributions to $z_t := z_{ij,t}$.

At a synchronized configuration all tokens equal x_t and every softmax weight equals $1/N$. The weights sum to one, so their variations sum to zero: varying the logits does not change the weighted average $\tilde{m}_i$ to first order, while varying the averaged tokens changes it by $\bar{w} := N^{-1}\sum_k w^k$, whatever the index i. Hence $D\big(\mathrm{Attn}_\theta[X](X^i)\big)[w] = V\bar{w}$ for every i, and the attention noise contributes to z_t only through the variation of the projection, $DP_x[z] = -(zx^\top + xz^\top)$, applied to $\mathrm{Attn}_\theta[X_t](x_t) = Vx_t$. The MLP noise contributes $-(z_t x_t^\top + x_t z_t^\top)\,B(dt, x_t) + P_{x_t}\,\partial_{z_t}B(dt, x_t)$, where $\partial_z B$ denotes the derivative of the field in the direction z. The radial drift in equation 5 has coefficient $(d-1)K(1)$ at Δ_N, with $K(1) := \kappa + \sigma_V^2/d$ as in Lemma 28, and its variation in the configuration only produces a term along x_t. Altogether,

$$dz_t = -\big(z_t x_t^\top + x_t z_t^\top\big)dG_t + P_{x_t}\,dM_t - \tfrac{1}{2}(d-1)K(1)\,z_t\,dt + x_t\,\alpha_t\,dt,$$

with α a real process and

$$G_t := \int_0^t\!\!\int_\Theta Vx_s\,W(d\theta, ds) + \int_0^t B(ds, x_s), \qquad M_t := \int_0^t \partial_{z_s}B(ds, x_s).$$

By Lemma 11 with $s_{ij} = \sigma_V^2$ at Δ_N and by equation 4, $d\langle G^k, G^l\rangle_t = K(1)\delta_{kl}\,dt$. Differentiating $\langle h(x), h(y)\rangle = \kappa(\langle x, y\rangle)$ as in Appendix A gives, for $u, v \in T_x\mathbb{S}^{d-1}$, $\langle D_u h(x), D_v h(x)\rangle = \kappa'(1)\langle u, v\rangle$ and $\langle h(x), D_v h(x)\rangle = \kappa'(1)\langle x, v\rangle = 0$. Hence $d\langle M^k, M^l\rangle_t = \kappa'(1)\|z_t\|^2\delta_{kl}\,dt$, and M has zero covariation with the MLP part of G and, by independence, with its attention part. Lemma 26 with $a = K(1)$ and $b = \kappa'(1)$ gives

$$dY_{ij,t} = \big((d-1)\kappa'(1) - (d-3)K(1)\big)Y_{ij,t}\,dt + d\mathcal{N}_{ij,t}, \qquad d\langle\mathcal{N}_{ij}\rangle_t = 4\big(K(1) + \kappa'(1)\big)Y_{ij,t}^2\,dt.$$

By equation 25, $(d-1)\kappa'(1) - (d-3)K(1) = -\Lambda^*$ and $4(K(1) + \kappa'(1)) = \Sigma$.

Suppose $Y_{ij,0} > 0$. Up to the first time $Y_{ij,t}$ reaches zero, Itô's formula after localization gives

$$\log\frac{Y_{ij,t}}{Y_{ij,0}} = \big(-\Lambda^* - \tfrac{1}{2}\Sigma\big)t + \int_0^t Y_{ij,s}^{-1}\,d\mathcal{N}_{ij,s}.$$

The local martingale on the right has quadratic variation Σt. It has a finite limit at every finite stopping time up to which this quadratic variation remains bounded, so the logarithm cannot tend to $-\infty$ in finite time. Thus $Y_{ij,t} > 0$ at all finite times. The process

$$B_t^{ij} := \frac{1}{\sqrt{\Sigma}}\int_0^t Y_{ij,s}^{-1}\,d\mathcal{N}_{ij,s}$$

is consequently defined for all $t \geq 0$ and has quadratic variation t. Lévy's characterization identifies it as a standard Brownian motion, which proves equation 31. The logarithmic identity and $B_t^{ij}/t \to 0$ prove the pairwise growth limit. Finally, if $w_0 \in (T_{X_0}\Delta_N)^\perp$, transversality gives $\|w_t\|^2 = N^{-1}\sum_{i<j} Y_{ij,t}$. For the fixed nonzero w_0, this finite sum has at least one initially positive term, and every such term has the same almost sure logarithmic rate. Taking its logarithm and dividing by $2t$ proves the limit for $\|w_t\|$. □

Since $Y_{ij,t} = Y_{ij,0}\exp\big((-\Lambda^* - \tfrac{1}{2}\Sigma)t + \sqrt{\Sigma}B_t^{ij}\big)$, the second moment and the mean logarithm are explicit:

$$\mathbb{E}[Y_{ij,t}] = e^{-\Lambda^* t}\,Y_{ij,0}, \qquad \log\mathbb{E}[Y_{ij,t}] - \mathbb{E}[\log Y_{ij,t}] = \tfrac{1}{2}\Sigma t, \tag{33}$$

the first because the exponential martingale $e^{\sqrt{\Sigma}B_t - \Sigma t/2}$ has mean one. For a transverse w_0, summing over pairs with $\|w_t\|^2 = N^{-1}\sum_{i<j} Y_{ij,t}$ gives $\mathbb{E}\|w_t\|^2 = e^{-\Lambda^* t}\|w_0\|^2$.

E.1 THE LONGITUDINAL BLOCK

At a synchronized configuration all softmax averages coincide with the common token, so attention acts as a linear map of that token. The longitudinal block is therefore the derivative of a one-token flow driven by a single isotropic field.

Lemma 28. *Let $K(r) := \kappa(r) + \sigma_V^2 r/d$, and let ϕ_t be the flow on $\mathbb{S}^{d-1}$ of*

$$dx_t = P_{x_t}\, dB_t^K(x_t) - \tfrac{1}{2}(d-1)K(1)\, x_t\, dt, \tag{34}$$

where B^K is the centered Gaussian Brownian field

$$B_t^K(x) := B_t(x) + \int_\Theta Vx\, W(d\theta, [0,t]), \qquad \mathbb{E}\big[B_t^K(x) B_s^K(y)^\top\big] = (t \wedge s) K(\langle x, y\rangle) I_d.$$

Then $\Phi_t(y,\dots,y) = (\phi_t(y),\dots,\phi_t(y))$ for every $y \in \mathbb{S}^{d-1}$. Consequently, for $X_0 = (x,\dots,x)$ and $v \in T_x\mathbb{S}^{d-1}$,

$$J_t(v,\dots,v) = (D\phi_t(x)v, \dots, D\phi_t(x)v).$$

Proof. If $X^1 = \cdots = X^N = y$, every softmax average equals y, so $\mathrm{Attn}_\theta[X](X^i) = Vy$ for all θ and i. The attention noise at such configurations is the linear field $x \mapsto \int_\Theta Vx\, W(d\theta, dt)$, whose covariance is $\int_\Theta (Vx)(Vy)^\top \rho_*(d\theta) = \frac{\sigma_V^2}{d}\langle x, y\rangle I_d$. Being independent of B, it adds to the MLP field to give covariance K. Since $\int_\Theta \|P_y V y\|^2 \rho_*(d\theta) = (d-1)\sigma_V^2/d$, the radial drift of equation 5 is then

$$-\tfrac{1}{2}\Big((d-1)\frac{\sigma_V^2}{d} + (d-1)\kappa\Big) y = -\tfrac{1}{2}(d-1)K(1)\, y.$$

Each token therefore solves equation 34 with the same driving field, and pathwise uniqueness gives $X_t^i = \phi_t(y)$ for all i. Differentiating the identity $\Phi_t(y,\dots,y) = (\phi_t(y),\dots,\phi_t(y))$ in y along v proves the last claim. □

Proposition 29. *Let $X_0 = (x,\dots,x)$, let $v \in T_x\mathbb{S}^{d-1}$ be fixed and nonzero, and set $v_t := D\phi_t(x)v$ and $Y_t := \|v_t\|^2$. With $\bar\lambda_K := (d-1)K'(1) - (d-3)K(1)$ and $\Sigma_K := 4\big(K(1) + K'(1)\big)$, there is a standard real Brownian motion B such that*

$$dY_t = \bar\lambda_K\, Y_t\, dt + \sqrt{\Sigma_K}\, Y_t\, dB_t.$$

Consequently, almost surely,

$$\lim_{t\to\infty} \frac{1}{t} \log \|v_t\| = \tfrac{1}{2}\bar\lambda_K - \tfrac{1}{4}\Sigma_K = \lambda_{\mathrm{MLP}} + \frac{2}{d-1}\lambda_{\mathrm{ATT}},$$

and

$$\mathbb{E}\|v_t\|^2 = e^{\bar\lambda_K t}\|v\|^2, \qquad \bar\lambda_K = \bar\lambda_{\mathrm{MLP}} - \frac{4}{d-1}\lambda_{\mathrm{ATT}}.$$

Proof. By Lemma 28, $x_t = \phi_t(x)$ solves equation 34. Differentiating in the initial point gives

$$dv_t = -\big(v_t x_t^\top + x_t v_t^\top\big) B^K(dt, x_t) + P_{x_t}\, \partial_{v_t} B^K(dt, x_t) - \tfrac{1}{2}(d-1)K(1)\, v_t\, dt,$$

and $\langle v_t, x_t\rangle = 0$. The field B^K has covariance $K(\langle x, y\rangle)I_d$, so the computation in the proof of Proposition 27, with κ replaced by K, shows that Lemma 26 applies with $a = K(1)$ and $b = K'(1)$. This yields the stated equation, the logarithmic limit $\frac{1}{2}(\bar\lambda_K - \frac{1}{2}\Sigma_K)$ and, as in equation 33, the second moment. Finally $K(1) = \kappa + \sigma_V^2/d$ and $K'(1) = \kappa'(1) + \sigma_V^2/d$ give $\bar\lambda_K = \bar\lambda_{\mathrm{MLP}} + 2\sigma_V^2/d$, and equation 11 gives $\frac{1}{2}\bar\lambda_K - \frac{1}{4}\Sigma_K = \lambda_{\mathrm{MLP}} - \sigma_V^2/d$. Since $\sigma_V^2/d = -2\lambda_{\mathrm{ATT}}/(d-1)$ by equation 7, both identities follow. □

E.2 Proof of Theorem 9

Proof of Theorem 9. Quenched rates. For a fixed nonzero transverse w_0, Proposition 27 and Proposition 15 give $t^{-1}\log\|J_t^{\perp}w_0\| \to \lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}$ almost surely. Choose an orthonormal basis (f_b) of the transverse space at X_0, of dimension $(N-1)(d-1)$. The limits hold simultaneously for these finitely many vectors, and

$$\max_b \|J_t^{\perp} f_b\| \le \|J_t^{\perp}\| \le \Big(\sum_b \|J_t^{\perp} f_b\|^2\Big)^{1/2}$$

gives the same limit for the operator norm. For the longitudinal block, let (e_a) be an orthonormal basis of $T_x\mathbb{S}^{d-1}$. The vectors $\hat{e}_a := N^{-1/2}(e_a,\dots,e_a)$ form an orthonormal basis of the longitudinal space, and Lemma 28 gives

$$\|J_t^{\parallel}\hat{e}_a\| = \|D\phi_t(x)e_a\|.$$

Proposition 29 and the same sandwich then give

$$\lambda(J^{\parallel}) = \lambda_{\mathrm{MLP}} + \frac{2}{d-1}\lambda_{\mathrm{ATT}}.$$

Since J_t maps the orthogonal decomposition at X_0 onto the orthogonal decomposition at X_t, it is block diagonal, so

$$\|J_t\| = \max\big(\|J_t^{\parallel}\|, \|J_t^{\perp}\|\big), \qquad \lambda(J) = \max\big(\lambda(J^{\parallel}), \lambda(J^{\perp})\big) = \lambda(J^{\parallel})$$

by equation 13. All limits are explicit constants, hence deterministic.

Annealed rates. By equation 33, $\mathbb{E}\|J_t^{\perp}\|_F^2 = \sum_b \mathbb{E}\|J_t^{\perp} f_b\|^2 = (N-1)(d-1)e^{-\Lambda^* t}$, so $\bar{\lambda}(J^{\perp}) = -\Lambda^* = \bar{\lambda}_{\mathrm{MLP}} - \sigma_V^2(d-3)/d = \bar{\lambda}_{\mathrm{MLP}} + 2(d-3)\lambda_{\mathrm{ATT}}/(d-1)$ by equation 25 and equation 7. By Proposition 29, $\mathbb{E}\|J_t^{\parallel}\|_F^2 = \sum_a \mathbb{E}\|D\phi_t(x)e_a\|^2 = (d-1)e^{\bar{\lambda}_K t}$, so $\bar{\lambda}(J^{\parallel}) = \bar{\lambda}_{\mathrm{MLP}} - 4\lambda_{\mathrm{ATT}}/(d-1)$. Block diagonality gives $\|J_t\|_F^2 = \|J_t^{\parallel}\|_F^2 + \|J_t^{\perp}\|_F^2$, hence $\bar{\lambda}(J) = \max(\bar{\lambda}(J^{\parallel}), \bar{\lambda}(J^{\perp})) = \bar{\lambda}(J^{\parallel})$, since the difference of the two rates is $-2\lambda_{\mathrm{ATT}} \ge 0$.

Transposition. The operator and Frobenius norms are invariant under transposition, so all four rates are unchanged. □

Remark 30. The cases excluded from the band can be read from equation 14. At $\bar{\lambda}(J^{\perp}) = 0$ the second moment of a transverse perturbation is constant while $\lambda(J^{\perp}) = -\Sigma/4 < 0$, whereas at $\lambda(J^{\perp}) = 0$ one has $\bar{\lambda}(J^{\perp}) = \Sigma/2 > 0$. In the latter case the linear calculation does not decide whether the nonlinear tokens synchronize, which is the critical case left open by Theorem 1.

E.3 Rate of collapse

The quenched transverse exponent also describes the nonlinear token distances on the non-exit event of the local theorem. For a finite initial separation the expansion equation 24 contains relative errors of order $M(X_t)$, and on that event the configuration synchronizes, so these errors tend to zero and their contribution to the time-averaged logarithmic drift vanishes.

Proposition 31. *Assume $\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}} < 0$ and let τ be the stopping time constructed in the proof of Theorem 1(ii) in Appendix C.1. On $\{\tau = \infty\}$, for every pair $i \ne j$ with $X_0^i \ne X_0^j$,*

$$\lim_{t\to\infty} \frac{1}{t}\log\|X_t^i - X_t^j\| = \lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}} \qquad \textit{almost surely}.$$

Proof. We compute the rate on $\{\tau = \infty\}$, where the proof of Theorem 1(ii) gives $M(X_t) \to 0$ and the construction of τ ensures $M(X_t) < \delta$ for all times. Fix a pair $i \ne j$ with $u_{ij}(X_0) > 0$. Since Lemma 20 keeps this deficit positive at finite times, we can apply Itô's formula to its logarithm up to τ. On the non-exit event, Lemma 14 gives

$$d\log u_{ij}(X_t) = \Big(\frac{\mathcal{L}u_{ij}}{u_{ij}} - \frac{1}{2}\frac{\Gamma(u_{ij},u_{ij})}{u_{ij}^2}\Big)dt + dn_t = \Big(2(\lambda_{\mathrm{MLP}} + \lambda_{\mathrm{ATT}}) + \varepsilon_t\Big)dt + dn_t,$$

where $|\varepsilon_t| \le \frac{3}{2}CM(X_t)$ and $d\langle n\rangle_t = (\Sigma + O(M(X_t)))dt$. The convergence of $M(X_t)$ to zero implies $\frac{1}{t}\int_0^t \varepsilon_s\,ds \to 0$ by Cesàro convergence. Moreover, $\langle n_{\cdot\wedge\tau}\rangle_t \le (\Sigma + C\delta)t$, so the law of

large numbers for continuous martingales gives $n_{t\wedge\tau}/t \to 0$ almost surely. Dividing the logarithmic identity by time therefore yields $\frac{1}{t}\log u_{ij}(X_t) \to 2(\lambda_{\mathrm{MLP}}+\lambda_{\mathrm{ATT}})$ on $\{\tau=\infty\}$, and $\|X_t^i - X_t^j\| = \sqrt{2u_{ij}(X_t)}$ gives the stated separation rate. □

E.4 Vanishing gradients of an early shared block

We make precise the training statement of Section 4. Let $F : \mathbb{R}^q \times \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$, $(\vartheta, y) \mapsto F_\vartheta(y)$, be C^1, and let $X_s = (y, \dots, y) \in \Delta_N$ be the synchronized configuration at depth $s > 0$. The block replaces it by $X_s(\vartheta) := (F_\vartheta(y), \dots, F_\vartheta(y))$, after which the dynamics equation 5 continues from depth s. Write $J_{t,s}$ for the derivative at $X_s(\vartheta)$ of the flow from depth s to depth t, and $\Pi^{\parallel}$ for the orthogonal projection onto the longitudinal space at X_t.

Proposition 32. *Fix ϑ and $\mathcal{R} \in C^1(\mathcal{M})$, and set $B := D_\vartheta X_s(\vartheta)$. Then*

$$\nabla_\vartheta \mathcal{R}(X_t) = B^\top \big(J_{t,s}^{\parallel}\big)^\top \Pi^{\parallel} \nabla\mathcal{R}(X_t), \qquad \|\nabla_\vartheta \mathcal{R}(X_t)\| \le \|B\|\, \|J_{t,s}^{\parallel}\| \sup_{\mathcal{M}} \|\nabla\mathcal{R}\|,$$

and almost surely

$$\limsup_{t\to\infty} \frac{1}{t} \log \|\nabla_\vartheta \mathcal{R}(X_t)\| \le \lambda_{\mathrm{MLP}} + \frac{2}{d-1}\lambda_{\mathrm{ATT}}.$$

In particular, if the right-hand side is negative, a gradient step $\vartheta \mapsto \vartheta - \eta\nabla_\vartheta \mathcal{R}(X_t)$ with η independent of t changes ϑ by an exponentially small amount as $t \to \infty$.

Proof. Since $X_s(\vartheta) \in \Delta_N$ for every ϑ, the columns of B are tangent to Δ_N, that is, longitudinal. The Jacobian preserves the longitudinal space along Δ_N, so $J_{t,s}B = J_{t,s}^{\parallel}B$, and the chain rule gives $D_\vartheta \mathcal{R}(X_t) = D\mathcal{R}(X_t)\, J_{t,s}^{\parallel} B$. Transposing, and using that $J_{t,s}^{\parallel}$ takes values in the longitudinal space at X_t, yields the formula with $\Pi^{\parallel}$. The bound follows from $\|B^\top\| = \|B\|$, $\|(J_{t,s}^{\parallel})^\top\| = \|J_{t,s}^{\parallel}\|$ and $\|\Pi^{\parallel}\| \le 1$, and $\nabla\mathcal{R}$ is bounded because $\mathcal{M}$ is compact. For the rate, the driving noises after depth s are independent of those up to s, so conditionally on the past the flow from s to t is distributed as Φ_{t-s} started at the synchronized point $X_s(\vartheta)$. Theorem 9 gives $(t-s)^{-1}\log\|J_{t,s}^{\parallel}\| \to \lambda(J^{\parallel})$ almost surely for every fixed synchronized starting point, hence, integrating over the starting point, almost surely. As $\|B\|$ does not depend on t, the claim follows. □