A multiscale analysis of mean-field transformers in the moderate interaction regime

Giuseppe Bruno
Department of Mathematics and Statistics
University of Bern
giuseppe.bruno@unibe.ch

Federico Pasqualotto
Department of Mathematics
University of California, San Diego
fpasqualotto@ucsd.edu

Andrea Agazzi
Department of Mathematics and Statistics
University of Bern
andrea.agazzi@unibe.ch

Abstract

We study the evolution of tokens through the depth of encoder-only transformer models at inference time by modeling them as a mean-field interacting particle system, and analyzing the corresponding dynamics. More specifically, we consider this problem in the *moderate interaction regime*, where the number N of tokens is large and the inverse temperature parameter β of the model scales together with N. In this regime, the dynamics of the system displays a multiscale behavior: a fast phase, where the token empirical measure collapses on a low-dimensional subspace, an intermediate phase, where the measure further collapses into clusters, and a slow phase, where such clusters sequentially merge into a single one. We characterize the limiting dynamics in each phase, exemplifying our results with some simulations.

1 Introduction

The transformer architecture [49], through its extensive use in Large Language Models, has played a crucial role in the recent, unprecedented developments in machine learning and artificial intelligence. One of the key innovations at the heart of this architecture are self-attention modules [6], allowing to capture long-range dependencies in the data, e.g., in prompts with a large number N of tokens. To further improve their performance, practitioners have implemented these models in different hyperparameter regimes, e.g., choosing the model's inverse temperature parameter β (a parameter that scales the query-key dot products in the self-attention layer) as a function of N [37, 43]. However, the groundbreaking empirical success of these machine learning models remains largely unexplained from the theoretical perspective. In particular, a precise mathematical description of the internal representations learned by transformers, and of how these representations behave in different hyperparameter regimes, is still lacking.

A promising approach to fill this gap was presented in the work [46], where the authors interpret tokens traveling through a deep stack of transformer layers as particles evolving in time and interacting in a mean-field way. A subsequent line of work [24] has then observed that tokens in this model tend to organize into clusters, offering - in a simplified setting - a compelling qualitative explanation of how transformer models build representation of complex input data.

Despite its apparent simplicity, this interacting particle system exhibits a remarkably rich dynamical behavior. Indeed, recent studies have identified distinct dynamical phases, characterized by quali-

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

tatively different clustering patterns, which depend on specific choices of parameters, timescales, and initial conditions. However, these results often rely on restrictive - and sometimes unrealistic - assumptions, and they typically capture only limited aspects of the collapse dynamics, providing partial views of the transformer's complex dynamical landscape that are difficult to reconcile into a consistent and global dynamical picture. Such a global characterization of the clustering phenomenon in a realistic parameter regime is arguably fundamental to understanding how internal representations form in deep models and how to operate such models in the optimal hyperparameter regimes.

Contributions In this work, we study the dynamics of the mean-field transformer model developed in [46, 25], constrained on the d-dimensional sphere, in the limit of large context size, i.e., when the number N of input tokens is large. Motivated by recent scaling strategies in state-of-the-art LLMs such as SSMax [37] and YaRN [43] (used respectively in Llama 4 and Qwen 3), we consider the setting where the inverse temperature parameter β grows with N. In this regime, our contributions can be summarized as follows:

1. We identify three distinct dynamical phases (respectively denoted the *alignment*, *heat* and *pairing* phases), corresponding to different scales of time as a function of the parameter β. In each phase the model dynamics displays, asymptotically in β, qualitatively distinct behavior, characterized by a different limiting equation.
2. In the alignment phase, occurring on a fast timescale of order $O(1)$, we prove our main technical result: under general assumptions on the parameter matrices, the finite particle dynamics converges to a linear transport equation modeling the collapse of the token measure onto a low-dimensional manifold dictated by the spectral properties of such matrices. To the best of the authors' knowledge, this phase was not yet identified in the literature.
3. In the heat and pairing phases, occurring respectively on timescales of order $O(\beta)$ and $O(e^{c\beta})$, we identify, under stronger conditions on the parameter values, the limiting dynamics as a forward or backward heat equation on the aligned manifold (leading, in the backward case, to metastable clustering) and a finite-dimensional system of ODEs describing sequential cluster merging along geodesics.

Together, these phases reconcile various previously identified dynamical regimes as different timescales of a single unified dynamical picture. Furthermore, our multi-phase analysis allows to relax some of the restrictive assumptions imposed by previous works, extending their applicability to more realistic scenarios.

Related works The model studied in this paper was introduced in [46], where the authors also identify the heat equation as describing the dynamics of the particle system in $\mathbb{R}^d$ in the large β regime. This limit emerges as a correction term in their analysis upon subtracting an appropriate leading order term from the prelimit equation. In this paper, by considering the dynamics on the sphere – resulting from the inclusion of the layer normalization in our model – we provide a justification for the spontaneous collapse of the system's state to a subspace where this correction term becomes of leading order, dominating the dynamics of the model on a certain timescale.

In [25, 24, 29, 13, 44], the authors identified the clustering behavior occurring in this and closely related models as $t \to \infty$ for β, N fixed. Analogous convergence results, under different assumptions, are provided in [18, 34], while quantitative contraction rates for such convergence are given in [16]. These works, however, do not address the dynamically meta-stable phases characterized by partial clustering numerically highlighted in [25]. This intermediate phase is explored in [9] in the large N limit for tokens distributed uniformly at initialization and in [23], where the authors study the formation of meta-stable clusters under the assumption that the system is initialized into well-separated configurations. Furthermore, in [23] the authors identify different dynamical timescales in the finite N case and characterize for the first time what we refer to in this paper as the pairing phase in the large β limit. In all these cases, however, the results are limited to the setting where the model's key, query and value matrices, Q, K and V, were multiples of the identity. More recently, the work [11] analyzes the stability of fixed points of the same model based on the eigendecomposition of Q, K, V under the weaker assumption that parameters satisfy a modified Wasserstein gradient flow condition ($Q^T K = V = D$), but does not study the dynamical landscape connecting such fixed points. Finally, in [3, 4], the authors discuss clustering for hardmax transformers. Our work provides a framework to combine the observations listed above in a unique dynamical picture.

Our modeling approach shares conceptual roots with the neural ODEs literature [14, 21]. However, a key distinction is that we consider N particles interacting through a mean-field PDE, as opposed to one in the previous references. This connects our work to the broader literature on mean-field models for neural networks [45, 36, 17, 2, 19], where timescale analysis has also been a subject of interest (see [7]). In contrast to these works, which typically focus on training dynamics, our study centers on the inference-time evolution of representations through network depth.

Finally, our research relates to the study of moderate scaling limits in interacting particle systems. For instance, Oelschläger [39] proved the convergence of certain systems to the porous medium equation with noise. These results were subsequently extended to cases without noise [41, 40], with different exponents [22], or employing different techniques and equations [12, 10, 42]. Another relevant line of work investigates the convergence of specific interacting particle systems to the heat equation, explored both numerically and theoretically [20, 8, 33].

2 Framework and notation

We consider the framework introduced in [46, 25, 24], modeling the transformer architecture as a discrete-time dynamical system governing the evolution of N *tokens* $\{x_i(\ell)\}_{i=1,\dots,N}$ through its layers via:

$$\begin{cases} x_i(\ell+1) = \mathcal{N}\left(x_i(\ell) + \dfrac{1}{Z_{\beta,i}(\ell)} \sum_{j=1}^N e^{\beta\langle Q_\ell x_i(\ell), K_\ell x_j(\ell)\rangle} V_\ell x_j(\ell) \right), \quad \ell = 0,\dots,L-1, \\ x_i(0) = x_i, \end{cases} \tag{1}$$

where $\mathcal{N} : \mathbb{R}^d \to \mathbb{S}^{d-1}$ denotes the normalization operator on the $d-$dimensional unit sphere $\mathbb{S}^{d-1}$, L denotes the depth of the transformer architecture and $Z_{\beta,i}(\ell) = \sum_{j=1}^N e^{\beta\langle Q_\ell x_i(\ell), K_\ell x_j(\ell)\rangle}$ is a normalization constant. The dynamics depends on the parameters with matrix values Q_ℓ, K_ℓ, and V_ℓ that represent the query, key, and value matrices at each layer, respectively.

In the spirit of neural ODEs [14], the authors then consider the infinite-depth limit of (1), leading to the following continuous-time model, describing the evolution of $x_i(t) : [0,\infty) \to \mathbb{S}^{d-1}$:

$$\dot{x}_i(t) = P_{x_i(t)} \left(\frac{1}{Z_{\beta,i}(t)} \sum_{j=1}^N e^{\beta\langle Q_t x_i(t), K_t x_j(t)\rangle} V_t x_j(t) \right). \tag{SA}$$

Here and throughout, $P_x y := y - \langle x, y\rangle x$ denotes the orthogonal projection of y onto the tangent space $T_x\mathbb{S}^{d-1}$, $\langle \cdot, \cdot \rangle$ is the Euclidean inner product in $\mathbb{R}^d$, and $Z_{\beta,i}(t) = \sum_{j=1}^N e^{\beta\langle Q_t x_i(t), K_t x_j(t)\rangle}$ is the time-dependent normalization factor. The parameter $\beta > 0$ is interpreted as the *inverse temperature*.

Remark 2.1. *The MLP would act as a drift term in the dynamics, whose consequence should still be investigated further. We expect different dynamical behavior depending on the relative scale of the MLP coefficients and the attention part. Although the framework allows for the inclusion of feedforward layers via a Lie-Trotter splitting scheme (see [26]), we choose to isolate exclusively the self-attention mechanism, both because our interest lies specifically in its dynamics, and for the sake of clarity. For the same reason, we assume, as in the works cited above, the parameter matrices to be shared across layers: $Q_t \equiv Q$, $K_t \equiv K$, and $V_t \equiv V$.*

As the positional information of each token is encoded in its initial condition, the dynamics (SA) is invariant under permutations of the particles' indices. This symmetry allows us to fully characterize the system's state through the particles' empirical measure $\mu(t) := \frac{1}{N}\sum_{i=1}^N \delta_{x_i(t)}$, where δ_x denotes the Dirac measure centered at x. The measure $\mu(t)$ evolves according to the continuity equation:

$$\begin{cases} \partial_t \mu + \operatorname{div}(\chi_\beta[\mu]\,\mu) = 0 & \text{on } \mathbb{R}_{\geq 0} \times \mathbb{S}^{d-1}, \\ \mu|_{t=0} = \mu(0) & \text{on } \mathbb{S}^{d-1}, \end{cases} \tag{2}$$

where the vector field $\chi_\beta[\mu] : \mathbb{S}^{d-1} \to T\mathbb{S}^{d-1}$ is defined as

$$\chi_\beta[\mu] = P_x \left(\frac{1}{Z_{\beta,\mu}(x)} \int_{\mathbb{S}^{d-1}} e^{\beta\langle Qx, Ky\rangle} Vy \,\mathrm{d}\mu(y) \right), \tag{3}$$

with $Z_{\beta,\mu}(x) := \int_{\mathbb{S}^{d-1}} e^{\beta\langle Qx, Ky\rangle} \mathrm{d}\mu(y)$. This formulation extends the token dynamics to a flow on the space $\mathcal{P}(\mathbb{S}^{d-1})$ of probability measures over the sphere $\mathbb{S}^{d-1}$, encompassing both empirical and absolutely continuous distributions.

3 Main results

As discussed in the introduction, in this paper we consider the limit as $\beta \to \infty$ of the dynamics (3). To present the dynamical scales arising in this limit, consider a formal Taylor expansion of the vector field $\chi^{\beta}[\mu]$ generated by a sufficiently smooth measure μ:

$$\chi_{\beta}[\mu](x) \approx \underbrace{\frac{\int e^{\beta\langle x',y\rangle} P_x V y \, \mu(x')\sigma(dy)}{\mu(x') \int e^{\beta\langle x',y\rangle}\sigma(dy)}}_{\text{(I)}} + \underbrace{\frac{\int e^{\beta\langle x',y\rangle} P_x V y \langle y - x', \nabla\mu(x')\rangle \, \sigma(dy)}{\mu(x') \int e^{\beta\langle x',y\rangle}\sigma(dy)}}_{\text{(II)}}, \tag{4}$$

where σ denotes the Lebesgue measure on $\mathbb{S}^{d-1}$ and $x' = K^{\top}Qx$. For large β, Laplace approximation suggests that (I) typically dominates (II) at initialization, giving rise to a first, fast dynamical phase:

- **Alignment Phase**: on a timescale of $O(1)$, the limiting dynamics are governed by a linear transport equation (Eq. (5) below) and the token distribution rapidly collapses onto a lower-dimensional subspace determined by the spectral properties of the matrix VK^TQ.

After the dynamics collapses to this low-dimensional subspace, we identify some classes of parameters for which the leading-order contribution to the vector field, approximated by term (I), vanishes. In such scenarios, the dynamics becomes governed by term (II), which involves the gradient of the measure μ. This gives rise to a second, intermediate phase:

- **Heat Phase**: operating on a timescale of $O(\beta)$ (achieved by rescaling time as $t' = t/\beta$), the dynamics within the previously identified subspace exhibits diffusive or anti-diffusive behavior. Depending on the model parameters (specifically the sign related to VK^TQ restricted to the subspace), this phase can lead to further concentration into distinct clusters (backward heat equation) or to smoothing/spreading of the distribution (forward heat equation).

In the attractive case, we identify the limiting dynamics up until the formation of clusters. We expect the clusters to be invariant in this timescale, and to interact only on much longer ones.

- **Pairing Phase**: on an exponentially long timescale in β (e.g., $O(e^{c\beta})$ for some $c > 0$, where c depends on the distance between clusters), the clusters formed in the previous phase sequentially merge. Typically, the closest pair of clusters collapses first, governed by a system of ODEs describing their interaction, eventually leading to a single clustered state.

We refer to Appendix E for a graphical representation of the three phases introduced above.

We outline the structure of the remainder of this Section. In in Section 3.1, we recall a quantitative result connecting the large N behavior of the ODE system with the behavior of the corresponding PDE in the relevant timescale. This will allow us to focus solely on the PDE analysis when we describe the three main dynamical phases in Sections 3.2, 3.3, and 3.4.

3.1 Large N convergence

To connect the timescales analysis above to the N-particle system of ODEs (SA), we consider the regime where $N \to \infty$ and $\beta = \beta_N \to \infty$ *slowly enough with respect to* N. This is relevant for context scaling techniques and LLMs (see introduction). To proceed, we use the following lemma:

Lemma 3.1. *Assume that the initial tokens* $\{x_i(0)\}_{i\in[N]}$ *are sampled independently and identically distributed from a reference measure* $\mu_0 \in \mathcal{P}(\mathbb{S}^{d-1})$*. Let* $\mu_t^{N,\beta}$ *be the empirical measure for particles* $\{x_i(t)\}_{i\in[N]}$ *evolving via the ODEs* (SA)*, and let* μ_t^{β} *be the solution to the continuity equation* (2) *with initial condition* μ_0*. Fix a time interval* $[0, T_\beta]$ *where* T_β *is a* β*-dependent timescale. If* $\beta = \beta_N$ *depends on* N *and diverges slowly enough as* $N \to \infty$*, then:*

$$W_1(\mu_t^{N,\beta}, \mu_t^{\beta}) \to 0 \quad \textit{as } N \to \infty,$$

uniformly on $[0, T_\beta]$.

Proof. This follows from the Dobrushin-type stability estimate: $W_1(\mu_t^{N,\beta}, \mu_t^{\beta}) \leq W_1(\mu_0^N, \mu_0)\, e^{L_\beta t}$, where L_β is a positive constant depending on the Lipschitz constant of the vector field χ_β, as discussed in [9]. The claimed convergence follows from $W_1(\mu_0^N, \mu_0) \to 0$ provided that $L_{\beta_N} T_{\beta_N}$ grows sufficiently slowly with N such that the overall term tends to zero. □

Our goal is to understand the behavior of $\mu_t^{N,\beta}$ in the joint limit $N, \beta_N \to \infty$. We denote the limiting distribution of μ_t^β as $\beta \to \infty$ by μ_t^∞. Following an argument analogous to that in [22], though in a different setting, we can decompose the convergence problem as:

$$W_1(\mu_t^{N,\beta}, \mu_t^\infty) \leq W_1(\mu_t^{N,\beta}, \mu_t^\beta) + W_1(\mu_t^\beta, \mu_t^\infty).$$

In our regime, Lemma 3.1 guarantees that the first term vanishes as $N \to \infty$. Consequently, the analysis of the N-particle system in this coupled limit reduces to studying the behavior of the solution μ_t^β to the continuity equation (2) as $\beta \to \infty$. The PDE analysis in this limit will therefore be the focus of the following sections.

3.2 The Alignment Phase

To characterize the limiting dynamics in the first phase we make the following assumptions:

Assumption 1. *Q, K, V are invertible square matrices.*

Assumption 2. *The probability measure μ_0 on S^{d-1} is absolutely continuous with respect to the Lebesgue measure on $\mathbb{S}^{d-1}$. Its density is bounded from above and below ($\min_{x\in\mathbb{S}^{d-1}} \mu_0(x) > 0$) and Lipschitz continuous.*

These technical assumptions, significantly milder than the ones made in most related works, are needed to guarantee that the terms appearing in the analysis of the limiting equation, e.g., the denominator in (5), are sufficiently well behaved. Under these conditions, we show that the limiting dynamics in this regime coincides with the formal Laplace approximation of term (I) in (4), i.e., the integrals in the definition of the vector fields can be replaced by the value of the integrand at the maximum point $x' = K^TQx/|K^TQx|$, leading to the significantly simplified expression (5) below.

Theorem 3.2. *Let Assumptions 1, 2 hold, then the solutions $\{\mu_\beta\}_\beta$ of the continuity equation (2) converge in $\mathcal{C}([0,T], \mathcal{P}(\mathbb{S}^{d-1}))$ to the solution μ^∞ of the partial differential equation:*

$$\begin{cases} \partial_t \mu(x) &= -\operatorname{div}\left(\mu(x) P_x \frac{VK^TQx}{|K^TQx|}\right), \\ \mu(0,x) &= \mu_0(x), \quad x \in \mathbb{S}^{d-1}. \end{cases} \tag{5}$$

Proof Sketch. To establish the result we must prove well-posedness of the family of equations leading to the desired limit and obtain sufficient regularity uniformly in β to ensure that the formal simplifications from (4)(I) to (5) are allowed. This is particularly important as the derivatives of the kernel tend to infinity in the limit $\beta \to \infty$. The core argument proceeds in three main steps. First, we establish the relative compactness of the family of trajectories $\{\mu^\beta\}_{\beta>0}$ in the space $\mathcal{C}([0,T], \mathcal{P}(\mathbb{S}^{d-1}))$ using a variant of Ascoli-Arzelà theorem and the boundedness of the vector field $\chi_\beta[\mu^\beta]$. The second, crucial step involves deriving uniform in β estimates on the regularity (i.e., Lipschitz bounds) of the vector field $\chi_\beta[\mu^\beta]$ along the solution trajectories μ^β. This is achieved by analyzing the concentration behavior of the kernel $e^{\beta\langle Qx,Ky\rangle}$ as $\beta \to \infty$, leveraging properties related to the cumulants of the Von Mises-Fisher distribution, and employing a continuation argument to propagate regularity over time. Finally, using the compactness and uniform regularity, we pass to the limit $\beta \to \infty$ in the weak formulation of the continuity equation (2). The uniform estimates allow us to conclude that μ^∞ is a solution of (5), while uniqueness follows from [5]. The full proof is deferred to Appendix A.1. □

A consequence of Theorem 3.2 is that in the large-β limit, the tokens, to leading order, evolve independently of each other, driven primarily by the structure of the Q, K, and V matrices. In this regime, self-attention behaves like a composition of linear layers followed by layer normalization, with minimal influence from inter-token interactions.

Combining the above result with Lemma 3.1 we obtain the following convergence result:

Corollary 3.3. *Under Assumptions 1, 2, for every $t > 0$ we have $W_1(\mu_t^{N,\beta}, \mu_t^\infty) \to 0$ as $N \to \infty$, provided that $\beta_N \to \infty$ slowly enough.*

Having established that μ^∞ is a solution of equation (5), we can investigate its long-time behavior. In particular, we show below that the support of μ_t^∞ is asymptotically flattened onto a lower-dimensional subspace determined by the spectral properties of the matrix $VK^\top Q$,

Proposition 3.4. *Let μ_0 be a probability measure on $\mathbb{S}^{d-1}$ absolutely continuous with respect to the Lebesegue measure, and let μ_t be the corresponding solution of* (5)*. Then for every $\nu \in \omega(\mu_0)$ (the ω-limit set of μ_0) it holds:*

$$supp(\nu) \subseteq E_{max} \cap \mathbb{S}^{d-1},$$

where E_{max} is the generalized eigenspace associated to the eigenvalue of VK^TQ with largest real part.

Proof. The proof is provided in Appendix A.2, where we reduce the analysis to a linear system of ODEs in $\mathbb{R}^d$ with matrix VK^TQ, identifying the corresponding asymptotics with the ones of (5). □

Remark 3.5. *At a first glance, this result might appear inconsistent with those of [11], since in some cases measures supported on $E_{\max}$ do not maximize the energy. However, this apparent discrepancy is a consequence of the order of the limits being taken, with $\beta \to \infty$ preceding $t \to \infty$ in our case.*

Proposition 3.4 demonstrates that the token representations rapidly collapse onto a lower-dimensional subspace determined by the model's matrices. This can be interpreted as the initial phase of the inference process, where information is compressed into a smaller, more relevant subspace. This phenomenon is consistent with the rank collapse observed, e.g., in [38, 27].

Remark 3.6. *Apart from the collapse to $E_{\max}$, one cannot in general conclude the existence of a limiting (stationary) dynamics for* (5)*. Indeed, it is not difficult to construct examples where the particles continue to rotate on the sphere indefinitely, e.g., when V is a rotation and $Q^TK = \mathrm{Id}$.*

Remark 3.7. *Recent works have studied transformer models with stochastic perturbations [48], where the token dynamics is influenced by random noise. In this setting, the convergence to the corresponding equation (5) (with an additional Laplacian term) is typically easier to establish due to the regularizing effect of the noise (see [39]).*

3.3 The Heat Phase

Having established the rapid collapse onto the subspace $E_{max} \cap \mathbb{S}^{d-1}$, we now investigate the slower evolution within this subspace, assuming that the initial measure μ_0 is supported in $E_{\max} \cap \mathbb{S}^{d-1}$ as a consequence of the previous analysis:

Assumption 3. *The initial condition μ_0 in the heat phase satisfies $supp(\mu_0) \subseteq E_{max} \cap \mathbb{S}^{d-1}$.*

Since the intersection $E_{max} \cap \mathbb{S}^{d-1}$ can be identified with a lower-dimensional sphere, specifically $\mathbb{S}^{\dim(E_{\max})-1}$, we will, with a slight abuse of notation, continue to denote it by $\mathbb{S}^{d-1}$.

To demonstrate that the heat equation, described in a different setting in [46], emerges as an intermediate dynamical phase due to the spherical geometry induced by LayerNorm, we assume:

Assumption 4. *$Q^TK|_{E_{max}} = \lambda_1 I$ and $V|_{E_{max}} = \pm\lambda_2 I$ when restricted to E_{max}, with $\lambda_1, \lambda_2 > 0$.*

Under this condition, E_{max} is an invariant subspace for Eq. (2) and, without loss of generality, we can suppose $\lambda_1, \lambda_2 = 1$.

Remark 3.8. *Assumption 4, for example, is satisfied under the global assumption $Q^\top K = S$ and $V = \pm S$, with S symmetric definite positive matrix. This is a fairly standard assumption in recent studies within this framework and it endows the model with an additional structure of gradient flow on $\mathcal{P}(\mathbb{S}^{d-1})$ with respect to a modified metric (see [11, 25]).*

In this regime, the vector field $\chi^\beta[\mu]$ vanishes on the support of μ as $\beta \to \infty$, but its rescaled version admits the formal limit (see Corollary B.2):

$$\beta\chi_\beta[\mu](x) \to \gamma\frac{\nabla_{\mathbb{S}^{d-1}}\mu}{\mu}(x), \quad \text{as } \beta \to \infty,$$

where $\gamma := \pm 1$, depending on the sign choice in the definition of V. This scaling of the vector field by β corresponds to a time rescaling $dt = \beta ds$, explaining the phase duration of order $O(\beta)$.

Proposition 3.9. *Let Assumption 4 hold and let $\mu_0^\infty \in \mathcal{P}(\mathbb{S}^{d-1})$ be the initial measure. Assume that there exist $T > 0$, k positive integer, and $\mu_t^\infty \in C([0,T], C^{k+3}(\mathbb{S}^{d-1}))$, with $\min_{x\in\mathbb{S}^{d-1}} \mu_t^\infty(x) > 0$ for all $t \in [0,T]$, such that μ_t^∞ solves the heat equation on $[0,T] \times \mathbb{S}^{d-1}$:*

$$\begin{cases} \partial_t\mu &= -\gamma\Delta\mu, \\ \mu(0) &= \mu_0, \end{cases} \tag{6}$$

where Δ denotes the Laplace-Beltrami operator on $\mathbb{S}^{d-1}$. Then, for large β, μ_t^∞ solves the mean-field PDE:

$$\begin{cases} \partial_t \mu & = -\operatorname{div}(\mu\, \beta \chi_\beta[\mu]) + R_\beta, \\ \mu(0) & = \mu_0, \end{cases}$$

where the residual term satisfies $R_\beta \to 0$ in $C([0,T], C^k(\mathbb{S}^{d-1}))$ as $\beta \to \infty$.

This proposition, whose proof is provided in Appendix B, characterizes the limiting dynamics within the lower-dimensional manifold, connecting the transformer model with a heat flow on the sphere, thereby justifying the name of this phase. Remarkably, this connection holds without the need for correction terms, in contrast to [46]. We now need to distinguish between two different cases:

- **Forward diffusion**. When $\gamma < 0$ in equation (6), the dynamics corresponds to a forward heat equation. In this setting, local existence and regularity for μ_t^∞ (and in particular the assumptions of Prop. 3.9) are automatically satisfied due to the smoothing properties of forward diffusion, provided that $\mu_0 \in C^{k+3}(\mathbb{S}^{d-1})$. Notably, interacting particle systems of the specific form given by equation (SA) (under the assumption $Q^\top K = \mathrm{Id} = -V$) have been studied in the literature and are known as *diffusion-velocity methods*; see, for instance, [33, 20, 8, 31, 32, 35].
- **Backward diffusion**. When $\gamma > 0$, the dynamics corresponds to a backward heat equation. In this case, the regularity assumptions on μ_0^∞ ensuring local existence and regularity are significantly more restrictive (e.g. requiring that μ_0 is in the Gevrey-$\frac{1}{2}$ space). Nonetheless, we construct explicit examples of solutions below. The backward heat equation is a prototypical ill-posed problem, which explains why the statement of Proposition 3.9 is necessarily weaker than that of Theorem 3.2.

A family of initial conditions μ_0 that satisfies the assumptions of Proposition 3.9 is given by

$$\mu_0 = \sum_{j=1}^{M} \alpha_j \mathcal{N}_{\mathbb{S}^{d-1}}(m_j, \sigma_j^2),$$

where $\alpha_j \geq 0$, $\sum_{j=1}^M \alpha_j = 1$, and $\mathcal{N}_{\mathbb{S}^{d-1}}(m, \sigma^2)$ denotes the heat kernel (the forward-in-time evolution under the heat semi-group $\exp(t\Delta)$ of a Dirac delta, analogous to a Gaussian $\mathcal{N}_{\mathbb{R}^d}(m, \sigma^2)$ in Euclidean space) centered at $m \in \mathbb{S}^{d-1}$ with concentration related to σ^2. By linearity of Δ, the explicit solution to $\partial_t \mu = -\gamma \Delta \mu$ is then given by:

$$\mu_t^\infty = \sum_{j=1}^{M} \alpha_j \mathcal{N}(m_j, \sigma_j^2 - \gamma t). \tag{7}$$

For forward diffusion ($\gamma < 0$), this solution is a smooth function for all $t \geq 0$. while in the backward case ($\gamma > 0$) this only holds for $t \in [0, T_{\min})$, where $T_{min} = \min_j \sigma_j^2$ is the time at which the first Gaussian component collapses to a Dirac delta δ_{m_j}. A more general class in which local existence and well-posedness hold in both the forward and backward directions is the set of positive, Gevrey-$1/2$, functions.

Motivated by the aggregation behavior observed in the finite-β particle system, we conjecture that the collapsed δ_{m_j} remains invariant under the limiting dynamics, while other components continue evolving independently according to the backward heat equation until their respective collapse times. From the practical perspective, this observation suggests that the transformer's behavior in this regime can be interpreted as a form of regularized denoising (when β is finite) acting on the input. This aligns with the clustering phenomena extensively studied in previous works on the model. The dynamics in this phase, governed by a heat equation, drive the formation of distinct token clusters (via backward diffusion) or the smoothing of the token distribution (via forward diffusion). This behavior can be interpreted as a representation refinement stage, where tokens are organized into more defined semantic groups.

Remark 3.10. *In [25], a simplified model, referred to as the Unnormalized Self-Attention (USA) model, is proposed, where the normalization factor $Z_{\beta,\mu}(x)$ is replaced by a constant Z_β, significantly simplifying the mathematical analysis. By choosing $Z_\beta = \frac{1}{\beta} \int_{\mathbb{S}^{d-1}} e^{\beta\langle x,y\rangle}\, d\sigma(y)$ (or equivalently by rescaling time), the limiting behavior of the model no longer yields the heat equation, but rather the porous medium equation: $\partial_t \mu = \Delta(\mu^2)$. Even in this case, the convergence of particles system to this nonlinear PDE has been extensively studied (see for example [22, 33, 39, 41] or [47] for $\mathbb{S}^{d-1}$).*

3.4 Pairing Phase

The initial conditions we consider for the dynamics on longer timescales must be compatible with the steady states of the preceding phase. Motivated by the discussion at the end of the previous section, we therefore formulate the following assumption:

Assumption 5. *The initial condition μ_0 in the pairing phase can be written as $\mu_0 = \sum_{j=1}^{m} \alpha_j \delta_{x_j}$ for an $m \in \mathbb{N}$, with $x_j \in \mathbb{S}^{d-1}$, $\alpha_j > 0 \, \forall j \in \{1, ..., m\}$ and $\sum_{j=1}^{m} \alpha_j = 1$.*

Under this assumption, further supposing for the sake of clarity that $\alpha_j = 1/m$ for all $j \in \{1, \dots, m\}$, we can interpret each cluster as a particle, and the dynamics of the system is given by the set of ODEs (SA). In the regime of large β, clusters interact very weakly due to their separation and the exponential tails (in β) of the interaction kernel, resulting in exponentially long timescales for the nontrivial dynamics. Here, analogously to [3], interactions are dominated by the closest pair of clusters $(\underline{i}, \underline{j})$, assumed unique, satisfying $\langle x_{\underline{i}}, x_{\underline{j}} \rangle = \max_{i \neq j} \langle x_i, x_j \rangle$ at initialization. We note that this hardmax particle interaction, as well as the timescale where it arises in the large β limit, was introduced in [23, Section 6] in the case $d = 2$. We present an analogous result here in arbitrary dimension, without claiming originality, to provide a complete dynamical picture across phases.

Proposition 3.11. *The solutions $x_i(t)$ of the ODE system* (SA)*, under Assumptions 4 and positive V, with the rescaled time $dt = e^{\beta(1-\langle x_{\underline{i}}, x_{\underline{j}} \rangle)} ds$, converge as $\beta \to \infty$ to the solutions of the system:*

$$\begin{cases} \dot{y}_k(t) = \begin{cases} P_{y_{\underline{i}}}(y_{\underline{j}}) & \textit{if } k = \underline{i}, \\ P_{y_{\underline{j}}}(y_{\underline{i}}) & \textit{if } k = \underline{j}, \\ 0 & \textit{otherwise}, \end{cases} \\ y_i(0) = x_i(0) \end{cases}$$

on finite intervals $[0, T_\epsilon]$, with T_ϵ such that $\langle y_{\underline{i}}, y_{\underline{j}} \rangle \leq 1 - \epsilon$ throughout the interval, for any $\epsilon > 0$.

In other words, all clusters remain stationary except for the closest pair, which collapses along the geodesic connecting them, in a time exponential in β. Note that this result only holds up to an arbitrary moment before the first collapse. We refer to [23, Section 6] for a detailed explanation of the challenges to bypass this limit and a proof of an analogous result until and beyond the collapse time in a related but simplified model. The above proposition is proven for completeness in Appendix C.

This final, slow phase models the sequential merging of the closest token clusters. This can be interpreted as the construction of higher-order abstractions, where previously formed groups are hierarchically combined to create more complex representations.

4 Numerical experiments

This section presents numerical simulations of the transformer model in Eq. (1). All experiments are conducted in dimension $d = 3$ or or $d = 2$ to facilitate visualization and are designed to validate our theoretical findings. The attention mechanism is implemented using the official PyTorch function `torch.nn.functional.scaled_dot_product_attention()` and the experiments are performed on a single Nvidia H100. The code is available at [28].

First Phase Dynamics. Figure 1 illustrates the dynamics of the alignment phase, showing distinct behaviors based on the parameters choices for Q, K, V. For both scenarios presented in Figure 1, the initial state consists of $N = 10^4$ tokens sampled independently and identically uniformly from the sphere $\mathbb{S}^2$. We set the inverse temperature parameter $\beta = 30$ and use a time step of $dt = 10^{-2}$.

- **Scenario 1a (collapse to 1D subspace).** The matrix VK^TQ is chosen such that it possesses a unique eigenvalue with maximal real part. As predicted by our theory, this configuration leads to the tokens collapsing onto a one-dimensional subspace (i.e. two antipodal points).
- **Scenario 1b (non-gradient flow dynamics and rotation).** This example employs a parameter choice for Q, K, V that falls outside the gradient flow regime. Nevertheless, the tokens are observed to collapse toward a two-dimensional subspace (a great circle), accompanied by a collective rotation of the particles along this circle.

Both observed behaviors are consistent with the results in Theorem 3.2 and Proposition 3.4.

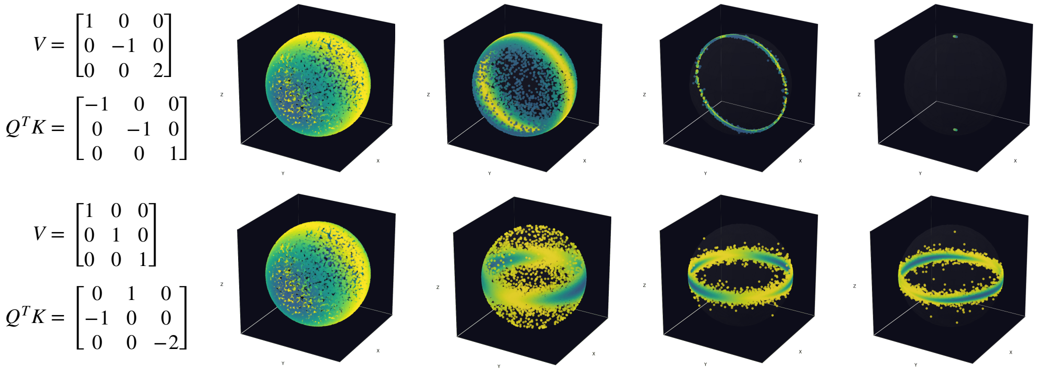

Figure 1: Simulations of two different scenarios (one per row, four timesteps) for the first phase.

Second Phase Dynamics. We support the conclusions of section 3.3 through two examples, comparing empirical dynamics with analytical solutions of backward and forward diffusion equations.

- **Scenario 2a (collapse to 2D subspace and backward diffusion).** For the experiment in Figure 2 we set $\beta = 10$. The initial configuration comprises $N = 10^4$ i.i.d. tokens. Their elevation angle ψ is sampled uniformly on $[-\frac{\pi}{2}, \frac{\pi}{2}]$, while their azimuthal angles, $\theta_i \in [0, 2\pi)$, are distributed according to the mixture density $g(\theta)$:
$$g(\theta) := 0.2 \cdot \mathcal{N}(\theta; \pi/2, \sigma_0) + 0.5 \cdot \mathcal{N}(\theta; 0, \sigma_0) + 0.3 \cdot \mathcal{N}(\theta; 4\pi/3, \sigma_0)$$
where $\mathcal{N}(\cdot; \mu, \sigma_0)$ denotes the probability density function of a wrapped normal distribution on $\mathbb{S}^1$ with mean μ and standard deviation $\sigma_0 = 0.2$. The parameters Q, K, and V are chosen so that, after the first phase, the tokens collapse onto the xy-plane, with distribution $g(\theta)$. The analytical solution to the backward heat equation with initial condition $g(\theta)$ (computed as in Eq. (7)) is plotted as a red curve in Figure 2. The positions of the clusters agree with this solution, numerically confirming the predictions of Proposition 3.9.
- **Scenario 2b (forward diffusion comparison).** In Figure 3, we compare the empirical token distribution with the analytical solution of the forward heat equation characterizing a possible example of the second phase of the dynamics. Specifically, we simulate the evolution of 5×10^4 tokens, initially sampled from a superposition of three Gaussian densities on $\mathbb{S}^1$, through the transformer model with parameters $\beta = 50$, $d = 2$, $Q = K = \mathrm{Id}$, $V = -\mathrm{Id}$, and $dt = 10^{-3}$. The analytical solution of the forward heat equation (in red) closely matches the token distribution histogram (in blue) over time (i.e., depth). Note that, as expected, the forward diffusion process is significantly more stable numerically than the backward one.

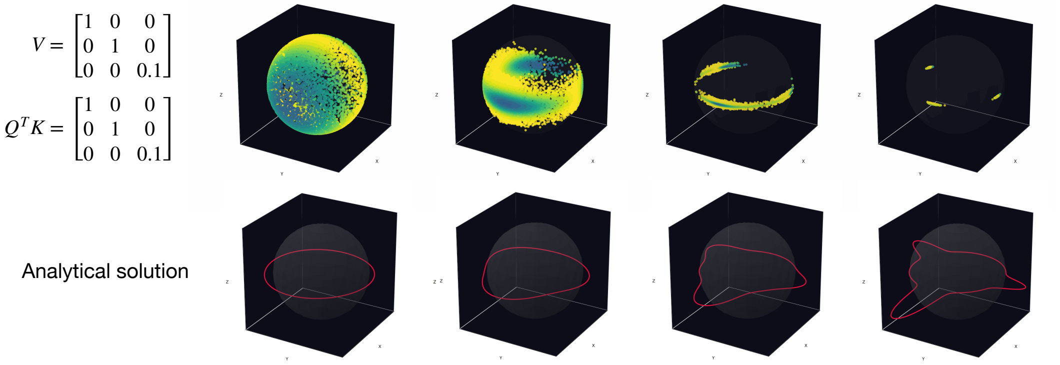

Figure 2: Numerical simulation of the backward scenario for the second phase on $\mathbb{S}^2$.

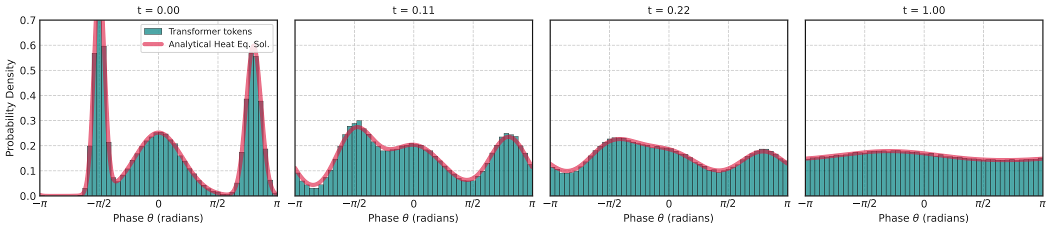

Figure 3: Evolution of the tokens distribution in the forward scenario for the second phase in $\mathbb{S}^1$.

5 Conclusions

This work provides a mathematical analysis of token dynamics in mean-field transformer models within the moderate interaction regime, where interaction strength (β) scales with context size (N), motivated by scaling practices in modern LLMs. Our study reveals a fundamental multiscale structure governing the evolution of token representations through network depth in this setting. More specifically, we showed that, under this scaling, the system progresses through a sequence of three different dynamical phases characterized by qualitatively distinct dynamical behavior, operating on separated timescales. Through our analysis, we offer a unified dynamical picture describing how deep transformers might achieve progressive representation refinement.

This unified dynamical picture is, however, not yet fully rigorous. Indeed, while we establish convergence results for the alignment phase and provide partial justification for the intermediate and late phases under specific assumptions, a complete mathematical treatment of the full dynamics - particularly of the backward heat regime beyond the first collapse and of the slow clustering interactions - remains an open challenge due to significant technical difficulties in the analysis of the strongly unstable limiting equations. Additionally, characterizing phase transitions in self-attention for different N, β scalings is an interesting separate question, with progress made in [15] under assumptions on inter-token angles.

Furthermore, while the first dynamical phase has quite general assumptions on the parameter matrices, the following phases still require relatively limited assumptions (although less limited than in most previous works). Relaxing these assumptions further, in particular in the case of non-gradient dynamics, would constitute an interesting, but also technically quite challenging, avenue of future research.

There are several important directions in which our work could be extended. Most notably, incorporating the MLP, which could be interpreted as introducing a drift term in the dynamics, acting independently on each token without accounting for mutual interactions. Another natural extension involves studying the dynamics under more general parameter settings. For instance, during the heat phase, we assume that $Q^T K = S$ is symmetric positive definite, which holds, for example, when $Q = K$. This “shared-QK” assumption is not novel and has been adopted in prior empirical work (e.g., in [30]). While different choices of these parameters (both MLP and Attention matrices) can have a dramatic effect on model behavior, with adversarial choices potentially leading to qualitatively different dynamics from the one predicted in this paper, we believe our results to be a relevant first step towards understanding the development of representations in transformers, capturing some important qualitative features of these models as shown in [25].

A further direction of future research consists in providing sufficient conditions for the stability of the space E_{max} emerging in the alignment phase under the prelimit model (i.e., for large but finite β), thereby justifying Assumption 3 and, ultimately, connecting in a rigorous way the alignment and heat phases identified in this paper.

While the path from this theoretical analysis to direct application is not immediate, we believe our work opens several potential avenues for future investigation. The characterization of the alignment phase, for instance, offers a potential mechanism for interpreting how token representations evolve into learned subspaces. Finally, by focusing on the "moderate interaction regime", we hope our analysis provides a theoretical foundation for a more principled understanding of parameter scaling, particularly as models are adapted for longer contexts.

Acknowledgements

AA and GB acknowledge partial support by the Institute of Mathematical Statistics and Actuarial Sciences at the University of Bern. FP acknowledges support by the American Mathematical Society through the Bergman fellowship. AA thanks B. Geshkovski, J. Mattingly, Y. Polyanskiy, and P. Rigollet for interesting discussions. All authors thank the anonymous reviewers for comments that have improved the quality of the paper. Calculations were performed on UBELIX, the HPC cluster at the University of Bern.

References

[1] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1948.

[2] Andrea Agazzi and Jianfeng Lu. Global optimality of softmax policy gradient with single hidden layer neural networks in the mean-field regime. In *International Conference on Learning Representations (ICLR 2021)*, 2021.

[3] Albert Alcalde, Giovanni Fantuzzi, and Enrique Zuazua. Clustering in pure-attention hardmax transformers and its role in sentiment analysis. *SIAM Journal on Mathematics of Data Science*, 7(3):1367–1393, 2025.

[4] Albert Alcalde, Borjan Geshkovski, and Domènec Ruiz-Balet. Attention's forward pass and frank-wolfe. *arXiv preprint arXiv:2508.09628*, 2025.

[5] Luigi Ambrosio, Luis Caffarelli, Michael G Crandall, Lawrence C Evans, and Nicola Fusco. Transport equation and cauchy problem for non-smooth vector fields. *Calculus of Variations and Nonlinear Partial Differential Equations: With a historical overview by Elvira Mascolo*, pages 1–41, 2008.

[6] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[7] Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *Foundations of Computational Mathematics*, pages 1–84, 2024.

[8] Yann Brenier. Geometric diffusions of 1-currents. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 26, pages 831–846, 2017.

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

[10] Martin Burger and Antonio Esposito. Porous medium equation and cross-diffusion systems as limit of nonlocal interaction. *Nonlinear Analysis*, 235:113347, 2023.

[11] 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 A*, 383(2298):20240233, 2025.

[12] José Antonio Carrillo, Katy Craig, and Francesco S Patacchini. A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58:1–53, 2019.

[13] Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers. *arXiv preprint arXiv:2501.18322*, 2025.

[14] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

[15] Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Critical attention scaling in long-context transformers. *arXiv preprint arXiv:2510.05554*, 2025.

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

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

[18] Christopher Criscitiello, Quentin Rebjock, Andrew D McRae, and Nicolas Boumal. Synchronization on circles and spheres with nonlinear interactions. *arXiv preprint arXiv:2405.18273*, 2024.

[19] Valentin De Bortoli, Alain Durmus, Xavier Fontaine, and Umut Simsekli. Quantitative propagation of chaos for sgd in wide neural networks. *Advances in Neural Information Processing Systems*, 33:278–288, 2020.

[20] Pierre Degond and Francisco-José Mustieles. A deterministic approximation of diffusion equations using particles. *SIAM Journal on Scientific and Statistical Computing*, 11(2):293–310, 1990.

[21] Weinan E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 1(5):1–11, 2017.

[22] Alessio Figalli and Robert Philipowski. Convergence to the viscous porous medium equation and propagation of chaos. *ALEA Lat. Am. J. Probab. Math. Stat*, 4:185–203, 2008.

[23] Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model. *arXiv preprint arXiv:2410.06833*, 2024.

[24] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36, 2024.

[25] 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.

[26] Borjan Geshkovski, Philippe Rigollet, and Domènec Ruiz-Balet. Measure-to-measure interpolation using transformers. *arXiv preprint arXiv:2411.04551*, 2024.

[27] Alessio Giorlandino and Sebastian Goldt. Two failure modes of deep transformers and how to avoid them: a unified theory of signal propagation at initialisation. *arXiv preprint arXiv:2505.24333*, 2025.

[28] GitHub-Repository. `https://github.com/gbruno16/multiscale_transformers`.

[29] Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking. *Advances in Neural Information Processing Systems*, 37:115652–115681, 2024.

[30] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[31] Gilles Lacombe. Analyse d'une équation de vitesse de diffusion. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 329(5):383–386, 1999.

[32] Gilles Lacombe and Sylvie Mas-Gallic. Presentation and analysis of a diffusion-velocity method. In *Esaim: Proceedings*, volume 7, pages 225–233. EDP Sciences, 1999.

[33] Pierre-Louis Lions and Sylvie Mas-Gallic. Une méthode particulaire déterministe pour des équations diffusives non linéaires. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 332(4):369–376, 2001.

[34] Johan Markdahl, Johan Thunberg, and Jorge Gonçalves. Almost global consensus on the n-sphere. *IEEE Transactions on Automatic Control*, 63(6):1664–1675, 2017.

[35] Sylvie Mas-Gallic. The diffusion velocity method: a deterministic way of moving the nodes for solving diffusion equations. *Transport Theory and Statistical Physics*, 31(4-6):595–605, 2002.

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

[37] Ken M Nakanishi. Scalable-softmax is superior for attention. *arXiv preprint arXiv:2501.19399*, 2025.

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

[39] Karl Oelschläger. A law of large numbers for moderately interacting diffusion processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 69(2):279–322, 1985.

[40] Karl Oelschläger. Large systems of interacting particles and the porous medium equation. *Journal of differential equations*, 88(2):294–346, 1990.

[41] Karl Oelschläger. A sequence of integro-differential equations approximating a viscous porous medium equation. *Zeitschrift für Analysis und ihre Anwendungen*, 20(1):55–91, 2001.

[42] Thierry Paul and Emmanuel Trélat. Universal approximations of quasilinear pdes by finite distinguishable particle systems. *arXiv preprint arXiv:2501.11387*, 2025.

[43] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[44] Yury Polyanskiy, Philippe Rigollet, and Andrew Yao. Synchronization of mean-field models on the circle. *arXiv preprint arXiv:2507.22857*, 2025.

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

[46] Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré. Sinkformers: Transformers with doubly stochastic attention. In *International Conference on Artificial Intelligence and Statistics*, pages 3515–3530. PMLR, 2022.

[47] Anna Shalova. Noisy gradient flows: with applications in machine learning. *PhD Thesis*, 2025.

[48] Anna Shalova and André Schlichting. Solutions of stationary mckean-vlasov equation on a high-dimensional sphere and other riemannian manifolds. *arXiv preprint arXiv:2412.14813*, 2024.

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

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Abstract and introduction summarize the claims made in the paper that are proved or discussed in the following sections.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Yes, see section conclusions.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The assumptions are reported in the corresponding sections, while the proofs are in the supplementary materials

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The details of the experiments are reported in the corresponding section.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: An anonymous github repository is provided.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [NA]

   Justification: The experiments don't have training and test.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [NA]

   Justification: The kind of experiments does not need statistical significance, they are solution of ODEs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: See the section about numerical experiments.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: Anonymity and all the other rules have been respected.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [NA]

    Justification: It is a theoretical paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: It is a theoretical paper.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [NA]

    Justification: It is a theoretical paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: See numerical experiments section and the github link.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Proofs of the alignment phase

This section is divided into two parts. The first part contains the proof of Theorem 3.2, while the second one contains the proof of Proposition 3.4.

A.1 Moderate scaling limit

Consider the family $\{\mu^\beta\}_{\beta\geq 0}$ of solutions to the usual continuity equation (2):

$$\begin{cases} \partial_t \mu^\beta & = -div(\mu^\beta \chi_\beta[\mu^\beta]), \\ \mu^\beta(0,x) & = \mu_0(x), \quad x \in \mathbb{S}^{d-1}, \end{cases}$$

where χ_β is the vector field given by:

$$\chi_\beta[\mu](x) = \frac{\int_{S^{d-1}} e^{\beta\langle x,By\rangle} P_x V y \mu(dy)}{\int_{S^{d-1}} e^{\beta\langle x,By\rangle} \mu(dy)}.$$

Remark A.1. *For notational simplicity, we will refer to the matrix $Q^t K$ in the main body as B.*

Remark A.2. *In the following C will be a constant depending only on V, B, d and μ_0. Its value may change line by line.*

Under assumptions 1, 2, i.e.:

- A1: Q, B are invertible square matrices.
- A2: The probability measure μ_0 on S^{d-1} is absolutely continuous with respect to the Lebesgue measure. Its density is bounded and Lipschitz continuous and its minimum satisfies $\min_{x\in\mathbb{S}^{d-1}} \mu_0(x) > 0$.

one can prove Theorem 3.2: μ_β converges weakly to μ_∞ in $C([0,T];\mathcal{P}(S^{d-1}))$ where μ_∞ is the unique solution of the PDE:

$$\begin{cases} \partial_t \mu & = -div(\mu \frac{P_x V B^T x}{|B^T x|}), \\ \mu(0,x) & = \mu_0(x), \quad x \in \mathbb{S}^{d-1}. \end{cases}$$

and the metric in $\mathcal{C}([0,T],\mathcal{P}(\mathbb{S}^{d-1}))$ is given by:

$$d(\mu,\nu) := \sup_{t\in[0,T]} \rho(\mu_t,\nu_t) = \sup_{t\in[0,T]} \sup_{f\in BL(\mathbb{S}^{d-1})} \left| \int_{\mathbb{S}^{d-1}} f(x)\mu_t(dx) - f(x)\nu_t(dx) \right|,$$

where $BL(\mathbb{S}^{d-1})$ is the set of all the Lipschitz continuous functions on $\mathbb{S}^{d-1}$ which are bounded together with their Lipschitz constant by 1.

The idea of the proof follows five steps:

- Relative compactness in $C([0,T];\mathcal{P}(S^{d-1}))$,
- Bounds on cumulants of the Von Mises-Fisher distribution,
- Prove a relationship between the derivatives of μ and the regularity of the vector field,
- Apply a continuation argument to show the uniform regularity of $\chi_\beta[\mu_\beta]$,
- Use the regularity to pass to the limit in the PDE.

A.1.1 Relative compactness

Proposition A.3. *The set $\{\mu_\beta\}_{\beta\geq 0}$ is relatively compact in $\mathcal{C}([0,T],\mathcal{P}(\mathbb{S}^{d-1}))$.*

Proof. By Prokhorov's theorem, since $\mathbb{S}^{d-1}$ is compact, we can conclude that $\mathcal{P}(\mathbb{S}^{d-1})$ is weakly compact. Since ρ metricizes the weak topology, then also $(\mathcal{P}(\mathbb{S}^{d-1}),\rho)$ is compact. To apply Ascoli-Arzelà theorem, we just need the equicontinuity of the set $\{\mu^\beta\}_{\beta>0}$.

Given $0 \leq s \leq t \leq T$ and $\beta > 0$:

$$\begin{aligned}\rho(\mu_s^\beta, \mu_t^\beta) &= \sup_{f \in BL(\mathbb{S}^{d-1})} \left| \int_{\mathbb{S}^{d-1}} f(x)(\mu_t^\beta(dx) - \mu_s^\beta(dx)) \right| \\ &\leq \sup_{f \in BL(\mathbb{S}^{d-1})} \int_s^t \left| \int_{\mathbb{S}^{d-1}} \langle \nabla f(x), \chi^\beta[\mu_u^\beta](x) \rangle \mu_u^\beta(dx) du \right| \\ &\leq \int_s^t \int_{\mathbb{S}^{d-1}} |\chi^\beta[\mu_u^\beta](x)| \mu_u^\beta(dx) du \leq |t-s|.\end{aligned}$$

This is sufficient to conclude the proof. □

A.1.2 Bounds on the vector field

The aim of the following paragraphs is to obtain some bounds on $D^i \chi_\beta[\mu]$, $i = 0, 1, 2$.

To fix the notation we define the probability measure $\nu_x^{\mu,\beta,B}$ on $\mathbb{S}^{d-1}$ as:

$$\nu_x^{\mu,\beta,B}(dy) := \frac{e^{\beta\langle x, By \rangle} \mu(dy)}{\int_{S^{d-1}} e^{\beta\langle x, By \rangle} \mu(dy)}.$$

Remark A.4. *The measure $\nu_x^{\sigma,\beta,B}$ is the Von Mises-Fisher distribution with mean direction $\frac{B^T x}{|B^T x|}$ and concentration parameter $\beta|B^T x|$. Some properties of this distribution are studied later.*

Then the vector field $\chi_\beta[\mu]$ can be written as:

$$\chi_\beta[\mu](x) = P_x V \left(\mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y] \right) = P_x V \left(\mathbb{E}_{\nu_x^{\mu,1,Id}}[Y] \right) \circ (\beta B^T x),$$

where $\circ$ denotes the composition with respect to the parameter x of the measure $\nu_x^{\mu,\beta,B}$.

Lemma A.5. *The derivatives of the vector field $\chi_\beta[\mu]$ are bounded by:*

$$\begin{aligned}&|\chi_\beta[\mu](x)| \leq C, \\ &|D_x^1 \chi_\beta[\mu](x)| \leq C \left(1 + \beta\, |\mathbb{E}_{\nu_x^{\mu,\beta,B}}[\left(Y - \mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y] \right)^{\otimes 2}]| \right), \\ &|D_x^2 \chi_\beta[\mu](x)| \leq C \left(1 + \beta\, |\mathbb{E}_{\nu_x^{\mu,\beta,B}}[\left(Y - \mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y] \right)^{\otimes 2}]| + \beta^2\, |\mathbb{E}_{\nu_x^{\mu,\beta,B}}[\left(Y - \mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y] \right)^{\otimes 3}]| \right)\end{aligned}$$

where C is a constant depending only on V, B, d.

Proof. Let's compute the derivatives of $\chi_\beta[\mu]$:

$$\begin{aligned}|\chi_\beta[\mu](x)| &\leq |P_x V| |\mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y]| \leq C_V, \\ |D_x^1 \chi_\beta[\mu](x)| &\leq |D_x^1 P_x V| | \left(\mathbb{E}_{\nu_x^{\mu,1,Id}}[Y] \circ (\beta B^T x) \right) | + |P_x V| |D_x^1 \left(\mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y] \right) | \\ &\leq C_V \left(1 + |D_x^1 \left(\mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y] \right) | \right), \\ |D_x^2 \chi_\beta[\mu](x)| &\leq |D_x^2 P_x V| |\mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y]| + 2 |D_x^1 P_x V| |D_x^1 \left(\mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y] \right) | + |P_x V| |D_x^2 \left(\mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y] \right) | \\ &\leq C_V (1 + 2 |D_x^1 \left(\mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y] \right) | + |D_x^2 \left(\mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y] \right) |\end{aligned}$$

Hence we need to compute the derivatives with respect to x of $\mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y] = \mathbb{E}_{\nu_x^{\mu,1,Id}}[Y] \circ (\beta B^T x)$. Thanks to the Faa di Bruno formula:

$$D_x^n \mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y] = \sum_{\pi \in \Pi_n} \left((D_x^{|\pi|} \mathbb{E}_{\nu_x^{\mu,1,Id}}[Y])|_{\beta B^T x} \circ \bigotimes_{P \in \pi} D^P(\beta B^T x) \right),$$

with Π_n the set of all the possible partitions of $\{1, ..., n\}$ and $\otimes$ the tensor product. The previous expression can be bounded by $C_B \sum_{l=0}^n \beta^l \|(D_x^l \mathbb{E}_{\nu_x^{\mu,1,Id}}[Y])|_{\beta B^T x}\|$, where C_B is a constant depending on the matrix B.

Thus, the aim is to compute a bound for $D_x^n \mathbb{E}_{\nu_x^{\mu,1,Id}}[Y]$. This is related to the $d-$ dimensional cumulants (tensors) of the distribution $\nu_x^{\mu,1,Id}$. Indeed, we can write:

$$\begin{aligned}
D_x^n \left(\mathbb{E}_{\nu_x^{\mu,1,Id}}[Y]\right)_x &= D_v^n \left(\mathbb{E}_{\nu_{x+v}^{\mu,1,Id}}[Y]\right)_{v=0} \\
&= D_v^n \left(\frac{\int e^{\langle x+v,y\rangle} y\ \mu(dy)}{\int e^{\langle x+v,y\rangle}\ \mu(dy)}\right)_{v=0} \\
&= D_v^n \left(\frac{\int e^{\langle v,y\rangle} y\ e^{\langle x,y\rangle} \mu(dy)}{\int e^{\langle x,y\rangle}\ \mu(dy)} \frac{\int e^{\langle x,y\rangle}\ \mu(dy)}{\int e^{\langle v,y\rangle} e^{\langle x,y\rangle}\ \mu(dy)}\right)_{v=0} \\
&= D_v^n \left(\frac{\int e^{\langle v,y\rangle} y\ \nu_x^{\mu,1,Id}(dy)}{\int e^{\langle v,y\rangle}\ \nu_x^{\mu,1,Id}(dy)}\right)_{v=0} \\
&= D_v^{n+1} \left(\log \mathbb{E}_{\nu_x^{\mu,1,Id}}[e^{\langle v,Y\rangle}]\right)_{v=0}.
\end{aligned} \tag{8}$$

It is well known that the first three cumulants correspond to the central moments:

$$\begin{aligned}
D_x^0 \left(\mathbb{E}_{\nu_x^{\mu,1,Id}}[Y]\right)_x &= \mathbb{E}_{\nu_x^{\mu,1,Id}}[Y], \\
D_x^1 \left(\mathbb{E}_{\nu_x^{\mu,1,Id}}[Y]\right)_x &= \mathbb{E}_{\nu_x^{\mu,1,Id}}[\left(Y - \mathbb{E}_{\nu_x^{\mu,1,Id}}[Y]\right)^{\otimes 2}], \\
D_x^2 \left(\mathbb{E}_{\nu_x^{\mu,1,Id}}[Y]\right)_x &= \mathbb{E}_{\nu_x^{\mu,1,Id}}[\left(Y - \mathbb{E}_{\nu_x^{\mu,1,Id}}[Y]\right)^{\otimes 3}].
\end{aligned} \tag{9}$$

The thesis then follows by replacing these equalities in the initial bounds (after renaming C_V and C_B). □

Lemma A.6. *Let σ be the uniform measure on $\mathbb{S}^{d-1}$. Then:*

$$\begin{aligned}
|\mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[Y]| &\le 1, \\
|\mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[\left(Y - \mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[Y]\right)^{\otimes 2}]| &\le C\frac{1}{\beta}, \\
|\mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[\left(Y - \mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[Y]\right)^{\otimes 3}]| &\le C\frac{1}{\beta^2}.
\end{aligned}$$

Proof. By (8) and Schwarz's theorem the three tensors are invariant by permutations of the indices and by definition of $\nu_x^{\sigma,\beta,Id}$ they are also invariant by rotations that fix x. Hence (see lemma D.2) they must have the form:

$$\begin{aligned}
\mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[Y] &= \alpha_1 x, \\
\mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[\left(Y - \mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[Y]\right)^{\otimes 2}] = D_x^1 \left(\mathbb{E}_{\nu_x^{\sigma,1,Id}}[Y]\right)_{\beta x} &= \alpha_2 x \otimes x + \beta_2 I, \\
\mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[\left(Y - \mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[Y]\right)^{\otimes 3}] = D_x^2 \left(\mathbb{E}_{\nu_x^{\sigma,1,Id}}[Y]\right)_{\beta x} &= \alpha_3 x \otimes x \otimes x + \beta_3 Sym(x \otimes Id).
\end{aligned} \tag{10}$$

Where $Sym(x \otimes Id) = x_i\delta_{jk} + x_j\delta_{ik} + x_k\delta_{ij}$. We need to compute the coefficients $\alpha_1, \alpha_2, \beta_2, \alpha_3, \beta_3$. Define $A(\beta) = \int_{S^{d-1}} \langle x, y\rangle\ \nu_x^{\sigma,\beta,Id}(dy)$. Similarly to what has been done in (8), one can relate this to the cumulants of $\langle x, Y\rangle$ by noticing that:

$$\begin{aligned}
\partial_\beta^n \left(A(\beta)\right)_\beta = \partial_t^n \left(A(\beta + t)\right)_{t=0} &= \partial_t^n \left(\frac{\int e^{(\beta+t)\langle x,y\rangle} \langle x, y\rangle \sigma(dy)}{\int e^{(\beta+t)\langle x,y\rangle} \sigma(dy)}\right)_{t=0} \\
&= \partial_t^n \left(\frac{\int e^{t\langle x,y\rangle} \langle x, y\rangle e^{\beta\langle x,y\rangle} \sigma(dy)}{\int e^{t\langle x,y\rangle} e^{\beta\langle x,y\rangle} \sigma(dy)}\right)_{t=0} = \partial_t^n \left(\frac{\int e^{t\langle x,y\rangle} \langle x, y\rangle \nu_x^{\sigma,\beta,Id}(dy)}{\int e^{t\langle x,y\rangle} \nu_x^{\sigma,\beta,Id}(dy)}\right)_{t=0} \\
&= \partial_t^{n+1} \left(\log \mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[e^{t\langle x,Y\rangle}]\right)_{t=0}.
\end{aligned}$$

This give us immediately the following identities:

$$\begin{aligned} A(\beta) &= E_{\nu_x^{\sigma,\beta,Id}}[\langle x, Y\rangle], \\ A'(\beta) &= E_{\nu_x^{\sigma,\beta,Id}}[(\langle x, Y\rangle - A(\beta))^2], \\ A''(\beta) &= E_{\nu_x^{\sigma,\beta,Id}}[(\langle x, Y\rangle - A(\beta))^3]. \end{aligned} \tag{11}$$

Suppose without loss of generality that $x = e_1$. Then α_1 is given by:

$$\alpha_1 = \mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[\langle e_1, Y\rangle] = A(\beta). \tag{12}$$

The coefficients α_2 and β_2 can be obtained comparing the representations in (10) with the representations in (9), and exploiting the relations in (11):

$$\begin{aligned} \alpha_2 + \beta_2 &= D_x^1\left(\mathbb{E}_{\nu_x^{\sigma,1,Id}}[Y]\right)_{\beta x}[e_1, e_1] = \mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[(\langle e_1, Y\rangle - A(\beta))^2] = A'(\beta), \\ (d-1)\beta_2 &= \sum_{i>1}^{d} D_x^1\left(\mathbb{E}_{\nu_x^{\sigma,1,Id}}[Y]\right)_{\beta x}[e_i, e_i] = \sum_{i>1}^{d} \mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[(\langle e_i, Y\rangle)^2] \\ &= 1 - \mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[(\langle e_1, Y\rangle)^2] = 1 - \mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[(\langle e_1, Y\rangle - A)^2] - A^2 \\ &= 1 - A' - A^2. \end{aligned} \tag{13}$$

And the same can be done for α_3 and β_3:

$$\begin{aligned} \alpha_3 + 3\beta_3 &= D_x^2\left(\mathbb{E}_{\nu_x^{\sigma,1,Id}}[Y]\right)_{\beta x}[e_1, e_1, e_1] = \mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[(\langle e_1, Y\rangle - A(\beta))^3] = A''(\beta), \\ (d-1)\beta_3 &= \sum_{i>1}^{d} D_x^2\left(\mathbb{E}_{\nu_x^{\sigma,1,Id}}[Y]\right)_{\beta x}[e_1, e_i, e_i] = \sum_{i>1}^{d} \mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[(\langle e_i, Y\rangle)^2(\langle e_1, Y\rangle - A)] \\ &= -\mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[(\langle e_1, Y\rangle)^2(\langle e_1, Y\rangle - A)] \\ &= -\mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[(\langle e_1, Y\rangle - A)^3] - 2A\mathbb{E}_{\nu_x^{\sigma,\beta,Id}}[(\langle e_1, Y\rangle - A)^2] \\ &= -A'' - 2AA'. \end{aligned}$$

To conclude it is sufficient to show that $1 - A^2 = O(\frac{1}{\beta})$ and $A', A'' = O(\frac{1}{\beta^2})$. Now, using the identity:

$$Z_\beta = \int_{S^{d-1}} e^{\beta\langle x,y\rangle} d\sigma(y) = C_d \beta^{1-d/2} I_{d/2-1}(\beta),$$

we can explicitly compute $A(\beta)$ as:

$$\begin{aligned} A(\beta) = \frac{\partial_\beta Z_\beta}{Z_\beta} &= \frac{(1-\frac{d}{2})\beta^{-d/2} I_{d/2-1}(\beta) + \beta^{1-d/2}(I_{d/2}(\beta) + (\frac{d}{2}-1)\frac{1}{\beta} I_{d/2-1}(\beta))}{\beta^{1-d/2} I_{d/2-1}(\beta)} \\ &= \frac{I_{d/2}(\beta)}{I_{d/2-1}(\beta)} \approx 1 - \frac{d-1}{2\beta}, \end{aligned} \tag{14}$$

where we used the derivatives rules for the modified Bessel function:

$$\begin{aligned} I'_\nu(z) &= I_{\nu-1}(z) - \frac{\nu}{z} I_\nu(z), \\ I'_\nu(z) &= I_{\nu+1}(z) + \frac{\nu}{z} I_\nu(z), \end{aligned}$$

and its asymptotic behavior (in both cases see [1]).

In a similar way we can also compute:

$$\begin{aligned}
A'(\beta) &= \frac{I_{d/2}(\beta)'}{I_{d/2-1}(\beta)} - \frac{I_{d/2}(\beta)}{\left(I_{d/2-1}(\beta)\right)^2} I_{d/2-1}(\beta)' \\
&= \frac{I_{d/2-1}(\beta) - \frac{d/2}{\beta} I_{d/2}(\beta)}{I_{d/2-1}(\beta)} - \frac{I_{d/2}(\beta)}{I_{d/2-1}(\beta)^2}\left(I_{d/2}(\beta) + \frac{d/2-1}{\beta} I_{d/2-1}(\beta)\right) \\
&= 1 - \frac{d/2}{\beta} A(\beta) - A(\beta)^2 - \frac{d/2-1}{\beta} A(\beta) \\
&= 1 - A^2(\beta) - \frac{d-1}{\beta} A(\beta) \approx \frac{(d-1)^2}{4\beta^2},
\end{aligned} \tag{15}$$

and

$$A''(\beta) = -2A(\beta)A'(\beta) - \frac{d-1}{\beta} A'(\beta) \approx O(\frac{1}{\beta^2}).$$

where we used the asymptotics in (14). This is sufficient to conclude the proof. □

Lemma A.7. *Given a strictly positive probability measure μ the following holds:*

$$|\nu_x^{\mu,\beta,B}(y) - \nu_x^{\sigma,\beta,B}(y)| \le \left(\frac{\|\nabla\mu\|_\infty}{\min|\mu|}(|y - x_B| + C\beta^{-1/2})\right)\nu_x^{\sigma,\beta,B},$$

where $x_B = \frac{B^T x}{|B^T x|}$

Proof. Indeed:

$$\begin{aligned}
\nu_x^{\mu,\beta,B}(y) &= \frac{e^{\beta\langle x,By\rangle}\mu(y)}{\int_{S^{d-1}} e^{\beta\langle x,By\rangle}\mu(y)} \\
&= \frac{e^{\beta\langle x,By\rangle}\mu(x_B) + e^{\beta\langle x,By\rangle}(\mu(y) - \mu(x_B))}{\mu(x_B)\int_{S^{d-1}} e^{\beta\langle x,By\rangle} d\sigma(y) + \int_{S^{d-1}} e^{\beta\langle x,By\rangle}(\mu(y) - \mu(x_B)) d\sigma(y)} \\
&= \nu_x^{\sigma,\beta,B}(y)\left(\frac{1 + \frac{\mu(y)-\mu(x_B)}{\mu(x_B)}}{1 + \frac{1}{\mu(x_B)}\int(\mu(y) - \mu(x_B)) d\nu_x^{\sigma,\beta,B}}\right) \\
&= \nu_x^{\sigma,\beta,B}(y)\left(1 + \frac{\frac{\mu(y)-\mu(x_B)}{\mu(x_B)} - \frac{1}{\mu(x_B)}\int(\mu(y) - \mu(x_B)) d\nu_x^{\sigma,\beta,B}(y)}{1 + \frac{1}{\mu(x_B)}\int(\mu(y) - \mu(x_B)) d\nu_x^{\sigma,\beta,B}(y)}\right) \\
&= \nu_x^{\sigma,\beta,B}(y)\left(1 + (R_1 + R_2)(1 + R_2)\right) \\
&\le \nu_x^{\sigma,\beta,B}(y)\left(1 + \frac{\|\nabla\mu\|_\infty}{\min|\mu|}(|y - x_B| + C\beta^{-1/2})\right),
\end{aligned}$$

where we used:

$$\begin{aligned}
|R_1| \le &\frac{|\mu(y) - \mu(x_B)|}{\mu(x_B)} \le \frac{\|\nabla\mu\|_\infty}{\min|\mu|}|y - x|, \\
|R_2| \le &\frac{1}{\mu(x_B)}\int(\mu(y) - \mu(x_B)) d\nu_x^{\sigma,\beta,B} \le \frac{\|\nabla\mu\|_\infty}{\min|\mu|}\int |y - x_B| d\nu_x^{\sigma,\beta,B} \le C\beta^{-1/2}\frac{\|\nabla\mu\|_\infty}{\min|\mu|},
\end{aligned}$$

and the last inequality is a consequence of $\nu_x^{\sigma,\beta,B} = \nu_{x_B}^{\sigma,\beta|B^T x|,Id}$ and lemma D.1. □

Proposition A.8. *The derivatives of the vector field $\chi^\beta[\mu]$ satisfy:*

$$\begin{aligned}
\chi[\mu] &\le C, \\
D_x^1\chi[\mu] &\le C\left(1 + \frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2}\right), \\
D_x^2\chi[\mu] &\le C\left(1 + \frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2} + \frac{\|\nabla\mu\|_\infty}{\min|\mu|}\right).
\end{aligned}$$

Proof. Thanks to Lemma A.5 we just need to bound the cumulants. The first one is already done.

Second cumulant:

$$\begin{aligned}\left|\mathbb{E}_{\nu_x^{\mu,\beta,B}}[\Big(Y-\mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y]\Big)^{\otimes 2}]\right| &= \left|\int\int (y-z_1)\otimes(y-z_2)\nu_x^{\mu,\beta,B}(dy)\nu_x^{\mu,\beta,B}(dz_1)\nu_x^{\mu,\beta,B}(dz_2)\right| \\ &= \left|\mathbb{E}_{\nu_x^{\sigma,\beta,B}}[\Big(Y-\mathbb{E}_{\nu_x^{\sigma,\beta,B}}[Y]\Big)^{\otimes 2}]\right| + R.\end{aligned}$$

We have already shown in lemma A.6 that the first term is $\leq C\frac{1}{\beta}$. Now we need to bound the second term. R can be expanded by multi-linearity and using lemma A.7 the worst case is either of the form:

$$\begin{aligned}&\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2}\left|\int\int (y-z_1)\otimes(y-z_2)\nu_x^{\sigma,\beta,B}(dy)\nu_x^{\sigma,\beta,B}(dz_1)\nu_x^{\sigma,\beta,B}(dz_2)\right| \\ &\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2}\int \nu_x^{\sigma,\beta,B}(dy)\prod_{i=1}^{2}\int|y-z_i|\nu_x^{\sigma,\beta,B}(dz_i) \\ &= C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2}\int \nu_x^{\sigma,\beta,B}(dy)\left(\int|y-z_1|\nu_x^{\sigma,\beta,B}(dz_1)\right)^2 \\ &\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2}\int \nu_x^{\sigma,\beta,B}(dy)\int|y-z_1|^2\nu_x^{\sigma,\beta,B}(dz_1) \\ &\leq 4C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2}\int|y-x_B|^2\nu_x^{\sigma,\beta,B}(dy) \leq 4C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2}\beta^{-1} \leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\frac{1}{\beta^{3/2}},\end{aligned}$$

where in the last line we used lemma D.1, or of the form:

$$\begin{aligned}&\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\left|\int\int (y-z_1)\otimes(y-z_2)|y-z_B|\nu_x^{\sigma,\beta,B}(dy)\nu_x^{\sigma,\beta,B}(dz_1)\nu_x^{\sigma,\beta,B}(dz_2)\right| \\ &\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\int\int |y-z_1||y-z_2||y-z_B|\nu_x^{\sigma,\beta,B}(dy)\nu_x^{\sigma,\beta,B}(dz_1)\nu_x^{\sigma,\beta,B}(dz_2) \\ &\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\left(\int|y-z_B|^2\nu_x^{\sigma,\beta,B}(dy)\right)^{1/2}\left(\int\int|y-z|^4\nu_x^{\sigma,\beta,B}(dy)\nu_x^{\sigma,\beta,B}(dz)\right)^{1/2} \\ &\leq 4C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\left(\int|y-z_B|^2\nu_x^{\sigma,\beta,B}(dy)\right)^{1/2}\left(\int|y-z_B|^4\nu_x^{\sigma,\beta,B}(dy)+\int|z-z_B|^4\nu_x^{\sigma,\beta,B}(dz)\right)^{1/2} \\ &\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2}\beta^{-1} \leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\frac{1}{\beta^{3/2}},\end{aligned}$$

or of the form:

$$\begin{aligned}&\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\left|\int\int (y-z_1)\otimes(y-z_2)|z_1-z_B|\nu_x^{\sigma,\beta,B}(dy)\nu_x^{\sigma,\beta,B}(dz_1)\nu_x^{\sigma,\beta,B}(dz_2)\right| \\ &\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\int\int |y-z_1||y-z_2||z_1-z_B|\nu_x^{\sigma,\beta,B}(dy)\nu_x^{\sigma,\beta,B}(dz_1)\nu_x^{\sigma,\beta,B}(dz_2) \\ &\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\left(\int|z_1-z_B|^2\nu_x^{\sigma,\beta,B}(dz_1)\right)^{1/2}\left(\int\int|y-z_1|^2|y-z_2|^2\nu_x^{\sigma,\beta,B}(dy)\nu_x^{\sigma,\beta,B}(dz_1)\nu_x^{\sigma,\beta,B}(dz_2)\right)^{1/2} \\ &\leq 4C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\left(\int|y-z_B|^2\nu_x^{\sigma,\beta,B}(dy)\right)^{1/2}\left(\int|y-z_B|^4\nu_x^{\sigma,\beta,B}(dy)+\int|z-z_B|^4\nu_x^{\sigma,\beta,B}(dz)\right)^{1/2} \\ &\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2}\beta^{-1} \leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\frac{1}{\beta^{3/2}}.\end{aligned}$$

These are the worst cases because every $|y-z|$ produces an additional $\beta^{-1/2}$ by lemma D.1. Hence we proved, thanks to lemma A.5, that:

$$|D_x^1\chi[\mu]| \leq C\left(1+\frac{\|\nabla\mu_\beta\|_\infty}{\min|\mu|}\beta^{-1/2}\right).$$

The bound for the third cumulant is similar to what we have done above:

$$\begin{aligned}
&|\mathbb{E}_{\nu_x^{\mu,\beta,B}}[\Big(Y-\mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y]\Big)^{\otimes 3}]| \\
&\qquad =\left|\int\int\int\int (y-z_1)\otimes(y-z_2)\otimes(y-z_3)\nu_x^{\mu,\beta,B}(dy)\nu_x^{\mu,\beta,B}(dz_1)\nu_x^{\mu,\beta,B}(dz_2)\nu_x^{\mu,\beta,B}(dz_3)\right| \\
&\qquad =|\mathbb{E}_{\nu_x^{\sigma,\beta,B}}[\Big(Y-\mathbb{E}_{\nu_x^{\sigma,\beta,B}}[Y]\Big)^{\otimes 3}]|+R.
\end{aligned}$$

We have already shown that the first term is $O(\frac{1}{\beta^2})$. Now we need to bound the second term. R can be expanded again as in lemma A.7 and the worst case is either of the form:

$$\begin{aligned}
&\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2}\left|\int\int\int\int (y-z_1)\otimes(y-z_2)\otimes(y-z_3)\nu_x^{\sigma,\beta,B}(dy)\nu_x^{\sigma,\beta,B}(dz_1)\nu_x^{\sigma,\beta,B}(dz_2)\nu_x^{\sigma,\beta,B}(dz_3)\right| \\
&\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2}\int\nu_x^\sigma(dy)\prod_{i=1}^3\int|y-z_i|\nu_x^{\sigma,\beta,B}(dz_i) \\
&= C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2}\int\nu_x^\sigma(dy)\left(\int|y-z_1|\nu_x^{\sigma,\beta,B}(dz_1)\right)^3 \\
&\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2}\int\nu_x^\sigma(dy)\int|y-z_1|^3\nu_x^{\sigma,\beta,B}(dz_1) \\
&\leq 2C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2}\int|y-x_B|^3\nu_x^{\sigma,\beta,B}(dy)\leq C\beta^{-1/2}\beta^{-3/2}\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\frac{1}{\beta^2},
\end{aligned}$$

or of the form:

$$\begin{aligned}
&\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\left|\int\int\int\int (y-z_1)\otimes(y-z_2)\otimes(y-z_3)|y-z_B|\nu_x^{\sigma,\beta,B}(dy)\nu_x^{\sigma,\beta,B}(dz_1)\nu_x^{\sigma,\beta,B}(dz_2)\nu_x^{\sigma,\beta,B}(dz_3)\right| \\
&\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\int\int\int\int|y-z_1||y-z_2||y-z_3||y-z_B|\nu_x^{\sigma,\beta,B}(dy)\nu_x^{\sigma,\beta,B}(dz_1)\nu_x^{\sigma,\beta,B}(dz_2)\nu_x^{\sigma,\beta,B}(dz_3) \\
&\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\prod_{i=1}^3\left(\int\int|y-z_i|^4\nu_x^{\sigma,\beta,B}(dz_i)\nu_x^{\sigma,\beta,B}(dy)\right)^{1/4}\left(\int|y-z_B|^4\nu_x^{\sigma,\beta,B}(dy)\right)^{1/4} \\
&= C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\left(\int\int|y-z|^4\nu_x^{\sigma,\beta,B}(dz)\nu_x^{\sigma,\beta,B}(dy)\right)^{3/4}\left(\int|y-z_B|^4\nu_x^{\sigma,\beta,B}(y)\right)^{1/4} \\
&\leq C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}(\beta^{-2})^{3/4}(\beta^{-2})^{1/4}=C\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-2}.
\end{aligned}$$

Thus, we can replace the bounds on the second and third cumulants that we obtained above in the estimates of Lemma A.5 to conclude that:

$$\begin{aligned}
|D_x^2\chi[\mu]|&\leq C\left(1+\beta\,|\mathbb{E}_{\nu_x^{\mu,\beta,B}}[\Big(Y-\mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y]\Big)^{\otimes 2}]|+\beta^2\,|\mathbb{E}_{\nu_x^{\mu,\beta,B}}[\Big(Y-\mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y]\Big)^{\otimes 3}]|\right) \\
&\leq C\left(1+\frac{\|\nabla\mu_\beta\|_\infty}{\min|\mu|}\beta^{-1/2}+\frac{\|\nabla\mu_\beta\|_\infty}{\min|\mu|}\right).
\end{aligned}$$

□

Lemma A.9. *If μ solves the PDE:*

$$\begin{cases}\partial_t\mu &= -div(\mu\chi[\mu]),\\ \mu(0) &= \mu_0.\end{cases}$$

then:

- $\partial_t\|\mu\|_\infty\leq\|\mu\|_\infty|D_x^1\chi[\mu]|$,

- $\partial_t \left(\min \mu^\beta\right) \geq -\left(\min \mu^\beta\right) |D_x^1 \chi[\mu]|$,
- $\partial_t \|\nabla \mu\|_\infty \leq \|\nabla \mu\|_\infty |D_x^1 \chi[\mu]| + \frac{1}{2}\|\mu\|_\infty |D_x^2 \chi[\mu]|$.

Proof. Let x_t be a point of maximum for $|\mu_t|$. Then $\nabla_{S^{d-1}}\mu_t(x_t) = 0$ and:

$$\partial_t \mu_t(x_t) = -div(\mu\chi[\mu])(x_t) + \langle \nabla \mu_t(x_t), x_t' \rangle = -\mu_t(x_t) div(\chi[\mu])(x_t).$$

And for $\min \mu$ we can use the same argument.

Now, let x_t be a point of maximum for $|\nabla_{S^{d-1}}\mu|^2$, then $H_{S^{d-1}}\mu(x_t)\nabla_{S^{d-1}}\mu(x_t) = 0$, hence:

$$\begin{aligned}\partial_t |\nabla \mu(x_t)|^2 =& - \langle \nabla \mu(x_t), \nabla div(\mu\chi[\mu])(x_t) \rangle + \langle \nabla \mu(x_t), H\mu(x_t) x_t' \rangle \\ =& - \langle \nabla \mu(x_t), \nabla(\nabla \mu \cdot \chi[\mu])(x_t) \rangle - \langle \nabla \mu(x_t), \nabla(\mu D_x^1 \chi[\mu]))(x_t) \rangle \\ =& - \langle \nabla \mu(x_t), H\mu(x_t)\chi[\mu](x_t) \rangle - \langle \nabla \mu(x_t), D_x^1\chi[\mu](x_t)\nabla\mu(x_t) \rangle \\ & - \langle \nabla \mu(x_t), D_x^1\chi[\mu](x_t)\nabla\mu(x_t) \rangle - \langle \nabla \mu(x_t), D_x^2\chi[\mu](x_t)\mu(x_t) \rangle \\ \leq& 2|\nabla\mu(x_t)|^2 |D_x^1\chi[\mu]| + |\nabla\mu(x_t)||\mu(x_t)||D_x^2\chi[\mu]|.\end{aligned}$$

Using that $\partial_t |\nabla\mu(x_t)|^2 = 2|\nabla\mu(x_t)|\partial_t|\nabla\mu(x_t)|$ and dividing by $|\nabla\mu(x_t)|$ we get the thesis. □

Lemma A.10. *Consider again μ_t solution of the PDE:*

$$\begin{cases}\partial_t \mu &= -div(\mu\chi_\beta[\mu]), \\ \mu(0) &= \mu_0.\end{cases}$$

Define:

$$\begin{aligned} C_1 &= 2C\|\mu_0\|_\infty e^{2CT}, \\ C_2 &= 2C\left(1 + \frac{\|\mu_0\|_\infty}{\min \mu_0} e^{4CT}\right), \end{aligned}$$

Then, for β large enough (depending just on μ_0 and C):

- $\|\mu_t\|_\infty \leq 2(\|\mu_0\|_\infty)e^{2Ct}$,
- $\min \mu_t \geq \frac{1}{2}(\min \mu_0)e^{-2Ct}$,
- $|\nabla \mu_t|_\infty \leq 2\left(\frac{C_1}{C_2} + |\nabla\mu_0|_\infty\right)e^{C_2 t}$.

Proof. The thesis is true at time $t = 0$. Let us assume that it is true on $[0, t]$. Then $\exists \beta$ big enough (where "big" depends only on C_1, C_2, i.e. just μ_0, B, V, d) such that $\frac{\|\nabla\mu\|_\infty}{\min|\mu|}\beta^{-1/2} \leq 1$ on $[0, t]$. Hence $D\chi[\mu_\beta] \leq 2C$ on $[0, t]$ thanks to proposition A.8. By Gronwall applied to the first two bounds in lemma A.9 we can conclude:

$$\begin{aligned} \|\mu_t\|_\infty &\leq (\|\mu_0\|_\infty)e^{2Ct}, \\ \min \mu_t &\geq (\min \mu_0)e^{-2Ct}, \end{aligned}$$

For $\|\nabla\mu\|_\infty$ we have, again by lemma A.9:

$$\begin{aligned}\partial_t \|\nabla\mu_t\|_\infty &\leq \|\nabla\mu_t\|_\infty |D_x^1\chi[\mu_t]| + \frac{1}{2}\|\mu\|_\infty |D_x^2\chi[\mu_t]| \\ &\leq 2C\|\nabla\mu_t\|_\infty + \left(\|\mu_0\|_\infty e^{2CT}\right) C\left(1 + 1 + \frac{\|\nabla\mu_t\|_\infty}{\min|\mu|}\right) \\ &\leq 2C\|\mu_0\|_\infty e^{2CT} + C\left(2 + 2\frac{\|\mu_0\|_\infty}{\min\mu_0}e^{4CT}\right)\|\nabla\mu_t\|_\infty \\ &= C_1 + C_2\|\nabla\mu_t\|_\infty.\end{aligned}$$

where in the second row we used the assumption on $[0, t]$ and proposition A.8. Hence by Gronwall:

$$\|\nabla \mu_t\|_\infty \leq \left(\frac{C_1}{C_2} + \|\nabla \mu_0\|_\infty \right) e^{C_2 t}.$$

This concludes the continuation argument and the proof. □

Corollary A.11. *For β large enough (depending on μ_0, B, V, d) the vector fields $\{\chi_\beta[\mu]\}_{\beta,t}$ are jointly Lipschitz in β and $t \in [0, T]$.*

Proof. This is a consequence of lemma A.10 and proposition A.8. □

Corollary A.12. *For every $x \in \mathbb{S}^{d-1}$ and $t \in [0, T]$:*

$$\chi_\beta[\mu_t^\beta](x) \to \frac{P_x V B^T x}{|B^T x|} \quad \textit{as} \quad \beta \to \infty$$

Proof. With the usual notations one can write:

$$\chi_\beta[\mu^\beta](x) = P_x V \left(\mathbb{E}_{\nu_x^{\mu,\beta,B}}[Y] \right) = P_x V \left(\mathbb{E}_{\nu_x^{\sigma,\beta,B}}[Y] \right) + R.$$

The reminder R is bounded using lemma A.7 by:

$$\begin{aligned} |R| &\leq C \left(\beta^{-1/2} + \frac{\|\nabla \mu_t^\beta\|_\infty}{\min \mu_t^\beta} \int |y - x_B| \nu_x^{\sigma,\beta,B} \right) \\ &\leq C \left(\beta^{-1/2} + \frac{\|\nabla \mu_t^\beta\|_\infty}{\min \mu_t^\beta} \beta^{-1/2} \right) = O(\beta^{-1/2}), \end{aligned}$$

where the last line follows from lemma D.1 and lemma A.10. Hence, the proof can be concluded by noticing that:

$$P_x V \left(\mathbb{E}_{\nu_x^{\sigma,\beta,B}}[Y] \right) = P_x V \left(\mathbb{E}_{\nu_{x_B}^{\sigma,\beta|B^T x|,Id}}[Y] \right) = P_x V (A(\beta) x_B) = (1 + O(\beta^{-1})) P_x \frac{V B^T x}{|B^T x|},$$

where we used the identities in equations 10, 12, 14. □

We can finally pass to the limit in the PDE. Indeed consider a subsequence μ_β of solutions to the PDE converging in $\mathcal{C}([0, T], \mathcal{P}(\mathbb{S}^{d-1}))$ to a certain probability measure μ^∞. If we define the vector field $\chi_\infty(x) := P_x V \frac{B^T x}{|B^T x|}$, then for every $f \in C_b^2(\mathbb{S}^{d-1})$:

$$\langle f, \mu_t^\infty \rangle - \langle f, \mu_0^\infty \rangle - \int_0^t \langle \nabla f, \chi_\infty \mu_s^\infty \rangle \, ds| \leq \langle f, \mu_t^\infty - \mu_t^\beta \rangle| + \int_0^t |\langle \nabla f, \chi_\infty \mu_s^\infty - \chi_\beta[\mu_s^\beta] \mu_s^\beta \rangle| \, ds,$$

where we used that $\mu_0^\infty = \mu_0 = \mu_0^\beta$ and that the PDE in weak form for μ^β is:

$$\langle f, \mu_t^\beta \rangle - \langle f, \mu_0^\beta \rangle - \int_0^t \langle \nabla f, \chi^\beta[\mu_s^\beta] \mu_s^\beta \rangle \, ds = 0.$$

Moreover, as $\beta \to \infty$:

$$|\langle f, \mu_t^\infty - \mu_t^\beta \rangle| \to 0$$

thanks to the fact that f is Lipschitz and by the definition of convergence in $\mathcal{C}([0, T], \mathcal{P}(\mathbb{S}^{d-1}))$. For the second term:

$$\begin{aligned} \int_0^t |\langle \nabla f, \chi_\infty \mu_s^\infty - \chi_\beta[\mu_s^\beta] \mu_s^\beta \rangle| \, ds &\leq \int_0^t |\langle \nabla f, (\chi_\infty - \chi_\beta[\mu_s^\beta]) \mu_s^\infty \rangle| \, ds \\ &+ \int_0^t |\langle \nabla f, \chi_\beta[\mu_s^\beta] (\mu_s^\infty - \mu_s^\beta) \rangle| \, ds. \end{aligned}$$

The first part goes to 0 by dominated convergence ($\nabla f, \chi_\infty, \chi^\beta[\mu_\beta]$ are bounded, and $\chi_\beta[\mu_\beta] \to \chi_\infty$ point-wise by lemma A.12). The second part goes to 0 by definition of the convergence $\mu_\infty \to \mu_\beta$ and by equi-lipschitzianity of $\chi_\beta[\mu_\beta]$ (see corollary A.11).

The uniqueness is standard, since μ is a probability measure and the vector field is smooth (see, for example, [5]).

A.2 Asymptotic behavior

This section studies the asymptotic behavior of the support of the solution to the partial differential equation:

$$\partial_t \mu = -div\left(\mu \frac{P_x V B^T x}{|B^T x|}\right) \tag{16}$$

and in particular, we prove Proposition 3.4.

Lemma A.13. *The ODE:*

$$\frac{d}{dt} x(t) = \frac{P_x V B^T x(t)}{|B^T x(t)|} \tag{17}$$

is a time-reparameterization of:

$$\frac{d}{dt} y(t) = P_y V B^T y(t)$$

where $y(t) = x(f^{-1}(t))$ and $f(t) = \int_0^t \frac{1}{|B^T x(s)|} ds$.

Remark A.14. *The reparameterization is well defined since B is invertible.*

Proof. : We have:

$$\begin{aligned}\frac{d}{dt} y(t) &= x'(f^{-1}(t)) \cdot \frac{d}{dt} f^{-1}(t) \\ &= \frac{P_x(V B^T x)}{|B^T x|}(f^{-1}(t)) \cdot |B^T x(f^{-1}(t))| = P_y(V B^T y)\end{aligned}$$

This shows that the ODE is a time-reparameterization of the ODE for $y(t)$. □

Lemma A.15. *If $z(t)$ solves:*

$$\frac{d}{dt} z(t) = V B^T z(t),$$

then $y(t) = \frac{z(t)}{|z(t)|}$.

Proof. We have:

$$\begin{aligned}\frac{d}{dt} y(t) = \frac{d}{dt}\left(\frac{z(t)}{|z(t)|}\right) &= \frac{z'}{|z|} - \frac{1}{|z|^2}\frac{1}{2|z|} 2\langle z, z'\rangle z \\ &= \frac{V B^T z}{|z|} - \frac{\langle z, V B^T z\rangle}{|z|^3} z = V B^T y - \langle y, V B^T y\rangle y \\ &= P_y V B^T y.\end{aligned}$$

This concludes the proof. □

Corollary A.16. *For Lebesgue almost every x_0, the ω-limit set $\omega(x_0) \subset E_{max}$.*

Proof. This is a consequence of lemma A.13, lemma A.15 and of the classical theory for linear ODEs, after reducing to the Jordan canonical form of the matrix $V B^T$. □

Remark A.17. *This technical result parallels Lemma 3.1 in [29], which was used to analyze the dynamics of the first token in causal attention.*

And now we can finally prove Proposition 3.4:

Proof. Denote Φ_t the flow of the ODE (17) and let $\phi \in C_b^2(\mathbb{S}^{d-1})$ be a test function with $supp(\phi) \subset E_{max}^C \cap \mathbb{S}^{d-1}$. Fix $\mu_\infty \in \omega(\mu_0)$. Then there exists a divergent sequence of times $\{t_k\}_k$ such that $\mu_{t_k} \to \mu_\infty$ weakly. As a consequence:

$$\begin{aligned}\int_{\mathbb{S}^{d-1}} \phi(x)\mu_\infty(dx) &= \lim_{k\to\infty}\int_{\mathbb{S}^{d-1}} \phi(x)\mu_{t_k}(dx) = \lim_{k\to\infty}\int_{\mathbb{S}^{d-1}} \phi(x)\Phi_{t_k\#}\mu_0(dx)\\ &= \lim_{k\to\infty}\int_{\mathbb{S}^{d-1}} \phi(\Phi_{t_k}(x))\mu_0(dx) = 0,\end{aligned}$$

where we used corollary A.16 and the dominated convergence theorem. □

B Proofs of the heat phase

In this section, we prove Proposition 3.9, which characterizes the second phase using the heat equation on the sphere.

Lemma B.1. *Given a measure $\mu \in C^2(\mathbb{S}^{d-1}) \cap \mathcal{P}(\mathbb{S}^{d-1})$ strictly positive, then the following holds:*

$$\beta\chi_\beta[\mu](x) = \frac{\nabla_x\mu(x)}{\mu(x)} + \frac{\|\nabla\mu\|_\infty}{\min\mu}\left(1 + \frac{\|H_x\mu\|_\infty}{\min\mu}\right) O_{L^\infty(\mathbb{S}^{d-1})}(\beta^{-1/2}),$$

with the gradient ∇_x and Hessian H_x defined with respect to the standard Riemannian metric on $\mathbb{S}^{d-1}$.

Proof.

$$\begin{aligned}\beta\chi_\beta[\mu](x) &= \beta\frac{\int e^{\beta\langle x,y\rangle}P_x y\ \mu(dy)}{\int e^{\beta\langle x,y\rangle}\mu(dy)}\\ &= \beta\frac{\int e^{\beta\langle x,y\rangle}P_x y\ \mu(dy)}{\mu(x)\int e^{\beta\langle x,y\rangle}\sigma(dy)}\frac{1}{1+R_1}\\ &= \beta\left(\frac{\int e^{\beta\langle x,y\rangle}P_x y\ (\mu(x) + \langle y-x, \nabla\mu(x)\rangle)\ \sigma(dy)}{\mu(x)\int e^{\beta\langle x,y\rangle}\sigma(dy)} + R_2\right)\frac{1}{1+R_1},\end{aligned}$$

where, by lemma D.1:

$$\begin{aligned}|R_1| :&= \left|\frac{\int e^{\beta\langle x,y\rangle}\mu(dy)}{\mu(x)\int e^{\beta\langle x,y\rangle}\sigma(dy)} - 1\right| \le \frac{||\nabla\mu||_\infty}{\min\mu}\int |y-x|\nu_x^{\sigma,\beta,I}(dy) = \frac{||\nabla\mu||_\infty}{\min\mu}O(\beta^{-1/2}),\\ |R_2| :&= \left|\frac{\int e^{\beta\langle x,y\rangle}P_x y\mu(dy) - \int e^{\beta\langle x,y\rangle}P_x y\ (\mu(x) + \langle y-x,\nabla\mu(x)\rangle)\ \sigma(dy)}{\mu(x)\int e^{\beta\langle x,y\rangle}\sigma(dy)}\right|\\ &\le \frac{||H\mu||_\infty}{\min\mu}\int |y-x|^3\nu_x^{\sigma,\beta,I}(dy) = \frac{||H\mu||_\infty}{\min\mu}O(\beta^{-3/2}).\end{aligned}\tag{18}$$

Hence:

$$\beta\chi_\beta[\mu](x) = \beta\left(P_x\int y\nu_x^{\sigma,\beta,I}(dy) + P_x\int (y-x)^{\otimes 2}\nu_x^{\sigma,\beta,I}(dy)\frac{\nabla_x\mu(x)}{\mu(x)} + R_2\right)\frac{1}{1+R_1},$$

and noticing that $\mathbb{E}_{\nu_x^{\sigma,\beta,I}}[Y]$ is parallel to x (see proof of lemma A.6):

$$\begin{aligned}\beta\chi_\beta[\mu](x) &= \beta\left(P_x\int (y - \mathbb{E}_{\nu_x^{\sigma,\beta,I}}[Y])^{\otimes 2}\nu_x^{\sigma,\beta,I}(dy)\frac{\nabla_x\mu(x)}{\mu(x)} + R_2\right)\frac{1}{1+R_1}\\ &= \beta\left(P_x\mathbb{E}_{\nu_x^{\sigma,\beta,I}}[(Y - \mathbb{E}_{\nu_x^{\sigma,\beta,I}}[Y])^{\otimes 2}]\frac{\nabla_x\mu(x)}{\mu(x)} + R_2\right)\frac{1}{1+R_1}\\ &= \beta\left(P_x(\alpha_2 x\otimes x + \beta_2 I)\frac{\nabla_x\mu(x)}{\mu(x)} + R_2\right)\frac{1}{1+R_1}\\ &= \beta\left(\frac{1 - A'(\beta) - A(\beta)^2}{d-1}\frac{\nabla_x\mu(x)}{\mu(x)} + R_2\right)\frac{1}{1+R_1},\end{aligned}$$

where α_2, β_2 are defined in the proof of Lemma A.6) and we used equation 13. To conclude, it suffices to replace equations 18 and the asymptotic estimates 14 and 15:

$$
\begin{aligned}
&= \beta \left(\frac{1}{\beta} \frac{\nabla_x \mu(x)}{\mu(x)} + \frac{\|\nabla \mu\|_\infty}{\min \mu} O(\beta^{-3/2}) \right) \frac{1}{1 + \frac{\|H_x \mu\|_\infty}{\min \mu} O(\beta^{-1/2})} \\
&= \left(\frac{\nabla_x \mu(x)}{\mu(x)} + \frac{\|\nabla \mu\|_\infty}{\min \mu} O(\beta^{-1/2}) \right) \left(1 + \frac{\|H_x \mu\|_\infty}{\min \mu} O(\beta^{-1/2}) \right) \\
&= \frac{\nabla_x \mu(x)}{\mu(x)} + \frac{\|\nabla \mu\|_\infty}{\min \mu} \left(1 + \frac{\|H_x \mu\|_\infty}{\min \mu} \right) O(\beta^{-1/2}).
\end{aligned}
$$

□

Corollary B.2. *Given a family $\{\mu^\beta\}_\beta$ of probability measures on $\mathbb{S}^{d-1}$, suppose that there exist $c, C > 0$ such that $\|\mu^\beta\|_{C^2(\mathbb{S}^{d-1})} \leq C$ and $\mu^\beta \geq c$ for every $\beta \geq 0$. Moreover, assume there exists μ^∞ such that $\mu^\beta \to \mu^\infty$ in $C^1(\mathbb{S}^{d-1})$. Then*

$$
\beta \chi_\beta[\mu^\beta](x) \to \frac{\nabla_x \mu^\infty(x)}{\mu^\infty(x)} \quad \forall x \in \mathbb{S}^{d-1},
$$

where ∇_x is the gradient with respect to the standard Riemannian metric on $\mathbb{S}^{d-1}$.

Proof of Proposition 3.9. Without loss of generality, set $\gamma = -1$, the other case is analogous. The residual term is given by:

$$
R_\beta = \beta \operatorname{div}(\mu \chi_\beta[\mu]) - \Delta \mu = \operatorname{div} \left(\mu \left[\beta \chi_\beta[\mu] - \frac{\nabla \mu}{\mu} \right] \right).
$$

It is sufficient to show that:

$$
\left\| \beta \chi_\beta[\mu] - \frac{\nabla \mu}{\mu} \right\|_{C^{k+1}(\mathbb{S}^{d-1})} \to 0 \quad \text{as } \beta \to \infty.
$$

Corollary B.2 guarantees convergence in $C^0(\mathbb{S}^{d-1})$ thanks to the assumptions on μ_t. To improve this to higher regularity, we can use an interpolation argument through uniform bounds in C^{k+2}.

Define the kernel $W_\beta(t) := \frac{e^{\beta t}}{K_\beta}$, where $K_\beta := \int_{\mathbb{S}^{d-1}} e^{\beta \langle x, y \rangle} \, d\sigma(y)$. Then,

$$
\chi_\beta[\mu](x) = \frac{\nabla (W_\beta * \mu)(x)}{(W_\beta * \mu)(x)}.
$$

By the product rule, for every $0 \leq j \leq k+2$, there exists a polynomial p_j such that:

$$
\left\| D_x^j \left(\frac{\nabla (W_\beta * \mu)}{W_\beta * \mu} - \frac{\nabla \mu}{\mu} \right) \right\| \leq p_j \left(\|W_\beta * \mu\|_{C^{j+1}}, \|\mu\|_{C^{j+1}}, \min_{x \in \mathbb{S}^{d-1}} \mu \right),
$$

where we used that $\min_{x \in \mathbb{S}^{d-1}} W_\beta * \mu \geq \min_{x \in \mathbb{S}^{d-1}} \mu$. The only thing left is to notice that

$$
\|W_\beta * \mu\|_{C^j} \leq C_k \|\mu\|_{C^j},
$$

though proving this on $\mathbb{S}^{d-1}$ requires some care.

Consider the case $j = 1$ ($j > 1$ follows by induction) and fix $v \in T_x(\mathbb{S}^{d-1})$. Let $A \in \mathfrak{so}(d)$ (a skew-symmetric matrix) satisfying $Ax = v$, and define $R(t) := e^{tA}$. In such a way $R(0)x = x$ and $R'(0)x = v$. Then:

$$
\begin{aligned}
\nabla_x (W_\beta * \mu)[v] &= \frac{d}{dt} (W_\beta * \mu)(R(t)x) \Big|_{t=0} \\
&= \frac{d}{dt} \int_{\mathbb{S}^{d-1}} W_\beta(\langle R(t)x, y \rangle) \mu(y) \, d\sigma(y) \Big|_{t=0} \\
&= \frac{d}{dt} \int_{\mathbb{S}^{d-1}} W_\beta(\langle x, R(t)^T y \rangle) \mu(y) d\sigma(y)|_{t=0} \\
&= \frac{d}{dt} \int_{\mathbb{S}^{d-1}} W_\beta(\langle x, z \rangle) \mu(R(t)z) d\sigma(y)|_{t=0} \\
&= \int_{\mathbb{S}^{d-1}} W_\beta(\langle x, z \rangle) \nabla_z \mu[Az] \, d\sigma(z),
\end{aligned}
$$

where we used the change of variable $z = R(t)^T y$, and the invariance of the measure on the sphere. Since $\|W_\beta\|_{L^1} = 1$, it follows that

$$\|\nabla_x(W_\beta * \mu)\|_{C^1(\mathbb{S}^{d-1})} \leq C\|\mu\|_{C^1(\mathbb{S}^{d-1})}.$$

Higher derivatives follow similarly, completing the proof. □

C Proofs of the pairing phase

In this section we provide the proof of Proposition 3.11. Consider the interacting particle system on $\mathbb{S}^{d-1}$ described by the following ODEs corresponding to the case ($Q^T K = V = Id$):

$$\dot{x}_i(t) = \frac{1}{Z_\beta(x_i)} \sum_{j=1}^N e^{\beta\langle x_i, x_j\rangle} P_{x_i}(x_j).$$

where $Z_\beta(x_i) = \sum_{j=1}^N e^{\beta\langle x_i, x_j\rangle}$ and $P_{x_i}(x_j) = x_j - \langle x_i, x_j\rangle x_i$ is the projection on the hyperplane orthogonal to x_i. Suppose that there exists a unique pair $(\underline{i}, \underline{j})$ such that at initialization $\langle x_{\underline{i}}, x_{\underline{j}}\rangle = \max_{i\neq j}\langle x_i, x_j\rangle$ and denote $\langle x_{\underline{i}}(t), x_{\underline{j}}(t)\rangle := d_t$. Define also $m_t := \max\{\langle x_i, x_j\rangle | i \neq j \text{ and } \{i,j\} \neq \{\underline{i}, \underline{j}\}\}$.

Let $\alpha := \arccos(m_0) - \arccos(d_0) > 0$ and consider the time rescaling given by the inverse of $d\tau = e^{\beta(1-d_t)}dt$, that we will still denote by t. Then:

$$\dot{x}_i(t) = \frac{e^{\beta(1-d_t)}}{Z_\beta(x_i)} \sum_{j=1}^N e^{\beta\langle x_i, x_j\rangle} P_{x_i}(x_j).$$

As usual the constant C can change from line to line, but it does not depend on α or β.

Lemma C.1. *If β is such that $Ce^{-\beta(1-\cos(\alpha/4))}T \leq \frac{1}{2}\alpha$, then:*

$$d_t - m_t \geq 1 - \cos\left(\frac{\alpha}{4}\right) \text{ on } [0, T].$$

Proof. We proceed by a standard continuation argument. At $t = 0$ we have $d_0 - m_0 \geq 1 - \cos(\alpha) \geq 1 - \cos(\alpha/4)$. Suppose the thesis holds on $[0, t]$. Then we have:

- if $i \neq \underline{i}$ and $j \neq \underline{j}$:

$$\begin{aligned}
\partial_t \arccos(\langle x_i, x_j\rangle) =& -\frac{1}{\sqrt{1-\langle x_i(t), x_j(t)\rangle^2}}\partial_t\langle x_i(t), x_j(t)\rangle \\
=& -\frac{1}{\sqrt{1-\langle x_i, x_j\rangle^2}}\frac{e^{\beta(1-d_t)}}{Z_\beta(x_i)}\sum_{k=1}^N e^{\beta\langle x_i, x_k\rangle}\langle P_{x_i}(x_k), x_j\rangle \\
&-\frac{1}{\sqrt{1-\langle x_i, x_j\rangle^2}}\frac{e^{\beta(1-d_t)}}{Z_\beta(x_j)}\sum_{k=1}^N e^{\beta\langle x_j, x_k\rangle}\langle P_{x_j}(x_k), x_i\rangle \\
=& -\frac{e^{\beta(1-d_t)}}{Z_\beta(x_i)}\sum_{k=1}^N e^{\beta\langle x_i, x_k\rangle}\langle P_{x_i}(x_k), \frac{P_{x_i}x_j}{|P_{x_i}x_j|}\rangle \\
&-\frac{e^{\beta(1-d_t)}}{Z_\beta(x_j)}\sum_{k=1}^N e^{\beta\langle x_j, x_k\rangle}\langle P_{x_j}(x_k), \frac{P_{x_j}x_i}{|P_{x_j}x_i|}\rangle \\
\geq& -Ce^{\beta(m_t-d_t)} \geq -Ce^{-\beta(1-\cos(\alpha/4))}.
\end{aligned}$$

- if $i = \underline{i}$ and $j \neq \underline{j}$:

$$\begin{aligned}
\partial_t \arccos(\langle x_{\underline{i}}, x_j\rangle) =& -\frac{1}{\sqrt{1-\langle x_{\underline{i}}, x_j\rangle^2}}\partial_t\langle x_{\underline{i}}(t), x_j(t)\rangle \\
=& -\frac{1}{\sqrt{1-\langle x_{\underline{i}}, x_j\rangle^2}}\frac{e^{\beta(1-d_t)}}{Z_\beta(x_{\underline{i}})}\sum_{k=1}^N e^{\beta\langle x_{\underline{i}}, x_k\rangle}\langle P_{x_{\underline{i}}}(x_k), x_j\rangle \\
&- \frac{1}{\sqrt{1-\langle x_{\underline{i}}, x_j\rangle^2}}\frac{e^{\beta(1-d_t)}}{Z_\beta(x_j)}\sum_{k=1}^N e^{\beta\langle x_j, x_k\rangle}\langle P_{x_j}(x_k), x_{\underline{i}}\rangle \\
=& -\frac{e^{\beta(1-d_t)}}{Z_\beta(x_{\underline{i}})}\sum_{k=1}^N e^{\beta\langle x_{\underline{i}}, x_k\rangle}\langle P_{x_{\underline{i}}}(x_k), \frac{P_{x_{\underline{i}}}x_j}{|P_{x_{\underline{i}}}x_j|}\rangle \\
&- \frac{e^{\beta(1-d_t)}}{Z_\beta(x_j)}\sum_{k=1}^N e^{\beta\langle x_j, x_k\rangle}\langle P_{x_j}(x_k), \frac{P_{x_j}x_{\underline{i}}}{|P_{x_j}x_{\underline{i}}|}\rangle \\
\geq& -\frac{e^{\beta}}{Z_\beta(x_{\underline{i}})}|P_{x_{\underline{i}}}x_{\underline{j}}| \\
&- \frac{e^{\beta(1-d_t)}}{Z_\beta(x_{\underline{i}})}\sum_{k\neq \underline{j}}^N e^{\beta\langle x_{\underline{i}}, x_k\rangle}\langle P_{x_{\underline{i}}}(x_k), \frac{P_{x_{\underline{i}}}x_j}{|P_{x_{\underline{i}}}x_j|}\rangle \\
&- \frac{e^{\beta(1-d_t)}}{Z_\beta(x_j)}\sum_{k=1}^N e^{\beta\langle x_j, x_k\rangle}\langle P_{x_j}(x_k), \frac{P_{x_j}x_{\underline{i}}}{|P_{x_j}x_{\underline{i}}|}\rangle,
\end{aligned}$$

by Cauchy-Schwarz inequality, $\langle P_{x_{\underline{i}}}x_{\underline{j}}, \frac{P_{x_{\underline{i}}}x_j}{|P_{x_{\underline{i}}}x_j|}\rangle \leq |P_{x_{\underline{i}}}x_{\underline{j}}|$, hence:

$$\begin{aligned}
\geq& -\frac{e^{\beta}}{Z_\beta(x_{\underline{i}})}|P_{x_{\underline{i}}}x_{\underline{j}}| - \frac{e^{\beta}}{Z_\beta(x_{\underline{j}})}|P_{x_{\underline{j}}}x_{\underline{i}}| \\
&- \frac{e^{\beta(1-d_t)}}{Z_\beta(x_{\underline{i}})}\sum_{k\neq \underline{j}}^N e^{\beta\langle x_{\underline{i}}, x_k\rangle}\langle P_{x_{\underline{i}}}(x_k), \frac{P_{x_{\underline{i}}}x_j}{|P_{x_{\underline{i}}}x_j|}\rangle \\
&- \frac{e^{\beta(1-d_t)}}{Z_\beta(x_j)}\sum_{k=1}^N e^{\beta\langle x_j, x_k\rangle}\langle P_{x_j}(x_k), \frac{P_{x_j}x_{\underline{i}}}{|P_{x_j}x_{\underline{i}}|}\rangle \\
\geq& \partial_t \arccos(\langle x_{\underline{i}}, x_{\underline{j}}\rangle) - Ce^{\beta(m_t-d_t)}.
\end{aligned}$$

And in the last line we used that:

$$\partial_t \arccos(\langle x_{\underline{i}}, x_{\underline{j}}\rangle) = -\left(\frac{e^{\beta}}{Z_\beta(x_{\underline{i}})}|P_{x_{\underline{i}}}x_{\underline{j}}| + \frac{e^{\beta}}{Z_\beta(x_{\underline{j}})}|P_{x_{\underline{j}}}x_{\underline{i}}|\right) + O(e^{\beta(m_t-d_t)}).$$

In both cases the following holds:

$$\partial_t \arccos(\langle x_i, x_j\rangle) \geq \partial_t \arccos(\langle x_{\underline{i}}, x_{\underline{j}}\rangle) - Ce^{-\beta(1-\cos(\alpha/4))},$$

hence:

$$\arccos(m_t) - \arccos(m_0) \geq \arccos(d_t) - \arccos(d_0) - Ce^{-\beta(1-\cos(\alpha/4))}T,$$

that implies

$$\arccos(m_t) - \arccos(d_t) \geq \arccos(m_0) - \arccos(d_0) - Ce^{-\beta(1-\cos(\alpha/4))}T \geq \frac{\alpha}{2}.$$

We can conclude:

$$d_t - m_t \geq 1 - \cos(\alpha/2) > 1 - \cos(\alpha/4).$$

This is sufficient to close the continuation argument. □

Remark C.2. *We used the fact that, if* $\arccos(x) - \arccos(y) \geq \alpha$*, then* $y - x \geq 1 - \cos(\alpha)$.

Now, recall Proposition 3.11:

Proposition C.3. *The solutions* $x_i(t)$ *of the ODE system* (SA)*, under Assumptions 4 and positive* V*, with the rescaled time* $dt = e^{\beta(1-\langle x_{\underline{i}}, x_{\underline{j}}\rangle)}ds$*, converge as* $\beta \to \infty$ *to the solutions of the system:*

$$\begin{cases} \dot{y}_k(t) = \begin{cases} P_{y_{\underline{i}}}(y_{\underline{j}}) & \textit{if } k = \underline{i}, \\ P_{y_{\underline{j}}}(y_{\underline{i}}) & \textit{if } k = \underline{j}, \\ 0 & \textit{otherwise}, \end{cases} \\ y_i(0) = x_i(0) \end{cases}$$

on finite intervals $[0, T_\epsilon]$*, with* T_ϵ *such that* $\langle y_{\underline{i}}, y_{\underline{j}}\rangle \leq 1 - \epsilon$ *throughout the interval, for any* $\epsilon > 0$.

First we need the following lemma:

Lemma C.4. *If* β *is large enough then* $\delta_t := \langle y_{\underline{i}}, y_{\underline{j}}\rangle \leq 1-c$ *on* $[0,T]$ *implies* $d_t = \langle x_{\underline{i}}, x_{\underline{j}}\rangle \leq 1-c/2$ *on* $[0,T]$.

Proof. We proceed again using a continuation argument. The derivatives of the differences are bounded by:

$$\begin{aligned} \partial_t(\delta(t) - d(t)) =& \partial_t\langle y_{\underline{i}}, y_{\underline{j}}\rangle - \partial_t\langle x_{\underline{i}}, x_{\underline{j}}\rangle \\ &= 2|P_{y_{\underline{i}}} y_{\underline{j}}|^2 - \Big(\frac{e^{\beta(1-d_t)}}{Z_\beta(x_{\underline{i}})}\sum_{k=1}^N e^{\beta\langle x_{\underline{i}}, x_k\rangle}\langle P_{x_{\underline{i}}}(x_k), x_{\underline{j}}\rangle \\ &\qquad + \frac{e^{\beta(1-d_t)}}{Z_\beta(x_{\underline{j}})}\sum_{k=1}^N e^{\beta\langle x_{\underline{j}}, x_k\rangle}\langle P_{x_{\underline{j}}}(x_k), x_{\underline{i}}\rangle\Big) \\ &\leq 2|P_{y_{\underline{i}}} y_{\underline{j}}|^2 - 2\frac{e^\beta}{e^\beta + e^{\beta d_t}}|P_{x_{\underline{i}}}(x_{\underline{j}})|^2 + Ce^{-\beta(1-\cos(\alpha/4))} \\ &\leq 2|P_{y_{\underline{i}}} y_{\underline{j}}|^2 - 2\frac{e^\beta}{e^\beta + e^{\beta d_t}}|P_{x_{\underline{i}}}(x_{\underline{j}})|^2 + Ce^{-\beta(1-\cos(\alpha/4))} \\ &= 2|\delta_t(t) - d_t(t)|^2 + 2|\frac{e^\beta}{e^\beta + e^{\beta d_t}} - 1| + Ce^{-\beta(1-\cos(\alpha/4))} \\ &\leq 2C|\delta_t(t) - d_t(t)| + 2e^{-\beta c/2} + Ce^{-\beta(1-\cos(\alpha/4))}. \end{aligned}$$

where we used $|P_x(y)|^2 = 1 - \langle x, y\rangle^2$ and the previous lemma. The conclusion is again an application of Gronwall's lemma. □

Proof of Proposition 3.11. Thanks to the previous lemma for β large enough, on $[0, T_\epsilon]$ we have $d_t < 1 - c_\epsilon$. Now we can proceed with the proof:

Consider the case $k = \underline{i}$.

$$\begin{aligned} |x_{\underline{i}}(t) - y_{\underline{i}}(t)| = \int_0^t |\dot{x}_{\underline{i}}(s) - \dot{y}_{\underline{i}}(s)|ds &= \int_0^t \left(\frac{e^{\beta(1-d_s)}}{Z_\beta(x_{\underline{i}})}\sum_{k=1}^N e^{\beta\langle x_{\underline{i}}, x_k\rangle}P_{x_{\underline{i}}}(x_k) - P_{y_{\underline{i}}}(y_{\underline{j}})\right)ds \\ &\leq \int_0^t |\frac{e^\beta}{Z_\beta(x_{\underline{i}})}P_{x_{\underline{i}}}(x_{\underline{j}}) - P_{y_{\underline{i}}}(y_{\underline{j}})|ds + Ce^{-\beta(1-\cos(\alpha/4))} \\ &\leq L\int_0^t |x_{\underline{i}} - y_{\underline{i}}| + |x_{\underline{j}} - y_{\underline{j}}|ds + Ce^{-\beta(1-\cos(\alpha/4))}T. \end{aligned}$$

where we used the previous lemma and the fact that $\frac{e^{\beta}}{Z_{\beta}(x_{\underline{i}})} \approx 1 - e^{-\beta(1-d_s)} \approx 1$ thanks to the propertyon T_{ϵ}. The case $k \neq \underline{i}, \underline{j}$ is similar. Hence:

$$\sum_{k=1}^{N} |x_k(t) - y_k(t)| \leq L \int_0^T \sum_{k=1}^{N} |x_k(t) - y_k(t)| + Ce^{-\beta(1-\cos(\alpha/4))}T.$$

The conclusion is then just an application of Gronwall's lemma. □

D Useful lemmas

Lemma D.1. *Let $k > 0$, $\beta \to \infty$. Then:*

$$\int_{\mathbb{S}^{d-1}} (1 - \langle x, y \rangle)^{\frac{k}{2}} e^{\beta \langle x, y \rangle} dy \sim C_{d,k} \beta^{-\frac{d-1+k}{2}} e^{\beta}.$$

Proof. In the following $C_{d,k}$ is a constant that depends just on the dimension and on k and could change at each line:

$$\begin{aligned}
\int_{\mathbb{S}^{d-1}} (1 - \langle x, y \rangle)^{\frac{k}{2}} e^{\beta \langle x, y \rangle} dy &= C_{d,k} \int_{-1}^{1} (1-t)^{\frac{k}{2}} e^{\beta t} (1-t^2)^{\frac{d-3}{2}} dt \\
&= C_{d,k} \int_{0}^{1} (2-2u)^{\frac{k}{2}} e^{\beta(2u-1)} (4u(1-u))^{\frac{d-3}{2}} du \\
&= C_{d,k} e^{-\beta} \int_{0}^{1} (1-u)^{\frac{k+d-3}{2}} u^{\frac{d-3}{2}} e^{2\beta u} du \\
&= C_{d,k} e^{-c} M\left(\frac{d-1}{2}, \frac{k}{2} + d - 1, 2\beta\right),
\end{aligned}$$

where M is the Kummer's confluent hypergeometric function and its asymptotic behavior for $\beta \to \infty$ (see [1]) is given by:

$$I \sim C_{d,k} \beta^{-\frac{d-1+k}{2}} e^{\beta}.$$

□

Lemma D.2. *Suppose that T is a tensor such that:*

- *T is invariant under permutations of the indices.*
- *T is invariant under rotations that fix a unit vector x.*

Then if T is a 2-tensor, then it must be of the form:

$$T = \alpha(x \otimes x) + \beta Id.$$

If T is a 3-tensor, then it must be of the form:

$$T = \alpha(x \otimes x \otimes x) + \beta Sym(x \otimes I)$$

Proof. For simplicity, suppose $d \geq 5$. Without loss of generality we can assume $x = e_1$. Let's start with the case of the 2-tensor. For every $i \neq j$, we can consider another index $l \notin \{1, i, j\}$ and the rotation R such that $Re_i = -e_i$ (wlog $i \neq 1$, otherwise use j), $Re_l = -e_l$ and elsewhere is the identity. Then:

$$T[e_i, e_j] = T[Re_i, Re_j] = -T[e_i, e_j],$$

that implies $T[e_i, e_j] = 0$. If $i = j > 1$, then there exists a rotation R such that $Re_i = e_2$, $Re_2 = -e_i$ and the identity elsewhere:

$$T[e_i, e_i] = T[Re_i, Re_i] = T[e_2, e_2]$$

This conclude the proof for the 2-tensor.

For the 3-tensor: consider i, j, k such that $i > 1$ and ($i \notin \{j, k\}$ or $i = j = k$) . Consider another index $l \notin \{1, i, j, k\}$ and the rotation R such that $Re_i = -e_i$, $Re_l = -e_l$ and identity elsewhere. Then:

$$T[e_i, e_j, e_k] = T[Re_i, Re_j, Re_k] = -T[e_i, e_j, e_k],$$

that implies $T[e_i, e_j, e_k] = 0$. The only cases left are given by $i = j$ and $k = 1$ and their permutations. If $i = j > 1$ and $k = 1$, then construct the rotation R such that $Re_i = e_2$ and $Re_2 = -e_i$ to conclude, as above, that the tensor must be of the form $\alpha(x_i\delta_{jk} + x_j\delta_{ik} + x_k\delta_{ij}) + \beta x_i x_j x_k$. □

E Supplementary figures

This experiment uses the same settings as in Figure 2 in the backward regime, but with an initial distribution given by a mixture of four wrapped Gaussians on the xy-plane. The red curve shows the interaction energy of the system over time on a logarithmic timescale. We highlight the three distinct timescales and the corresponding behaviors discussed in the paper.

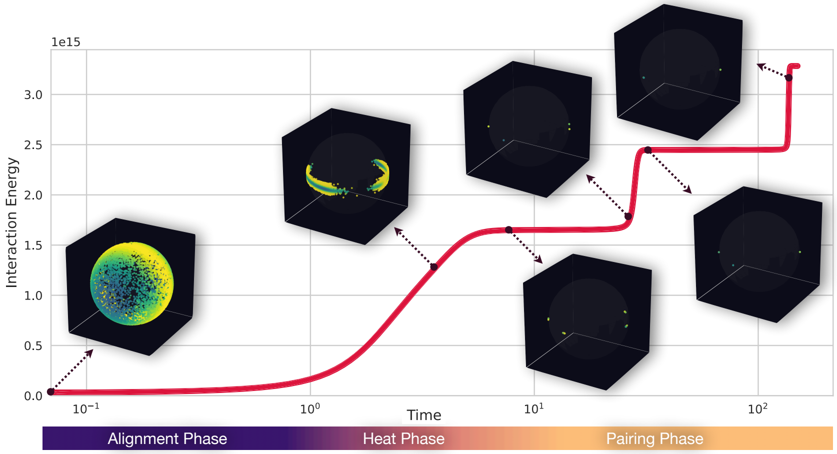

Figure 4: Evolution of the dynamics with $Q^t K = V = S$ definite positive.