QUANTITATIVE CLUSTERING IN MEAN-FIELD TRANSFORMER MODELS

SHI CHEN, ZHENGJIANG LIN, YURY POLYANSKIY, AND PHILIPPE RIGOLLET

Abstract. The evolution of tokens through a deep transformer models can be modeled as an interacting particle system that has been shown to exhibit an asymptotic clustering behavior akin to the synchronization phenomenon in Kuramoto models. In this work, we investigate the long-time clustering of mean-field transformer models. More precisely, we establish exponential rates of contraction to a Dirac point mass for any suitably regular initialization under some assumptions on the parameters of transformer models, any suitably regular mean-field initialization synchronizes exponentially fast with some quantitative rates.

Contents

1. Introduction

The (self-)attention mechanism, initially introduced by [BCB15], forms the foundation of the transformer architecture developed in [Vas17]. This revolutionary architecture has become fundamental for large language models (LLMs), catalyzing remarkable advances in artificial intelligence.

Recently, [GLPR25] proposed to study how a deep stack of attention layers processes information as a mean-field interacting particle system on the sphere $\mathbb{S}^{d-1}$ that exhibits long-time clustering properties; see also [SABP22, GLPR24, KPR24, KBH24, SS24, CRMB24, GKPR24, BPA25, AST24, BKK$^+$25, CACP25].

This model—called *attention dynamics*—captures the representation of *tokens* as they evolve through the successive layers of a transformer. In particular, the clustering phenomenon put forward in [GLPR24, GLPR25] is critical to understanding the structure of internal representations for these pervasive models. More specifically, in attention dynamics, n tokens $x_1, \ldots, x_n \in \mathbb{S}^{d-1}$ evolve as

$$\dot{x}_i(t) = \mathbf{P}_{x_i(t)}\Big[\frac{1}{n}\sum_{j=1}^{n} x_j(t) e^{\beta\langle x_i(t), x_j(t)\rangle}\Big] \qquad t \geqslant 0,\ i = 1, \ldots, n\,, \tag{1.1}$$

where $\beta \geqslant 0$, $\mathbf{P}_x[y] := y - \langle x, y\rangle x$ denotes the projection of $y \in \mathbb{R}^d$ onto the hyperplane $T_x\mathbb{S}^{d-1}$ tangent to the sphere at $x \in \mathbb{S}^{d-1}$. We refer to [GLPR25] for a derivation of this model and its relationship to the attention mechanism and layer normalization. When $d = 2$ and $\beta = 0$ attention dynamics coincide with the well-known Kuramoto model [Kur75, KK84, ABPV+05, BCM14]. It was observed and demonstrated in various situations that these tokens converge to a single token: $x_i(t) \to x_\infty$ as $t \to \infty$. This phenomenon is called *synchronization* or simply *clustering* and we use these terms interchangeably.

Note that the system of ODEs (1.1) is of the mean-field type. Indeed token i interact with all tokens only through their empirical distribution at time t. We denote this distribution by μ_t and recall that

$$\mu_t \coloneqq \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i(t)}\,.$$

In turn, the evolution of μ_t is governed by the continuity equation

$$\partial_t \mu_t + \mathring{\operatorname{div}}\,(\mu_t \mathcal{X}_{\mu_t,\beta}) = 0\,, \qquad \mathcal{X}_{\mu_t,\beta}(x) \coloneqq \int_{\mathbb{S}^{d-1}} \mathbf{P}_x[y] e^{\beta\langle x, y\rangle}\,\mathrm{d}\mu_t(y)\,, \tag{1.2}$$

where here and throughout the paper $\mathring{\operatorname{div}} = \mathring{\operatorname{div}}_{\mathbb{S}^{d-1}}$ denotes the divergence operator on the sphere.

As pointed out in [GLPR25], equation (1.2) describes a Wasserstein gradient flow that aims to maximize the functional

$$\mu \mapsto \mathsf{E}_\beta[\mu] \coloneqq \frac{1}{2\beta}\iint e^{\beta\langle x, y\rangle}\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\,, \tag{1.3}$$

where both integrals are over $\mathbb{S}^{d-1}$. It is easy to see that E is maximized at Dirac point masses δ_{x_0} for some $x_0 \in \mathbb{S}^{d-1}$. This maximum energy state corresponds to a clustering of the tokens into a single point. Thanks to these observations, clustering of n tokens hinges on three classical tools from finite dimensional dynamical systems theory: the dynamics for the the n-tuple $(x_1(t), \ldots, x_n(t)) \in (\mathbb{S}^{d-1})^n$ can be shown to (i) converge by the Łojasiewicz inequality, and (ii) avoid saddle points from almost every initialization by the center-stable manifold theorem. Moreover, all stationary points are saddle points except for the global maximizers where $x_1 = \cdots = x_n$; [GLPR25, KPR24, CRMB24, MTG17].

The number n of tokens that can be processed simultaneously by a transformer model is called *context length* and scaling up its value is a major engineering endeavor because of its direct impact on performance—current frontier models handle contexts with millions of tokens [Goo24]. However, past work on attention dynamics has largely focused on studying asymptotics where $t \to \infty$ and n remains finite implicitly assuming that $n \ll t$.

In this work, we investigate clustering properties for a *continuum* of tokens corresponding to $n = \infty$. The mean-field dynamics of the measure μ_t of tokens is governed by the continuity equation (1.2) but we focus on the case where it is initialized at a measure μ_0 that admits a density with respect to the uniform measure on the sphere $\mathbb{S}^{d-1}$. We call[1] this setup *mean-field attention dynamics*. Despite recent efforts [CACP25] to study convergence of the finite-particle system as $n \to \infty$, existing results do not imply asymptotic clustering for the mean-field attention dynamics for lack of a convergence that is uniform in time. Our results overcome this limitation by developing the infinite-dimensional tools necessary to studying directly the mean-field dynamics.

More precisely, our contributions are as follows. First, we show that, echoing the finite-dimensional case, stationary points for (1.2) are all saddle points for the interaction energy E except for global maxima given by point masses. In particular, our proof extends the approach of [CRMB24] by exhibiting escape directions for continuous measures. However, in the absence of a counterpart to the center-stable manifold theorem in infinite dimensions, this result is not sufficient to conclude to clustering. In fact, while infinite-dimensional versions of the Łojasiewicz inequality have been developed [Sim83, CM14], we show in Remark 2.2 that such inequalities cannot hold in general at critical points of the interaction energy E.

Nevertheless, we demonstrate that a stronger version of the Łojasiewicz inequality, known as the Polyak-Łojasiewicz (PL) inequality, holds around point masses for measures supported on a spherical cap. From such PL inequalities, it follows readily that the Wasserstein gradient flow (1.2) converges exponentially fast to a global maximizer of E when initialized on these measures with constrained support.

This PL inequality is employed in our main contribution, Theorem 2.4, which establishes exponential rates of convergence for the mean-field attention dynamics (1.2) initialized at *any* density $f_0 \in L^2(\mathbb{S}^{d-1})$ for sufficiently small temperature parameter $\beta < \beta_0$, where $\beta_0 > 0$ depends on f_0. Note that global convergence to point masses cannot hold at arbitrary temperatures. Indeed, for $\beta = 100$, we exhibit an equilibrium for mean-field attention dynamics that does *not* correspond to a single cluster in Example 2.6. This qualitative behavior is in sharp contrast with the Kuramoto model where $\beta = 0$ and for which it can be proved that any regular initialization converges to to a point mass exponentially fast; see [MP22].

Our main results for mean-field attention dynamics are stated in the next section. In fact, these results are corollaries for our general results stated in Section 3. These convergence results cover more general dynamics that correspond to less simplified versions of transformer models; see [GLPR25].

2. Clustering in mean-field attention dynamics

In this section, we present our main clustering results on mean-field attention dynamics (1.2).

Recall from [GLPR25] that the mean-field attention dynamics form a reverse Wasserstein gradient flow of the interaction energy E_β defined in (1.3): $\mathcal{X}_{\mu,\beta} = \mathbb{W}\mathsf{E}_\beta[\mu]$—see [CNWR24, AGS05] for an introduction to Wasserstein gradient flows.

[1] While the term "mean-field" technically applies to the Vlasov PDE (1.2) with any initialization, including a discrete one, it is common in the literature to use this term to denote such an evolution initialized at the measure that is absolutely continuous with respect to the uniform measure. To facilitate reading, we adopt the same abuse of language and use "mean-field" to indicate such an initialization.

Indeed, along (1.2) we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}_\beta[\mu_t] = \int_{\mathbb{S}^{d-1}} \|\mathcal{X}_{\mu_t,\beta}(x)\|_2^2 \,\mathrm{d}\mu_t(x) \geqslant 0\,, \tag{2.1}$$

with equality if and only if $\mathcal{X}_{\mu_t,\beta}(x) = 0$ for μ_t almost every $x \in \mathbb{S}^{d-1}$. This equality case characterizes critical points of the energy E_β. The next result shows that the only critical points that are local maxima for E_β are in fact single point masses.

Proposition 2.1. *Let $d \geqslant 3$. For any $\beta > 0$, any local maxima of the interaction energy E_β is a global maxima of the form $\mu = \delta_{x_0}$ for some $x_0 \in \mathbb{S}^{d-1}$.*

Proposition 2.1 is a direct consequence of Theorem 3.1 and holds for more general transformer models, including ones with learned parameters; see Section 3.

When $\mu \in \mathcal{P}(\mathbb{S}^{d-1})$ consists of only a finite number of tokens, similar results on the absence of nontrivial local maxima were described in [GLPR25, CRMB24] when $d \geqslant 3$ following [MTG17]. Our proof of Proposition 2.1 is adapted from [CRMB24]. We note that this technique only applies to $d \geqslant 3$ and the clustering of n tokens for attention dynamics has been recently extended to $d = 2$ in [PRY25] by refining the strategy initiated in [GLPR25].

As mentioned in the introduction, this result is not sufficient to conclude to a global convergence of the mean-field attention dynamics (1.2) to a point mass in absence because of the infinite dimensional nature of the problem. Nevertheless, using the Łojasiewicz structure theorem, one can see that critical points for E_β can only be supported on a finite union of submanifolds of $\mathbb{S}^{d-1}$ of dimension at most $d-2$; see for example Lemma E.5 of [BPA25]. In particular, no stationary points of the mean-field attention dynamics admits a density with respect to the uniform measure other than the uniform measure itself.

Additionally, even convergence of the mean-field attention dynamics to a single limiting stationary point is unclear because of the infinite-dimensional nature of the problem. Indeed, while it is a Wasserstein gradient flow, E_β lacks the Wasserstein geodesic convexity/concavity properties to ensure convergence. In finite dimensions, this limitation may be overcomed using Łojasiewicz inequality whenever the objective function, say f on a compact manifold is analytic. Indeed, in this case, [Łoj63] proved that for any critical point x_{crit} of f, there exists a neighborhood U of x_{crit} and constants $c_1 \in (1,\infty)$, $c_2 > 0$, such that for all $x \in U$,

$$|f(x) - f(x_{\mathrm{crit}})| \leqslant c_2\|\nabla f(x)\|_2^{c_1}\,. \tag{2.2}$$

As a direct corollary, we see that the critical values of f are locally discrete because if $x \in U$ and $\nabla f(x) = 0$, then (2.2) implies $f(x) = f(x_{\mathrm{crit}})$. This last observation is instrumental in establishing convergence of gradient flows of analytic functions. Unfortunately, this property does not hold in general for the energy functional E_β as illustrated by the following example.

Example 2.2 (No Łojasiewicz inequality for E_β)**.** *Let $d = 2$ and consider the energy function E_β for measures defined on the unit circle $\mathbb{S}^1 \subseteq \mathbb{R}^2$ identified to $\mathbb{R}/2\pi\mathbb{Z}$. Take the sequence of measures $\mu_\varepsilon = (1-\varepsilon)\delta_{\frac{\pi}{2}} + \varepsilon\delta_{-\frac{\pi}{2}}$, $\varepsilon \in (0,1)$. Observe that μ_ε forms a sequence of critical points for E_β because $\mathbb{\nabla}\mathsf{E}_\beta[\mu_\varepsilon](\cdot) = 0$ μ_ε almost everywhere. But $\mathsf{E}_\beta[\mu_\varepsilon] \neq \mathsf{E}_\beta[\mu_0]$ and $W_2(\mu_\varepsilon,\mu_0) \to 0$ as $\varepsilon \to 0$, where W_2 denotes the 2-Wasserstein distance. This implies that the critical values of E_β are not necessarily locally discrete. Hence, a Wasserstein version of (2.2) cannot hold for E_β on $\mathcal{P}(\mathbb{S}^{d-1})$ as argued above.*

Example 2.2 reveals a striking discrepancy between the mean-field dynamics studied here and the ones for a finite number of tokens. Indeed, the map:

$$(x_1, \ldots, x_n) \mapsto \frac{1}{n^2} \sum_{i,j=1}^{n} e^{\beta \langle x_i, x_j \rangle}$$

is analytic on the compact manifold $(\mathbb{S}^{d-1})^n$ so the Łojasiewicz inequality holds for a finite number n of tokens. This discrepancy stems from the infinite-dimensional nature of the space of probability measures.

The following result shows that if we rule out sequences that place mass outside of a spherical cap around x_0 then a strong version of the Łojasiewicz inequality, called Polyak-Łojasiewicz (PL) holds.

Theorem 2.3 (Polyak-Łojasiewicz inequality on a spherical cap)**.** *Fix $d \geqslant 2, \beta > 0, \alpha \in [0, \pi/2), u \in \mathbb{S}^{d-1}$ and let $S_\alpha^+(u) \subset \mathbb{S}^{d-1}$ denote the spherical cap of angle α around u defined by*

$$S_\alpha^+(u) \coloneqq \left\{ x \in \mathbb{S}^{d-1} \mid \langle x, u \rangle \geqslant \cos\alpha \right\}. \tag{2.3}$$

Let μ be a probability measure supported on $S_\alpha^+(u)$. Then if $10(1+\sqrt{\beta})\tan\alpha \leqslant 1$, the following PL inequality holds

$$\mathsf{E}_\beta[\delta_u] - \mathsf{E}_\beta[\mu] \leqslant 10 e^{-\beta} \int_{\mathbb{S}^{d-1}} \|\mathcal{X}_{\mu,\beta}(x)\|_2^2 \, \mathrm{d}\mu(x) \,.$$

As a result, the sequence of measures $\mu_t, t \geqslant 0$ initialized at $\mu_0 = \mu$ supported on $S_\alpha^+(u)$ and evolving according to (1.2) converges to a single point mass δ_{x_∞} with $\langle x_\infty, u \rangle \geqslant \cos\alpha$ at an exponential rate given by

$$W_2(\mu_t, \delta_{x_\infty}) \leqslant 20 e^{-\beta} e^{-\frac{e^\beta}{20} t} \left(\int_{\mathbb{S}^{d-1}} \|\mathcal{X}_{\mu,\beta}(x)\|_2^2 \, \mathrm{d}\mu(x) \right)^{\frac{1}{2}} .$$

Note that $\mathsf{E}_\beta[\delta_u] = \max_{\mu \in \mathcal{P}(\mathbb{S}^{d-1})} \mathsf{E}_\beta[\mu] = e^\beta$. When μ is a discrete measures supported on a hemisphere of $\mathbb{S}^{d-1}$, i.e., $\mathrm{supp}(\mu) \subseteq S_{\frac{\pi}{2}}^+(u)$ for some $u \in \mathbb{S}^{d-1}$, [GLPR25, Lemma 6.4] obtained a similar exponential synchronization result for transformer models, but the convergence rate there also depends on the initial positions of these tokens and becomes worse when the number of tokens increases. Similar hemisphere initial position assumptions are classical in Kuramoto models ($d = 2$ and $\beta = 0$); see, e.g., [HHK10, CHJK12, FL19, HKMP20, ABK+22].

We are now in a position to state our main result for initializations that need not be supported on a spherical cap. As observed in [MP22] for the Kuramoto model, such measures do not satisfy a PL inequality. Instead, the energy E_β satisfies a second-order differential inequality along the flow μ_t defined in (1.2) with a vanishing remainder term: if μ_0 has a density $f_0 \in L^2(\mathbb{S}^{d-1})$ with respect to the uniform measure,

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2} \mathsf{E}_\beta[\mu_t] \leqslant -\frac{\mathrm{d}}{\mathrm{d}t} \mathsf{E}_\beta[\mu_t] + C_0 e^{-d C_1 t}, \quad \text{for} \quad t > T_0$$

where C_0, T_0 are constants depending on μ_0, C_1 is a universal constant. In turn, this inequality enables us to establish the following result for any initial measure that admits a density $f_0 \in L^2(\mathbb{S}^{d-1})$ with respect to the uniform measure.

Theorem 2.4. *Fix $d \geqslant 2$. Let μ_t evolve according to the mean-field attention dynamics* (1.2) *initialized at μ_0 with mean such that*

$$R_0 := \Big\| \int_{\mathbb{S}^{d-1}} x \mathrm{d}\mu_0(x) \Big\|_2 > 0. \tag{2.4}$$

Assume that μ_0 admits a density $f_0 \in L^2(\mathbb{S}^{d-1})$ with respect to the uniform measure, then μ_t also admits a density $f_t \in L^2(\mathbb{S}^{d-1})$ for all $t > 0$. Moreover, there exist $\beta_0, C_0, T_0 > 0$, all depending on μ_0, such that if $|\beta| < \beta_0$, there exists an $x_\infty \in \mathbb{S}^{d-1}$ for which

$$W_2(\mu_t, \delta_{x_\infty}) \leqslant C_0 e^{-\frac{t}{100}}\,, \quad \text{for} \quad t > T_0\,. \tag{2.5}$$

Note that the convergence in Wasserstein distance to a point mass means that (i) the variance of μ_t converges to zero exponentially fast, and (ii) its mean also converges to x_∞.

The proof of Theorem 2.4 relies on the approximation $e^{\beta\langle x,y\rangle} \simeq 1$ for β small. As such, it can be generalized to more realistic scenarios where $e^{\beta\langle x,y\rangle}$ is replaced with $e^{\langle Q_t x, K_t y\rangle}$ in the definition of the vector field $\mathcal{X}_{\mu_t,\beta}(x)$ as long as $\|Q_t\|_2$, $\|\partial_t Q_t\|_2$, $\|K_t\|_2$, and $\|\partial_t K_t\|_2$ are bounded by a small enough constant, uniformly in time and space. We omit this extension in the present paper.

Theorem 2.4 follows as a special case of the more general Theorem 3.4 that handles broader attention mechanisms. When $\beta = 0$, a qualitative mean-field convergence result was proved in [FL19], and under additional symmetric assumptions on the entire flow $\{f_t(x)\}_t$, an exponential convergence rate of the system was also derived when $\beta = 0$. For Kuramoto models, [HKMP20, MP22] proved similar mean-field exponential convergence results for small frequency terms when $d = 2$ and $\beta = 0$. Long-time behaviors of Kuramoto models have also been extensively studied by [HHK10, HKPZ16, BCM14].

When $\beta > 0$, previous work on long-time convergence focused mostly on the case of delta masses of finitely many tokens instead of the mean-field setting considered here. To the best of our knowledge, Theorem 2.4 is the first result to provide quantitative rates of convergence for attention dynamics (mean-field or finite-particle) under a general initial condition like $R_0 > 0$. Indeed, for a finite number particles, convergence is either established using soft arguments that do not yield convergence rates [MTG17, PRY25] or exponential convergence is established under the assumption that particles are initialized a hemisphere of $\mathbb{S}^{d-1}$, as we mentioned after Theorem 2.3. However, as illustrated in Example 2.5, transformer models starting from delta masses and mean-field densities can have different asymptotic behavior.

Example 2.5 (No synchronization for finite particles)**.** *Fix $d = 2$. We construct an example on the unit circle $\mathbb{S}^1 \subseteq \mathbb{R}^2$ identified to $\mathbb{R}/2\pi\mathbb{Z}$ where particles do not converge to a single cluster despite being initialized at μ_0 that satisfies $R_0 > 0$; see Figure 1. To that end, define $\mu_0 = \frac{1}{50}\delta_{\frac{\pi}{2}} + \frac{49}{100}\delta_{-\frac{\pi}{2}-\xi} + \frac{49}{100}\delta_{-\frac{\pi}{2}+\xi}$ for an $\xi \in (0, \frac{1}{100})$.*

With this initialization, the initial velocity field $\mathcal{X}_{\mu_0,\beta}$ is given for any $\theta \in [0, 2\pi)$ by[2]

$$\mathcal{X}_{\mu_0,\beta}(\theta) = -\int_0^{2\pi} \sin(\theta - \omega) e^{\beta\cos(\theta-\omega)}\, \mathrm{d}\mu_0(\omega)\,. \tag{2.6}$$

[2]This expression follows from a simple change of variables; see [GLPR25, Section 7.1]

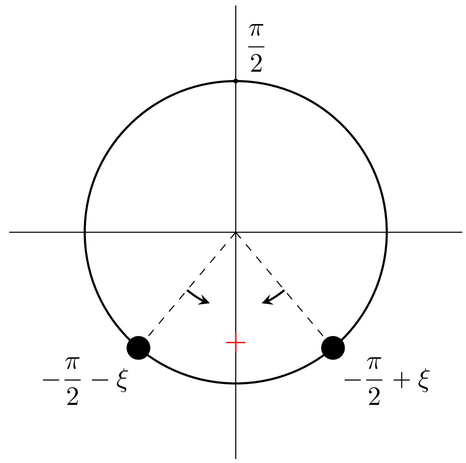

Figure 1. Illustration of μ_0 in Example 2.5 with $\xi = .7$. Circle radii are proportional to mass at each point. The cross indicates the mean of μ_0 with $R_0 > .7$ and the arrows indicate velocity fields for initial angles.

It is easy to see that $\mathcal{X}_{\mu_0,\beta}(\pi/2) = 0$ by symmetry. Moreover, $\mathcal{X}_{\mu_0,\beta}(-\frac{\pi}{2} - \xi) = -\mathcal{X}_{\mu_0,\beta}(-\frac{\pi}{2} + \xi) > 0$ so long as $\beta > 0$, hence the two particles in the south hemisphere get closer and μ_t initialized at μ_0 eventually converges to $\mu_\infty \coloneqq \frac{1}{50}\delta_{\frac{\pi}{2}} + \frac{49}{50}\delta_{-\frac{\pi}{2}}$ as $t \to \infty$. The point of this example is that, although μ_t initialized at μ_0 does not converge to a single point mass, Theorem 2.4 implies that when β is small, any initial measure with an $L^2(\mathbb{S}^{d-1})$-density, which may be arbitrary close to μ_0, contracts to a point mass at an exponential rate. In particular, we see that the contraction rate in Theorem 2.4 must depend on f_0.

We conclude this section by discussing an important limitation of Theorem 2.4, namely that β is required to be small enough. It turns out that this assumption is necessary, as there exists initializations μ_0 for which $R_0 > 0$ and that admit a density $f_0 \in L^2(\mathbb{S}^{d-1})$ for which the mean-field attention dynamics do not converge to a single point mass. We describe such an initialization on the circle in the following example.

Example 2.6 (No mean-field synchronization for large β)**.** *Fix $d = 2$. We construct an example on the unit circle $\mathbb{S}^1 \subseteq \mathbb{R}^2$ identified to $\mathbb{R}/2\pi\mathbb{Z}$ where particles do not converge to a single cluster when β is sufficiently large. To that end, let $\beta = 100$, and consider the flow (1.2) started at μ_0 that admits a density $f_0 \in L^2(\mathbb{R}/2\pi\mathbb{Z})$ with respect to the Lebesgue measure.*

We construct f_0 as follows. Fix $\eta, \xi \in (0, \frac{1}{100})$ and let h_1 be a positive, even, and smooth function supported on $[-\eta, \eta]$ such that h_1 is strictly increasing on $[-\eta, 0]$. We normalize h_1 such that $\int h_1 = 1/3$. Similarly, let h_2 be a positive, even, and smooth function supported on $[-\xi, \xi]$ such that h_2 is strictly increasing on $[-\xi, 0]$ and normalized as $\int h_2 = 2/3$. Finally, let f_0 be defined as $f_0(x) = h_1(x) + h_2(\pi + x)$; see Figure 2 for an illustration.

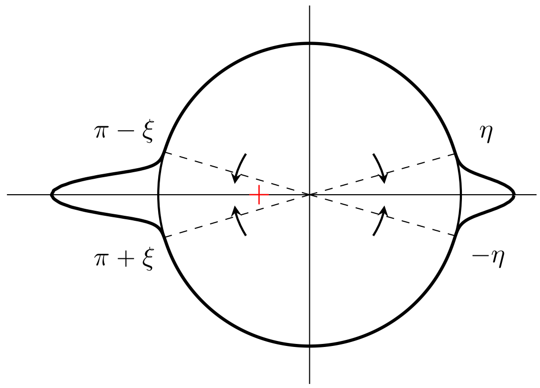

Figure 2. Illustration of f_0 in Example 2.6. The cross indicates the mean of f_0 with $R_0 > 0$, and the arrows indicate velocity fields at the boundaries of the support of f_0.

With this initialization, akin to Example 2.5, the initial velocity field $\mathcal{X}_{\mu_0,100}$ is given for any $\theta \in [0, 2\pi)$ by

$$\mathcal{X}_{\mu_0,100}(\theta) = -\int_0^{2\pi} \sin(\theta-\omega)e^{100\cos(\theta-\omega)} f_0(\omega)\,\mathrm{d}\omega\,. \tag{2.7}$$

Clearly, $\mathcal{X}_{\mu_0,100}(0) = \mathcal{X}_{\mu_0,100}(\pi) = 0$ by symmetry. Moreover, one can easily see that $\mathcal{X}_{\mu_0,100}$ pulls points $\theta \in [\pi-\xi, \pi+\xi]\backslash\{\pi\}$ towards π because the main contribution in $\mathcal{X}_{\mu_0,100}$ at those points comes from the integral on $[\pi-\xi, \pi+\xi]$ in (2.7)*. Similarly, even though $\int h_1 = 1/3 < 2/3 = \int h_2$ and points in $[\pi-\xi, \pi+\xi]$ are trying to pull points in $[-\eta, \eta]$ towards π, their contribution is negligible compared to the pull from antipodal points. Indeed, when $\xi \leqslant \eta$,*

$$\begin{aligned}
&\mathcal{X}_{\mu_0,100}(\eta) = \\
&= \int_{-\xi}^{\xi} \sin(\eta-\omega)e^{-100\cos(\eta-\omega)} h_2(\omega)\,\mathrm{d}\omega - \int_{-\eta}^{\eta} \sin(\eta-\omega)e^{100\cos(\eta-\omega)} h_1(\omega)\,\mathrm{d}\omega \\
&\leqslant \frac{2}{3}\sin(2\eta)e^{-100\cos(2\eta)} - \int_{-\eta}^{\eta/2} \sin(\eta-\omega)e^{100\cos(\eta-\omega)} h_1(\omega)\,\mathrm{d}\omega \\
&\leqslant \frac{2}{3}\sin(2\eta)e^{-100\cos(2\eta)} - \frac{1}{3}\sin\left(\frac{\eta}{2}\right)e^{100\cos(2\eta)} \lesssim -\eta\cdot 10^{38}.
\end{aligned}$$

By symmetry $\mathcal{X}_{\mu_0,100}(-\eta) = -\mathcal{X}_{\mu_0,100}(\eta)$ and we see that the edge of the interval $[-\eta, \eta]$ gets pulled towards 0*. Since trajectories of ODEs cannot cross, all the points in $[-\eta, \eta]$ get pulled towards 0. We can similarly discuss the case when $\xi \geqslant \eta$. Using a bootstrap argument, it can be shown that μ_t converges to $\mu_\infty = \frac{1}{3}\delta_0 + \frac{2}{3}\delta_\pi$.*

3. Clustering in general transformer models

Despite its simplicity, the previous section shows that mean-field attention dynamics (1.2) captures the clustering phenomenon observed in practice. In practice, the attention mechanism is parameterized by matrices that are learned from data during the training process. In general, the vector field $\mathcal{X}_{\mu_t,\beta}$ in (1.2) becomes

$$\widetilde{\mathcal{X}}_{\mu_t}(x) \coloneqq \int_{\mathbb{S}^{d-1}} \mathbf{P}_x[V_t y]\phi(\langle A_t x, y\rangle)\,\mathrm{d}\mu_t(y)\,, \tag{3.1}$$

where $\{V_t, A_t\}_t$ are learned matrices and ϕ is a known nonlinear function; see [GLPR25]. Such a time inhomogeneous system is difficult to study in full generality but we make progress in this direction by considering the case where[3] $V_t = V$ and $A_t = A$ for all t.

3.1. **Critical Points.** In the case where $V = A$, the vector field $\widetilde{\mathcal{X}}_\mu$ is a Wasserstein gradient flow for the energy functional

$$\mathsf{E}_\phi[\mu] \coloneqq \frac{1}{2}\iint \phi\left(\langle Ax, y\rangle\right)\mathrm{d}\mu(x)\,\mathrm{d}\mu(y). \tag{3.2}$$

Here and throughout this paper, unless explicitly stated otherwise, integrals are assumed to be taken over the set $\mathbb{S}^{d-1}$. We leverage this property to characterize the stationary points of the mean-field attention dynamics (1.2) with the more general vector field $\widetilde{\mathcal{X}}_\mu$ in Section 3.1 under additional assumptions on the matrix A.

Hereafter, we assume that ϕ is twice differentiable and that A is a $d \times d$ real symmetric matrix with eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_d$—we allow for some eigenvalues to be negative. The Wasserstein gradient of the interaction energy E_ϕ is given by

$$\widetilde{\mathcal{X}}[\mu](x) \coloneqq \mathbb{\nabla}\mathsf{E}_\phi[\mu](x) = \int_{\mathbb{S}^{d-1}} \mathbf{P}_x[Ay]\phi'\left(\langle Ax, y\rangle\right)\mathrm{d}\mu(y), \quad x \in \mathbb{S}^{d-1}\,. \tag{3.3}$$

Consider the general mean-field attention dynamics

$$\partial_t \mu_t + \mathring{\mathrm{div}}(\mu_t \widetilde{\mathcal{X}}[\mu_t]) = 0\,, \tag{3.4}$$

and observe that they collapse to (1.2) when $A = I_d$ and $\phi(z) = e^{\beta z}$ up to a time speed up.

The following theorem provides a partial resolution of Conjecture 2 in [KPR24] when adapted to non-causal attention dynamics.

Theorem 3.1. *Fix $d \geqslant 3$. Assume that the top three eigenvalues of A satisfy $\lambda_1 = \lambda_2 = \lambda_3 = \lambda > 0$, and $|\lambda_d| \leqslant \lambda$. Assume further that ϕ is twice differentiable, increasing and convex: $\phi' > 0$, $\phi'' \geqslant 0$. Then any local maxima of $\mathsf{E}_\phi[\mu]$ must be a global maximum, that is, a point mass δ_{x_0} for some $x_0 \in \mathbb{S}^{d-1}$ such that $Ax_0 = \lambda x_0$.*

In the rest of Section 3.1, we prove Theorem 3.1. It relies on the first and second variation formulas for $\mathsf{E}_\phi[\mu]$, which are of independent interest in the study of transformer models. We also note that if λ_d is significantly smaller than $-\lambda$, then a global maximizer of $\mathsf{E}_\phi[\cdot]$ need not be a Dirac measure, as shown by a counterexample in Remark 3.5 of [BKK+25].

---

[3]Employing the same weights across layers has been used to reduce the complexity of transformer models [LCG+20] and it has been shows that they demonstrate better reasoning properties in certain tasks [ZBB+23]. This is the model initially studied in [SABP22]

3.1.1. *First and Second Variation Formulas for* E_ϕ. Let $\mathcal{P}(\mathbb{S}^{d-1})$ denote the space of probability measures on $\mathbb{S}^{d-1}$ and let $\{\mathcal{X}_t(x)\,,\ t \geqslant 0,\ x \in \mathbb{S}^{d-1}\}$ be a family of vector fields on $\mathbb{S}^{d-1}$, continuously differentiable in (t,x) and such that $\mathcal{X}_t(x) \in T_x\mathbb{S}^{d-1}$ for all $(t,x) \in \mathbb{R}_{\geqslant 0} \times \mathbb{S}^{d-1}$. Note that $\{\partial_t \mathcal{X}_t(x)\,,\ t \geqslant 0,\ x \in \mathbb{S}^{d-1}\}$ is also a family of continuous vector fields in the tangent bundle of $\mathbb{S}^{d-1}$. Let $\{\mu_t\}_{t\geqslant 0}$ be a curve in $\mathcal{P}(\mathbb{S}^{d-1})$ starting from $\mu_0 \in \mathcal{P}(\mathbb{S}^{d-1})$ and evolving according the continuity equation driven by $\{\mathcal{X}_t\}_t$:

$$\partial_t \mu_t + \mathring{\mathrm{div}}\,(\mu_t \mathcal{X}_t) = 0, \quad t \geqslant 0. \tag{3.5}$$

The PDE (3.5) is understood in the distribution sense: for any smooth function $h(x)$ on $\mathbb{S}^{d-1}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{S}^{d-1}} h(x)\,\mathrm{d}\mu_t(x) = \int_{\mathbb{S}^{d-1}} \langle \mathring{\nabla}_x h(x), \mathcal{X}_t(x)\rangle\,\mathrm{d}\mu_t(x)\,.$$

Here and throughout this paper, $\mathring{\nabla}_x\, h(x)$ denotes the Riemannian gradient of h on $\mathbb{S}^{d-1}$. Note that viewing $\mathbb{S}^{d-1}$ as an embedded manifold in $\mathbb{R}^d$ considerably simplifies the Riemannian calculus on the sphere. Indeed, if H is a smooth extension of h to a neighborhood of $\mathbb{S}^{d-1}$ in $\mathbb{R}^d$, we have that $\mathring{\nabla}_x\, h(x) = \mathbf{P}_x \nabla_x H(x)$. In particular, since $\mathcal{X}_t(x) \in T_x\mathbb{S}^{d-1}$, we have $\langle \mathring{\nabla}_x h(x), \mathcal{X}_t(x)\rangle = \langle \nabla_x H(x), \mathcal{X}_t(x)\rangle$.

Lemma 3.2 (First Variation Formula for E_ϕ)**.**

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}_\phi[\mu_t] = \iint \phi'(\langle Ax, y\rangle)\langle Ay, \mathcal{X}_t(x)\rangle\,\mathrm{d}\mu_t(x)\,\mathrm{d}\mu_t(y).$$

Proof. Because A is a symmetric matrix, using (3.5), we have that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathsf{E}_\phi[\mu_t] = \iint \langle \nabla_x\left[\phi(\langle Ax, y\rangle)\right], \mathcal{X}_t(x)\rangle\,\mathrm{d}\mu_t(x)\,\mathrm{d}\mu_t(y)\,,$$

where we used the fact that $\mathcal{X}_t(x) \in T_x\mathbb{S}^{d-1}$. The proof follows readily by computing the gradient above. $\square$

Lemma 3.3 (Second Variation Formula for $\mathsf{E}_\phi[\,\cdot\,]$)**.**

$$\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathsf{E}_\phi[\mu_t] = \frac{1}{2}\iint \phi''\left(\langle Ax, y\rangle\right)\|\langle Ay, \mathcal{X}_t(x)\rangle + \langle Ax, \mathcal{X}_t(y)\rangle\|_2^2\,\mathrm{d}\mu_t(x)\,\mathrm{d}\mu_t(y) \tag{3.6}$$

$$+ \iint \phi'(\langle Ax, y\rangle)\langle A\mathcal{X}_t(x), \mathcal{X}_t(y)\rangle\,\mathrm{d}\mu_t(x)\,\mathrm{d}\mu_t(y) \tag{3.7}$$

$$- \frac{1}{2}\iint \phi'(\langle Ax, y\rangle)\langle Ax, y\rangle(\|\mathcal{X}_t(x)\|_2^2 + \|\mathcal{X}_t(y)\|_2^2)\,\mathrm{d}\mu_t(x)\,\mathrm{d}\mu_t(y) \tag{3.8}$$

$$+ \iint \phi'(\langle Ax, y\rangle)\left[\left\langle Ay, \partial_t \mathcal{X}_t(x) + \mathring{\nabla}_{\mathcal{X}_t(x)}\mathcal{X}_t(x)\right\rangle\right]\mathrm{d}\mu_t(x)\,\mathrm{d}\mu_t(y). \tag{3.9}$$

Proof. Taking the time derivative of the first variation formula, we get

$$\begin{aligned}\frac{\mathrm{d}^2}{\mathrm{d}t^2}\mathsf{E}_\phi[\mu_t] = & \underbrace{\iint \left\langle \nabla_x \left[\phi'(\langle Ax, y\rangle)\langle Ay, \mathcal{X}_t(x)\rangle\right], \mathcal{X}_t(x)\right\rangle \mathrm{d}\mu_t(x)\,\mathrm{d}\mu_t(y)}_{\triangleleft} \\ & + \underbrace{\iint \left\langle \nabla_y \left[\phi'(\langle Ax, y\rangle)\langle Ay, \mathcal{X}_t(x)\rangle\right], \mathcal{X}_t(y)\right\rangle \mathrm{d}\mu_t(x)\,\mathrm{d}\mu_t(y)}_{\triangleright} \\ & + \iint \phi'(\langle Ax, y\rangle)\langle Ay, \partial_t\mathcal{X}_t(x)\rangle \,\mathrm{d}\mu_t(x)\,\mathrm{d}\mu_t(y).\end{aligned}$$

The rest of the proof follows by direct computations, and we only highlight the key points.

For $\triangleleft$, as $\mathcal{X}_t(x) \in T_x\mathbb{S}^{d-1}$, we get that

$$\begin{aligned}&\langle \mathring{\nabla}_x \langle Ay, \mathcal{X}_t(x)\rangle, \mathcal{X}_t(x)\rangle = \langle \mathring{\nabla}_{\mathcal{X}_t(x)}[\mathbf{P}_x Ay], \mathcal{X}_t(x)\rangle + \langle \mathbf{P}_x Ay, \mathring{\nabla}_{\mathcal{X}_t(x)}\mathcal{X}_t(x)\rangle \\ &= \langle \nabla_{\mathcal{X}_t(x)}[\mathbf{P}_x Ay], \mathcal{X}_t(x)\rangle + \langle Ay, \mathring{\nabla}_{\mathcal{X}_t(x)}\mathcal{X}_t(x)\rangle \\ &= -\langle Ay, x\rangle \|\mathcal{X}_t(x)\|_2^2 + \langle Ay, \mathring{\nabla}_{\mathcal{X}_t(x)}\mathcal{X}_t(x)\rangle,\end{aligned}$$

where in the last equality, we used the fact that

$$\nabla_x[\mathbf{P}_x Ay] = -\nabla_x[\langle Ay, x\rangle x] = -(Ay)\otimes x - \langle Ay, x\rangle I_d,$$

and $\langle x, \mathcal{X}_t(x)\rangle = 0$.

Similarly, for $\triangleright$, as $\mathcal{X}_t(y) \in T_y\mathbb{S}^{d-1}$, we get that

$$\langle \mathring{\nabla}_y \langle Ay, \mathcal{X}_t(x)\rangle, \mathcal{X}_t(y)\rangle = \langle \nabla_y \langle Ay, \mathcal{X}_t(x)\rangle, \mathcal{X}_t(y)\rangle = \langle A\mathcal{X}_t(x), \mathcal{X}_t(y)\rangle.$$

The final form of the second variation formula can be obtained from the symmetric role of x and y. □

Equipped with the first and second variation formulas, we are now in a position to prove Theorem 3.1.

3.1.2. *Proof of Theorem 3.1.* Let μ_0 be a critical point of E_ϕ. We show that unless μ_0 is a point mass, there exists an escape direction, that is, a velocity field $\mathcal{X}_0$ such that if μ_t evolves according to (3.5), then the value of E_ϕ increases. Since μ_0 is a stationary point, it is sufficient to check that

$$\left.\frac{\mathrm{d}^2}{\mathrm{d}t^2}\right|_{t=0} \mathsf{E}_\phi[\mu_t] > 0.$$

Since μ_0 is a critical point, it follows from the first variation formula that

$$\iint \phi'(\langle Ax, y\rangle)\langle Ay, \mathcal{X}(x)\rangle \,\mathrm{d}\mu_0(x)\,\mathrm{d}\mu_0(y) = 0, \quad \forall\ \mathcal{X} \in C(T\mathbb{S}^{d-1}). \tag{3.10}$$

At such critical points, taking $\mathcal{X}(x) = \mathring{\nabla}_{\mathcal{X}_0(x)}\,\mathcal{X}_0(x)$ in (3.10), the second variation formula simplifies to

$$\begin{aligned}\left.\frac{\mathrm{d}^2}{\mathrm{d}t^2}\right|_{t=0} \mathsf{E}_\phi[\mu_t] = & \iint \frac{1}{2}\phi''\left(\langle Ax, y\rangle\right) \|\langle Ay, \mathcal{X}_0(x)\rangle + \langle Ax, \mathcal{X}_0(y)\rangle\|_2^2 \,\mathrm{d}\mu_0(x)\,\mathrm{d}\mu_0(y) \\ & + \iint \frac{1}{2}\phi'(\langle Ax, y\rangle)\left[2\langle A\mathcal{X}_0(x), \mathcal{X}_0(y)\rangle - \langle Ax, y\rangle(\|\mathcal{X}_0(x)\|_2^2 + \|\mathcal{X}_0(y)\|_2^2)\right] \mathrm{d}\mu_0(x)\,\mathrm{d}\mu_0(y).\end{aligned} \tag{3.11}$$

Recall that we assumed $\phi'' \geqslant 0$ so we focus on establishing the positivity of the second line in (3.11). To that end, following [MTG17, CRMB24], define $\mathcal{X}_0(x) = \mathbf{P}_x(w) = w - \langle w, x\rangle x$ where $w \in \mathbb{S}^{d-1}$. The second line in (3.11) becomes

$$\begin{aligned}&\iint \frac{1}{2}\phi'(\langle Ax, y\rangle)\left[2\langle A\mathcal{X}_0(x), \mathcal{X}_0(y)\rangle - \langle Ax, y\rangle(\|\mathcal{X}_0(x)\|_2^2 + \|\mathcal{X}_0(y)\|_2^2)\right] \mathrm{d}\mu_0(x)\,\mathrm{d}\mu_0(y)\\ &= \iint \frac{1}{2}\phi'(\langle Ax, y\rangle)\big[2\left(\langle Aw, w\rangle - \langle w, x\rangle\langle x, Aw\rangle - \langle w, y\rangle\langle y, Aw\rangle + \langle w, x\rangle\langle w, y\rangle\langle Ax, y\rangle\right)\\ &\quad - \langle Ax, y\rangle\left(2 - \langle w, x\rangle^2 - \langle w, y\rangle^2\right)\big]\,\mathrm{d}\mu_0(x)\,\mathrm{d}\mu_0(y).\end{aligned} \tag{3.12}$$

Pick $\{e_i\}_{i=1}^d$ as an orthonormal basis of $\mathbb{R}^d$ such that $Ae_i = \lambda_i e_i$ for $i = 1, \ldots, d$. We also write x, y in the coordinates of $\{e_i\}_{i=1}^d$, that is, $x = \sum_{i=1}^d x_i e_i$ and $y = \sum_{i=1}^d y_i e_i$. Choosing $w = e_i$ in (3.12) yields

$$\iint \frac{1}{2}\phi'(\langle Ax, y\rangle)\left[2\lambda_i(1 - x_i^2 - y_i^2) - \langle Ax, y\rangle\left(2 - x_i^2 - y_i^2 - 2x_iy_i\right)\right] \mathrm{d}\mu_0(x)\,\mathrm{d}\mu_0(y). \tag{3.13}$$

We aim to prove the following inequality:

$$\sum_{i=1}^{3}\iint \phi'(\langle Ax, y\rangle)\left[2\lambda_i(1 - x_i^2 - y_i^2) - \langle Ax, y\rangle\left(2 - x_i^2 - y_i^2 - 2x_iy_i\right)\right] \mathrm{d}\mu_0(x)\,\mathrm{d}\mu_0(y) \geqslant 0, \tag{3.14}$$

with equality if and only if $\mu_0 = \delta_u$ for some point $u \in \mathbb{S}^{d-1}$ satisfying $Au = \lambda u$. This inequality directly yields the desired conclusion, since (3.14) implies that there is an $i \in \{1, 2, 3\}$ such that (3.12) with $w = e_i$ is strictly positive unless $\mu_0 = \delta_u$ for some point $u \in \mathbb{S}^{d-1}$ with $Au = \lambda u$.

To prove (3.14), we build up the pointwise inequality: for any $x, y \in \mathbb{S}^{d-1}$,

$$\sum_{i=1}^{3} 2\lambda_i(1 - x_i^2 - y_i^2) - \langle Ax, y\rangle\left(2 - x_i^2 - y_i^2 - 2x_iy_i\right) \geqslant 0, \tag{3.15}$$

with equality if and only if the following conditions hold: for all j with $|\lambda_j| < \lambda$, we have $x_j = y_j = 0$; for j with $\lambda_j = \lambda$, we have $x_j = y_j$; for j with $\lambda_j = -\lambda$, we have $x_j = -y_j$.

We now use (3.15) to prove (3.14). Suppose there exists $u \in \mathrm{supp}(\mu_0)$ such that $Au \neq \lambda u$. Writing $u = \sum_{i=1}^d u_i e_i$, this implies that there exists some i such that $\lambda_i < \lambda$ and $u_i \neq 0$. Choose a small neighborhood $\mathcal{N} \subset \mathbb{S}^{d-1}$ around u, such that $|\langle x, e_i\rangle - u_i| < |u_i|/2$ for any $x \in \mathcal{N}$. Then, for any $x, y \in \mathcal{N}$, the inequality (3.15) is strictly positive because x, y don't satisfy the equality condition. Therefore, the integrand in (3.14) is strictly positive on $\mathcal{N} \times \mathcal{N}$ (since $\phi' > 0$), and nonnegative elsewhere by (3.15). This implies that (3.14) is strictly positive.

In the remaining case, suppose $Au = \lambda u$ for every $u \in \mathrm{supp}(\mu_0)$. Then, the equality conditions in (3.15) imply that μ_0 must be a point mass. Otherwise, there exists $x \neq y$ in the support of μ_0, and (3.15) becomes strictly positive, yielding a strictly positive value in (3.14). Hence, (3.14) is strictly positive unless $\mu_0 = \delta_u$ for some $u \in \mathbb{S}^{d-1}$ satisfying $Au = \lambda u$.

The rest of this proof is devoted to the proof of (3.15) and its equality conditions. Since $\lambda_1 = \lambda_2 = \lambda_3 =: \lambda > 0$, the left hand side of (3.15) becomes

$$\begin{aligned} &2\lambda\left(3-\sum_{i=1}^{3}(x_i^2+y_i^2)\right)-\sum_{i=1}^{3}\langle Ax,y\rangle\left(2-x_i^2-y_i^2-2x_iy_i\right)\\ &=2\left(\lambda-\langle Ax,y\rangle\right)\left(3-\sum_{i=1}^{3}(x_i^2+y_i^2)\right)+\langle Ax,y\rangle\sum_{i=1}^{3}(2x_iy_i-x_i^2-y_i^2). \end{aligned} \tag{3.16}$$

We claim that

$$2\left(\lambda-\langle Ax,y\rangle\right)\geqslant\lambda\sum_{i=1}^{3}(x_i^2+y_i^2-2x_iy_i). \tag{3.17}$$

Indeed $\|x\|_2^2=\|y\|_2^2=1$ and $\lambda_1=\lambda_2=\lambda_3=\lambda$ so that

$$\begin{aligned} 2(\lambda-\langle Ax,y\rangle)&=\lambda\|x\|^2+\lambda\|y\|^2-2\sum_{i=1}^{d}\lambda_ix_iy_i\\ &=\lambda\sum_{i=1}^{3}(x_i^2+y_i^2-2x_iy_i)+\lambda\sum_{i=4}^{d}(x_i^2+y_i^2)-2\sum_{i=4}^{d}\lambda_ix_iy_i\,. \end{aligned}$$

Hence, (3.17) is equivalent to

$$\sum_{i=4}^{d}\lambda(x_i^2+y_i^2)\geqslant2\sum_{i=4}^{d}\lambda_ix_iy_i, \tag{3.18}$$

which holds since $\lambda\geqslant|\lambda_i|$ for all i. Moreover, note that the equality holds if and only if: $x_i=y_i$ for all $i\geqslant4$ with $\lambda_i=\lambda$; $x_i=-y_i$ for all $i\geqslant4$ with $\lambda_i=-\lambda$; $x_i=y_i=0$ for all $i\geqslant4$ with $|\lambda_i|<\lambda$. Hence, by (3.16) and (3.17), we have that

$$\begin{aligned} &\sum_{i=1}^{3}2\lambda_i(1-x_i^2-y_i^2)-\langle Ax,y\rangle\left(2-x_i^2-y_i^2-2x_iy_i\right)\\ &\geqslant\left(\sum_{i=1}^{3}x_i^2+y_i^2-2x_iy_i\right)\left(3\lambda-\lambda\sum_{i=1}^{3}(x_i^2+y_i^2)-\langle Ax,y\rangle\right). \end{aligned} \tag{3.19}$$

Because $|\langle Ax,y\rangle|\leqslant\lambda\|x\|_2\|y\|_2=\lambda$, and $\sum_{i=1}^{3}(x_i^2+y_i^2)\leqslant\sum_{i=1}^{d}(x_i^2+y_i^2)=2$, we see that the right hand side of (3.19) is nonnegative, and it can only be 0 when $x_1=y_1$, $x_2=y_2$, $x_3=y_3$. Hence, we complete the proof for (3.15). By examining the equality conditions in (3.19) and (3.18), we conclude that equality in (3.15) holds if and only if the following is satisfied: $x_i=y_i=0$ for all i with $|\lambda_i|<\lambda$; $x_i=y_i$ for all i with $\lambda_i=\lambda$; and $x_i=-y_i$ for all i with $\lambda_i=-\lambda$.

3.2. **Long Time Behavior.** In this section, we consider a transformer model with vector field (3.1) where $V_t=I_d$ and $A_t=A$ for all $t\geqslant0$. Note that in absence of the preconditioner $V_t=A$, these dynamics may not be a Wasserstein gradient flow. Moreover, we only consider measures that admit a density with respect to the uniform measure on the sphere. To reflect this, it is convenient to consider the evolution of a density rather than evolution of the measure (3.4).

Let $\{\mu_t(x)\}_{t\geqslant0}$ be a curve of probability measures on $\mathbb{S}^{d-1}$ satisfying the continuity equation

$$\partial_t\mu_t+\mathring{\operatorname{div}}\left(\mu_t\mathcal{Y}[\mu_t]\right)=0, \tag{3.20}$$

where the vector field $\mathcal{Y}[\cdot]$ is defined for any positive measure ν on $\mathbb{S}^{d-1}$ by

$$\mathcal{Y}[\nu](x) \coloneqq \int_{\mathbb{S}^{d-1}} \mathbf{P}_x[y]\phi'\left(\langle Ax, y\rangle\right) \mathrm{d}\nu(y), \quad x \in \mathbb{S}^{d-1}, \tag{3.21}$$

where A is a $d \times d$ real symmetric matrix.

Here and in the rest of this section, ϕ' is a smooth positive function on the interval $[-\|A\|_2, \|A\|_2]$. In particular, we do not require monotonicity for ϕ' as in Section 3.1.

We often abuse notation and write $\mathcal{Y}_t = \mathcal{Y}[f_t] = \mathcal{Y}[\mu_t]$, and more generally, we liberally switch between f_t and μ_t if μ_t has density f_t. Define the C^1-norm of a continuously differentiable function h on an interval $S \subseteq \mathbb{R}$ as $\|h\|_{C^1(S)} \coloneqq \|h\|_{L^\infty(S)} + \|h'\|_{L^\infty(S)}$. The following theorem shows that, if ϕ' is close to the constant function 1 in C^1-norm on $S \coloneqq [-\|A\|_2, \|A\|_2]$, then the flow (3.20) converges to a delta mass exponentially fast. To that end, define

$$\varepsilon_\phi \coloneqq (\|A\|_2 + 2) \cdot \|\phi' - 1\|_{C^1(S)}. \tag{3.22}$$

Note that when $\varepsilon_\phi = 0$ and $A = I_d$, that is when $\phi' \equiv 1$, one recovers the Kuramoto dynamics on the sphere. Recall that R_0 measures the asymmetry of f_0 and is defined in (2.4) and also in (3.24).

Theorem 3.4. *Let $f_0 \in L^2(\mathbb{S}^{d-1})$ be a probability density on $\mathbb{S}^{d-1}$ and let $\{\mu_t(x)\}_{t\geqslant 0}$ denote the flow of probability measures where μ_t has density f_t evolving according to* (3.20)*. There exist universal constants $c_0, c_u > 0$, and two computable constants C_0, T_0 depending on $R_0, \|f_0\|_{L^2(\mathbb{S}^{d-1})}$ such that if $\varepsilon_\phi \leqslant c_u R_0^6$, then there exists an $x_\infty \in \mathbb{S}^{d-1}$ for which*

$$W_2\left(\mu_t, \delta_{x_\infty}\right) \leqslant C_0 e^{-c_0 t}, \quad \forall t \geqslant T_0.$$

3.2.1. *Main tools.* We adapt a technique developed in [DV05] to obtain quantitative convergence rates for non-convex (and non-concave) gradient flows. For Kuramoto models, that is, when $d = 2$ and $\phi' \equiv 1$, this technique was employed to derive a mean-field convergence result in [HKMP20, MP22]. Note, however, that this technique heavily depends on the form of the vector field driving the probability flow as already noted in [DV05]. In particular, the choice (3.21)—which is not a gradient flow—together with the complexity of dynamics on high-dimensional spheres brings substantial technical difficulties compared to the Kuramoto model on the circle. These difficulties manifest themselves most prominently in the proofs of Theorems 3.5 and 3.6.

For any $t \geqslant 0$, define

$$M_t \coloneqq \int_{\mathbb{S}^{d-1}} y \, \mathrm{d}\mu_t(y) \quad \text{and} \quad V_t(x) \coloneqq \mathbf{P}_x[M_t] = \int_{\mathbb{S}^{d-1}} \mathbf{P}_x[y] \, \mathrm{d}\mu_t(y). \tag{3.23}$$

Interestingly, M_t has a practical meaning: in encoder-only transformers such as BERT [DCLT19] the average token position M_t corresponds to the vector embedding called *mean-pooled embedding* of an input prompt that is often employed in further downstream tasks (classification, clustering, retrieval, etc.)

Moreover, define

$$R_t \coloneqq \|M_t\|_2, \quad U_t \coloneqq \frac{M_t}{R_t} \in \mathbb{S}^{d-1}\,, \tag{3.24}$$

and the following spherical caps with centers $\pm U \in \mathbb{S}^{d-1}$ for $\alpha \in (0, \pi/2)$,

(3.25)
$$S_\alpha^+(U) := \left\{x \in \mathbb{S}^{d-1} \mid \langle x, U\rangle \geqslant \cos\alpha\right\}, \quad S_\alpha^-(U) := \left\{x \in \mathbb{S}^{d-1} \mid \langle x, -U\rangle \geqslant \cos\alpha\right\}.$$

The spherical caps become smaller as $\alpha \to 0$. For simplicity, we also write $S_\alpha^+(t) = S_\alpha^+(U_t)$ and $S_\alpha^-(t) = S_\alpha^-(U_t)$.

Define I_t as

$$I_t := \int \|\mathcal{Y}_t(y)\|_2^2 \, \mathrm{d}\mu_t(y). \tag{3.26}$$

Direct calculation similar to Lemma 3.3 gives

$$\partial_t I_t = \iint Q_{\mu_t}(x, y) \, \mathrm{d}\mu_t(x) \, \mathrm{d}\mu_t(y), \tag{3.27}$$

where for any positive measure ν on $\mathbb{S}^{d-1}$,

(3.28)
$$\begin{aligned} Q_\nu(x, y) = {} & 2\left[\langle \mathcal{Y}[\nu](x), Ay\rangle + \langle \mathcal{Y}[\nu](y), Ax\rangle\right] \langle \mathcal{Y}[\nu](x), y\rangle \cdot \phi''(\langle Ax, y\rangle) \\ & + \left[2\langle \mathcal{Y}[\nu](x), \mathcal{Y}[\nu](y)\rangle - \langle x, y\rangle \left(\|\mathcal{Y}[\nu](x)\|_2^2 + \|\mathcal{Y}[\nu](y)\|_2^2\right)\right] \cdot \phi'(\langle Ax, y\rangle). \end{aligned}$$

We now state our two main tools.

Theorem 3.5. *If ϕ, A are such that $\varepsilon_\phi \leqslant 1/100$, where ε_ϕ is defined in (3.22), then for any $\alpha \in (0, \frac{\pi}{20})$, we have that*

$$\partial_t I_t \leqslant -I_t + 100\mu_t\left(\mathbb{S}^{d-1} \backslash S_\alpha^+(U_t)\right). \tag{3.29}$$

Theorem 3.5 holds for any measure along the flow, even for those that do not admit a density with respect to the uniform measure, but Theorem 3.6 below requires a initial density in $L^2(\mathbb{S}^{d-1})$.

Theorem 3.6. *Fix $\alpha = \pi/100$. Assume that μ_0 has density f_0 and $f_0 \in L^2(\mathbb{S}^{d-1})$. There exist two universal constant $c_u, c_1 > 0$, and two computable constants C_0, T_0 depending on $R_0, \|f_0\|_{L^2(\mathbb{S}^{d-1})}$ such that if $\varepsilon_\phi \leqslant c_u R_0^6$, it holds*

$$\mu_t\left(\mathbb{S}^{d-1} \backslash S_\alpha^+(U_t)\right) \leqslant C_0 e^{-(d-1)c_1 t}, \quad \forall t \geqslant T_0.$$

Now, we can combine Theorem 3.5 and Theorem 3.6 to prove Theorem 3.4.

3.2.2. *Proof of Theorem 3.4.* By Theorem 3.5 and Theorem 3.6, we see that for any $t \geqslant T_0$,

$$I_t + \partial_t I_t \leqslant 10^2 C_0 e^{-(d-1)c_1 t}. \tag{3.30}$$

Multiply by e^t on both sides and integrate T_0 to t to get that for any $t \geqslant T_0$,

$$I_t \leqslant I_{T_0} e^{T_0 - t} + 10^2 C_0 (t - T_0) e^{\max\{-(d-1)c_1 t, -t\}},$$

where we used the fact that for any $t \geqslant T_0$ and any $\kappa \in \mathbb{R}$,

$$\int_{T_0}^t e^{\kappa s} \, \mathrm{d}s \leqslant (t - T_0) e^{t \cdot \max\{\kappa, 0\}}.$$

We see that $I_t \to 0$ exponentially fast as $t \to +\infty$.

Also, recall that $\{\mu_t\}_t$ solves the continuity equation

$$\partial_t \mu_t(x) + \mathring{\operatorname{div}}\left(\mu_t(x) \mathcal{Y}_t(x)\right) = 0.$$

From [Vil09, Theorem 23.9], we have that for any $s \geqslant T_0$ and almost all $t \geqslant s$, the Wasserstein distance between μ_t and μ_s satisfies

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}W_2^2(\mu_t,\mu_s) = -\int_{\mathbb{S}^{d-1}}\langle\mathring{\nabla}\psi_{t\to s}(x),\mathcal{Y}_t(x)\rangle\,\mathrm{d}\mu_t(x),$$

where $\psi_{t\to s}(x)$ is a potential function associated with the Wasserstein geodesic connecting μ_t, μ_s, and $\mathring{\nabla}\,\psi_{t\to s}(x)$ satisfies that

$$\int_{\mathbb{S}^{d-1}}\|\mathring{\nabla}\psi_{t\to s}(x)\|_2^2\,\mathrm{d}\mu_t(x) = W_2^2(\mu_t,\mu_s).$$

By Cauchy-Schwarz, we see that for almost all $t \geqslant s \geqslant T_0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}W_2(\mu_t,\mu_s) \leqslant I_t^{\frac{1}{2}},$$

where the right hand side goes to 0^+ exponentially fast as we proved earlier. Hence, $\{\mu_t\}_{t\geqslant 0}$ is a Cauchy sequence in the Wasserstein space. By completeness of the Wasserstein space, there exists a probability measure $\mu_\infty \in \mathcal{P}(\mathbb{S}^{d-1})$ such that $\mu_t \to \mu_\infty$ in W_2, and μ_∞ satisfies that

$$\int_{\mathbb{S}^{d-1}}\|\mathcal{Y}_\infty(y)\|_2^2\,\mathrm{d}\mu_\infty(y) = 0.$$

By Theorem 3.6, $\mu_t\left(\mathbb{S}^{d-1}\backslash S_\alpha^+(U_t)\right) \to 0$ as $t \to +\infty$, so there is a $U_\infty \in \mathbb{S}^{d-1}$ such that $\mathrm{supp}(\mu_\infty) \subseteq S_\alpha^+(U_\infty)$, where we recall that $S_\alpha^+(U_\infty)$ is the spherical cap defined in (4.1). To conclude that $\mu_\infty = \delta_{x_\infty}$ for some $x_0 \in \mathbb{S}^{d-1}$, we use the following Lemma.

Lemma 3.7. *Let μ be a probability measure on $\mathbb{S}^{d-1}$ with support $\mathrm{supp}(\mu) \subseteq S_\alpha^+(U)$ for some $U \in \mathbb{S}^{d-1}, \alpha \in (0, \frac{\pi}{2})$ and such that*

$$\int_{\mathbb{S}^{d-1}}\|\mathcal{Y}[\mu](x)\|_2^2\,\mathrm{d}\mu(x) = 0, \tag{3.31}$$

Then $\mu = \delta_{x_0}$ for some $x_0 \in S_\alpha^+(U)$.

Proof. From (3.31), we know that

$$\mathcal{Y}[\mu](x) = \int_{\mathbb{S}^{d-1}}\mathbf{P}_x[y]\phi'\left(\langle Ax,y\rangle\right)\mathrm{d}\mu(y) = 0, \quad \forall x \in \mathrm{supp}(\mu).$$

Multiplying both sides by U we obtain the following

$$\int_{\mathbb{S}^{d-1}}(\langle y,U\rangle - \langle x,y\rangle\langle x,U\rangle)\phi'\left(\langle Ax,y\rangle\right)\mathrm{d}\mu(y) = 0, \quad \forall x \in \mathrm{supp}(\mu)\,. \tag{3.32}$$

Next, take $x = x_0$ to be any minimizer of $z \mapsto \langle z,U\rangle$ on the $\mathrm{supp}(\mu)$ so that $\langle y,U\rangle - \langle x_0,y\rangle\langle x_0,U\rangle \geqslant 0$ for any $y \in \mathrm{supp}(\mu)$. Thus, from (3.32) and $\phi' > 0$, we obtain that

$$\langle y,U\rangle - \langle x_0,y\rangle\langle x_0,U\rangle = 0, \quad \forall y \in \mathrm{supp}(\mu)\,.$$

Since $\mathrm{supp}(\mu) \subseteq S_\alpha^+(U)$, we know that $\langle x_0,U\rangle > 0$, and thus,

$$1 \leqslant \frac{\langle y,U\rangle}{\langle x_0,U\rangle} = \langle x_0,y\rangle \leqslant 1\,, \tag{3.33}$$

where in the first inequality, we use the definition of x_0. Hence, the inequalities in (3.33) are equalities, and then $\langle x_0,y\rangle = 1$ for all $y \in \mathrm{supp}(\mu)$, which implies that μ is a delta measure supported at x_0. ∎

Finally, we remark that C_0, T_0 in Theorem 3.4 and Theorem 3.6 can be sharpened as follows.

Theorem 3.8. *Fix $\alpha = \pi/100$. There exists a universal constant $c_u > 0$, and a $T_0 > 0$ such that if $\varepsilon_\phi \leqslant c_u R_0^6$, it holds*

$$\mu_t\left(\mathbb{S}^{d-1}\backslash S_\alpha^+(U_t)\right) \leqslant \|f_0\|_{L^2(\mathbb{S}^{d-1})} e^{-\frac{d-1}{16}(t-T_0)}\,, \quad \forall\ t \geqslant T_0\,.$$

Moreover, it is sufficient to take

$$T_0 \coloneqq \left[\frac{8}{R_0} \vee (d-1)\right] \cdot \left[10^{41}(d-1)R_0^{-14} + 10^{26}R_0^{-6}\log\|f_0\|^2_{L^2(\mathbb{S}^{d-1})}\right],$$

where $a \vee b \coloneqq \max\{a,b\}$ for $a, b \in \mathbb{R}$.

In Theorem 3.8, it is possible to achieve a better dependence on R_0, specifically R_0^{-2} using more involved arguments. We omit this result for the benefit of space and readability.

4. Łojasiewicz type inequality: Proof of Theorem 3.5 and Theorem 2.3

This section is mainly devoted to the proof of Theorem 3.5. The same proof together with Remark 4.2 gives the proof for Theorem 2.3.

Lemma 4.1. *Assume that ϕ, A are such that $\varepsilon_\phi \leqslant 1/100$, where ε_ϕ is defined in (3.22), and assume that a positive measure ν on $\mathbb{S}^{d-1}$ is such that there exists $U \in \mathbb{S}^{d-1}$ and $\alpha \in (0, \frac{\pi}{20})$, such that*

$$\operatorname{supp}(\nu) \subseteq S_\alpha^+(U) \coloneqq \left\{x \in \mathbb{S}^{d-1} \mid \langle x, U\rangle \geqslant \cos\alpha\right\}. \tag{4.1}$$

Then

$$\iint Q_\nu(x,y)\mathrm{d}\nu(x)\mathrm{d}\nu(y) \leqslant -\nu(\mathbb{S}^{d-1}) \int \|\mathcal{Y}[\nu](y)\|_2^2 \mathrm{d}\nu(y)\,, \tag{4.2}$$

where Q_ν is defined in (3.28). In particular, taking $\nu = \mu_t$, this implies that if $\operatorname{supp}(\mu_t) \subseteq S_\alpha^+(U)$, the following entropy production inequality holds:

$$\partial_t I_t \leqslant -I_t\,, \tag{4.3}$$

where I_t is defined in (3.26).

Remark 4.2. *To be consistent with the assumptions in Theorem 3.6, we eventually choose $\alpha = \frac{\pi}{100}$ and $\varepsilon_\phi \leqslant 1/100$ in our proof for Theorem 3.4. One can also prove the same result when $\varepsilon_\phi > /100$, but one needs to assume that α is less than a function in ε_ϕ, which goes to 0 as ε_ϕ goes to $+\infty$. For example, for the attention dynamics (1.2) where $A = \beta I_d$ and $\phi'(r) = e^r$ as in Theorem 2.3, the proof extends so long as $\tan\alpha \leqslant \frac{1}{10(1+\sqrt{\beta})}$, and β is any positive number. Also, one can replace the right-hand side of (4.2) with $-\frac{e^\beta}{10}\int_{\mathbb{S}^{d-1}} \|\mathcal{Y}[\mu](x)\|_2^2\, \mathrm{d}\mu(x)$, which is notably better when β is positive and large. Similar proofs and results in Lemma 4.1 also extend to the case when $\beta < 0$ in the dynamics (1.2).*

Proof of Lemma 4.1. Because both sides of (4.2) are homogeneous in constant multiplies of ν of degree 4, we can assume that ν is a probability measure, denoted μ for clarity, on $\mathbb{S}^{d-1}$. Also, in this proof, we simplify our notation to $\mathcal{Y} := \mathcal{Y}[\mu]$.

Take the standard orthonormal basis of $\mathbb{R}^d$ as $\{e_1, e_2, \ldots, e_d\}$. Without loss of generality, we assume that $U = e_d$. We adopt the gnomonic projection to rewrite

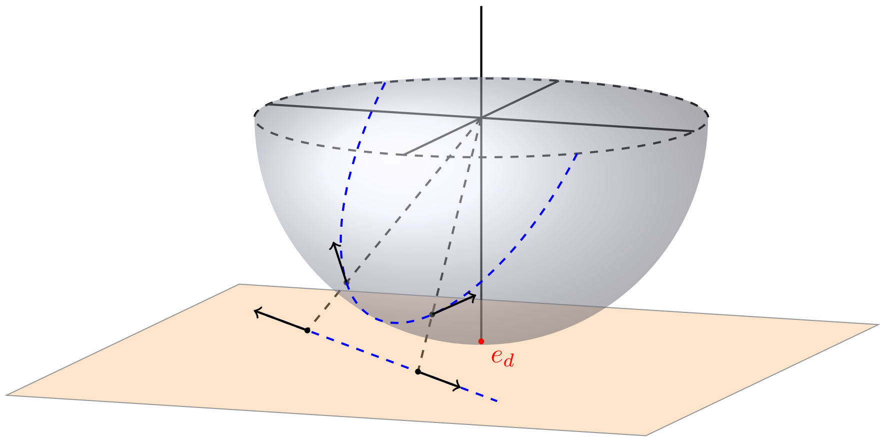

Figure 3. Illustration of gnomonic projection.

(4.2). Note that the gnomonic projection maps any geodesic (great circle) in the upper hemisphere of $\mathbb{S}^{d-1}$ to a geodesic (straight line) on the hyperplane $\mathbb{R}^{d-1}\times\{1\}$, so that the tangent vectors on $\mathbb{S}^{d-1}$ can be expressed as the difference of two points on $\mathbb{R}^{d-1}\times\{1\}$ under the inverse of the tangent map of the gnomonic projection. In particular, for any x, y in the upper hemisphere of $\mathbb{S}^{d-1}$, $\mathbf{P}_x[y]$ can be characterized by the geodesic connecting x, y, which enables us to rewrite $\mathbf{P}_x[y]$ in the definition of $\mathcal{Y}(x)$ in (3.21) in the following linear form (4.8), and gives an important equation (4.11) in this proof for Lemma 4.1. Such a property is not satisfied by stereographic projection and orthographic projection.

For an $x = (x_1, \dots, x_d)^\top \in S_\alpha^+(U) \subseteq \mathbb{S}^{d-1}$, we define

$$G(x) \coloneqq \left(\frac{x_1}{x_d}, \dots, \frac{x_{d-1}}{x_d}\right)^\top. \tag{4.4}$$

This map $G(x)$ (or the map $G(x) + e_d$, so that its image is in the hyperplane $\mathbb{R}^{d-1}\times\{1\}$), is called the gnomonic projection. G gives a diffeomorphism from $S_\alpha^+(U) \subseteq \mathbb{S}^{d-1}$ to the Euclidean ball $B_\alpha \subseteq \mathbb{R}^{d-1}$ centered at the origin and with radius $\tan\alpha$. Its inverse F is given by

$$F(u) \coloneqq \frac{1}{\sqrt{1+\|u\|_2^2}}(u + e_d), \quad \forall u \in B_\alpha. \tag{4.5}$$

Here we identify u with a vector in $\mathbb{R}^{d-1} \subseteq \mathbb{R}^d$. A direct computation shows that, the tangent map of F at u is given by

$$\mathrm{d}F_u(X) = \frac{(1+\|u\|_2^2)X - \langle X, u\rangle u - \langle X, u\rangle e_d}{\sqrt{(1+\|u\|_2^2)^3}}, \ \forall X \in T_u\mathbb{R}^{d-1} \cong \mathbb{R}^{d-1}. \tag{4.6}$$

For a $u \in B_\alpha$, we first find the preimage of $\mathcal{Y}(F(u))$ under $\mathrm{d}F_u$. By (4.5), one can first verify that, for any $v \in B_\alpha$,

$$\langle F(u), F(v)\rangle = \frac{\langle u, v\rangle + 1}{\sqrt{1+\|u\|_2^2}\sqrt{1+\|v\|_2^2}}, \tag{4.7}$$

and then by (4.6)

$$\mathbf{P}_{F(u)}[F(v)] = F(v) - \langle F(v), F(u)\rangle F(u) = \frac{\sqrt{1+\|u\|_2^2}}{\sqrt{1+\|v\|_2^2}}\,\mathrm{d}F_u(v-u). \tag{4.8}$$

Hence,

$$\mathcal{Y}(F(u)) = \mathrm{d}F_u(X(u)), \tag{4.9}$$

with

$$X(u) \coloneqq \int_{B_\alpha} (v-u)\frac{\sqrt{1+\|u\|_2^2}}{\sqrt{1+\|v\|_2^2}}\phi'\left(\langle AF(u), F(v)\rangle\right)\mathrm{d}G_\#\mu(v), \tag{4.10}$$

a vector in $T_u\mathbb{R}^{d-1} \cong \mathbb{R}^{d-1}$. Here, $G_\#\mu$ is the pushforward measure of μ induced by the gnomonic projection G. By symmetry of $u, v \in B_\alpha$ and because A is a symmetric matrix, we readily obtain the following important observation:

$$\begin{aligned} &\int_{B_\alpha} \frac{X(u)}{1+\|u\|_2^2}\,\mathrm{d}G_\#\mu(u) \\ &= \int_{B_\alpha}\int_{B_\alpha} (v-u)\frac{\phi'\left(\langle AF(u), F(v)\rangle\right)}{\sqrt{(1+\|u\|_2^2)(1+\|v\|_2^2)}}\,\mathrm{d}G_\#\mu(v)\,\mathrm{d}G_\#\mu(u) \\ &= 0. \end{aligned} \tag{4.11}$$

Next, we rewrite the left hand side of (4.2) (or $Q_\nu(x,y)$) in terms of $X(u)$'s by replacing $x, y \in S_\alpha^+(U)$ with $F(u), F(v)$ for $u, v \in B_\alpha$. By (4.6) and (4.9), we obtain the following identities:

$$\|\mathcal{Y}(F(u))\|_2^2 = \frac{\|X(u)\|_2^2}{1+\|u\|_2^2} - \frac{\langle X(u), u\rangle^2}{(1+\|u\|_2^2)^2}, \tag{4.12}$$

and

$$\begin{aligned} \langle \mathcal{Y}(F(u)), \mathcal{Y}(F(v))\rangle = {} & \frac{\langle X(u), X(v)\rangle}{\sqrt{(1+\|u\|_2^2)(1+\|v\|_2^2)}} - \frac{\langle X(u), u\rangle\langle X(v), u\rangle}{\sqrt{(1+\|u\|_2^2)^3(1+\|v\|_2^2)}} \\ & - \frac{\langle X(v), v\rangle\langle X(u), v\rangle}{\sqrt{(1+\|v\|_2^2)^3(1+\|u\|_2^2)}} + \frac{\langle X(u), u\rangle\langle X(v), v\rangle\left(\langle u, v\rangle + 1\right)}{\sqrt{(1+\|v\|_2^2)^3(1+\|u\|_2^2)^3}}. \end{aligned} \tag{4.13}$$

Before we proceed, let us first explain our main ideas. Recall that α and ε_ϕ are small parameters ($\alpha < \pi/20$, $\varepsilon_\phi < 1/100$) so terms of the form $\langle X(u), v\rangle$ are small when $v \in B_\alpha$. Hence, after the change of variables $(x,y) \mapsto (F(u), F(v))$, the leading term on the left hand side of (4.2) becomes

$$\begin{aligned} J_1 \coloneqq \int_{B_\alpha}\int_{B_\alpha} \phi'\left(\langle AF(u), F(v)\rangle\right)\Bigg[& 2\frac{\langle X(u), X(v)\rangle}{\sqrt{(1+\|u\|_2^2)(1+\|v\|_2^2)}} \\ & - \langle F(u), F(v)\rangle\left(\frac{\|X(u)\|_2^2}{1+\|u\|_2^2} + \frac{\|X(v)\|_2^2}{1+\|v\|_2^2}\right)\Bigg]\mathrm{d}G_\#\mu(u)\,\mathrm{d}G_\#\mu(v). \end{aligned}$$

We also notice that, when α is suitably small, $\|\mathcal{Y}(F(u))\|_2^2 \sim \frac{\|X(u)\|_2^2}{1+\|u\|_2^2}$. To simplify the notations in the followings, we assume that $\tan\alpha = \sqrt{\delta}$ for some $\delta \in (0,1)$ to be determined later. In the following estimates, we frequently use the fact that $\|u\|_2^2 \leqslant \tan^2\alpha = \delta$. We see that $J_1 = J_{11} + J_{12}$, where

$$J_{11} \coloneqq -\iint \phi'(\langle AF(u), F(v)\rangle)\sqrt{(1+\|u\|_2^2)(1+\|v\|_2^2)}\left\|\frac{X(u)}{1+\|u\|_2^2} - \frac{X(v)}{1+\|v\|_2^2}\right\|_2^2$$

$$J_{12} \coloneqq 2\iint \phi'(\langle AF(u), F(v)\rangle)\frac{\|u\|_2^2 - \langle u, v\rangle}{\sqrt{(1+\|u\|_2^2)(1+\|v\|_2^2)}}\frac{\|X(v)\|_2^2}{1+\|v\|_2^2}$$

and the double integrals above are over $B_\alpha \times B_\alpha$ and with respect to $G_\#\mu \otimes G_\#\mu$. Clearly,

$$\begin{aligned}
J_{11} &\leqslant -(1-\varepsilon_\phi)\int_{B_\alpha}\int_{B_\alpha}\left\|\frac{X(u)}{1+\|u\|_2^2} - \frac{X(v)}{1+\|v\|_2^2}\right\|_2^2 \mathrm{d}G_\#\mu(u)\,\mathrm{d}G_\#\mu(v) \\
&= -2(1-\varepsilon_\phi)\int_{B_\alpha}\frac{\|X(u)\|_2^2}{(1+\|u\|_2^2)^2}\,\mathrm{d}G_\#\mu(u) \\
&\leqslant -\frac{2(1-\varepsilon_\phi)}{1+\delta}\int_{B_\alpha}\frac{\|X(u)\|_2^2}{1+\|u\|_2^2}\,\mathrm{d}G_\#\mu(u),
\end{aligned}$$

where the equality is by (4.11). Also,

$$J_{12} \leqslant 4\delta(1+\varepsilon_\phi)\int_{B_\alpha}\frac{\|X(u)\|_2^2}{1+\|u\|_2^2}\,\mathrm{d}G_\#\mu(u).$$

By setting $\alpha \in (0, \frac{\pi}{20})$, so that $\delta \in (0, \tan^2\frac{\pi}{20})$, the above two displays imply that

$$J_1 \leqslant (-1.5 + 2.5\varepsilon_\phi)\int \|\mathcal{Y}[\mu]\|^2 \mathrm{d}\mu \tag{4.14}$$

which gives us a buffer to handle the remaining terms when establishing (4.2).

To control these terms, observe that

$$\iint Q_\mu(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) - J_1 = J_2 + J_3 + J_4,$$

where

$$\begin{aligned}
J_2 &\coloneqq \int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}} 2\left(\langle\mathcal{Y}(x), Ay\rangle + \langle\mathcal{Y}(y), Ax\rangle\right)\langle\mathcal{Y}(x), y\rangle \cdot \phi''(\langle Ax, y\rangle)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) \\
&\leqslant \|A\|_2 \cdot \|\phi' - 1\|_{C^1(S)}\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}} 2(\|\mathcal{Y}(x)\|_2^2 + \|\mathcal{Y}(x)\|_2\|\mathcal{Y}(y)\|_2)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) \\
&\leqslant 4\varepsilon_\phi\int_{\mathbb{S}^{d-1}}\|\mathcal{Y}(x)\|_2^2\,\mathrm{d}\mu(x),
\end{aligned}$$

and

$$\begin{aligned}
J_3 \coloneqq \int_{B_\alpha}\int_{B_\alpha} &\phi'(\langle AF(u), F(v)\rangle)\cdot\langle F(u), F(v)\rangle \\
&\cdot\left(\frac{\langle X(u), u\rangle^2}{(1+\|u\|_2^2)^2} + \frac{\langle X(v), v\rangle^2}{(1+\|v\|_2^2)^2}\right)\mathrm{d}G_\#\mu(u)\,\mathrm{d}G_\#\mu(v) \\
&\leqslant 2\delta(1+\varepsilon_\phi)\int_{B_\alpha}\frac{\|X(u)\|_2^2}{1+\|u\|_2^2}\,\mathrm{d}G_\#\mu(u),
\end{aligned}$$

and

$$
\begin{aligned}
J_4 \coloneqq 2\int_{B_\alpha}\int_{B_\alpha} \phi'\left(\langle AF(u),F(v)\rangle\right)\Bigg(&-\frac{\langle X(u),u\rangle\langle X(v),u\rangle}{\sqrt{(1+\|u\|_2^2)^3(1+\|v\|_2^2)}}\\
-\frac{\langle X(v),v\rangle\langle X(u),v\rangle}{\sqrt{(1+\|v\|_2^2)^3(1+\|u\|_2^2)}} &+ \frac{\langle X(u),u\rangle\langle X(v),v\rangle\left(\langle u,v\rangle+1\right)}{\sqrt{(1+\|v\|_2^2)^3(1+\|u\|_2^2)^3}}\Bigg)\,\mathrm{d}G_{\#}\mu(u)\,\mathrm{d}G_{\#}\mu(v)\\
\leqslant 6\delta(1+\varepsilon_\phi)\int_{B_\alpha}\frac{\|X(u)\|_2^2}{1+\|u\|_2^2}\,\mathrm{d}G_{\#}\mu(u).&
\end{aligned}
$$

Together with $\varepsilon_\phi < \frac{1}{100}, \delta \leqslant \tan^2(\frac{\pi}{20})$ and $\frac{\|X(u)\|_2^2}{1+\|u\|_2^2} \geqslant \|\mathcal{Y}(F(u))\|_2^2$ in (4.12), we get

$$
\begin{aligned}
&\int_{\mathbb{S}^{d-1}}\int_{\mathbb{S}^{d-1}} Q_\mu(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\\
&\leqslant -\left(1.5-2.5\varepsilon_\phi-4\varepsilon_\phi-8\delta(1+\varepsilon_\phi)\right)\int_{B_\alpha}\frac{\|X(u)\|_2^2}{1+\|u\|_2^2}\,\mathrm{d}G_{\#}\mu(u)\\
&\leqslant -\int_{B_\alpha}\|\mathcal{Y}(F(u))\|_2^2\,\mathrm{d}G_{\#}\mu(u) = -\int_{\mathbb{S}^{d-1}}\|\mathcal{Y}(x)\|_2^2\,\mathrm{d}\mu(x).
\end{aligned}
$$

This completes the proof of (4.2). □

Proof of Theorem 3.5. Fix $t>0$ and define the positive measures ν_1,ν_2 on $\mathbb{S}^{d-1}$ by

$$
\nu_1(\cdot)=\mu_t(\cdot\cap S_\alpha^+(t)),\quad \nu_2(\cdot)=\mu_t(\cdot\setminus S_\alpha^+(t))
$$

and let

$$
\overline{V}_1(x)=\int_{\mathbb{S}^{d-1}}\mathbf{P}_x[y]\phi'\left(\langle Ax,y\rangle\right)\mathrm{d}\nu_1(y),\quad \overline{V}_2(x)=\int_{\mathbb{S}^{d-1}}\mathbf{P}_x[y]\phi'\left(\langle Ax,y\rangle\right)\mathrm{d}\nu_2(y). \tag{4.15}
$$

We see that

$$
\mathcal{Y}_t(x)=\overline{V}_1(x)+\overline{V}_2(x).
$$

By the explicit formula (4.15), we have the estimates that $\|\mathcal{Y}_t(x)\|_2 \leqslant (1+\varepsilon_\phi)$, $\|\overline{V}_1(x)\|_2 \leqslant (1+\varepsilon_\phi)$, and $\|\mathcal{Y}_t(x)-\overline{V}_1(x)\|_2 = \|\overline{V}_2(x)\|_2 \leqslant (1+\varepsilon_\phi)\mu_t\left(\mathbb{S}^{d-1}\setminus S_\alpha^+(t)\right)$ for any $x\in\mathbb{S}^{d-1}$. These bounds imply that

$$
\begin{aligned}
&\left|\langle\mathcal{Y}_t(x),\mathcal{Y}_t(y)\rangle-\langle\overline{V}_1(x),\overline{V}_1(y)\rangle\right|\\
&=\left|\langle\overline{V}_1(x),\overline{V}_2(y)\rangle+\langle\overline{V}_2(x),\overline{V}_1(y)\rangle+\langle\overline{V}_2(x),\overline{V}_2(y)\rangle\right|\\
&\leqslant 3(1+\varepsilon_\phi)^2\mu_t\left(\mathbb{S}^{d-1}\setminus S_\alpha^+(t)\right).
\end{aligned} \tag{4.16}
$$

Hence, by (3.27), we have

$$
\begin{aligned}
\partial_t I_t \leqslant \iint\Big[&2\left(\langle\overline{V}_1(x),Ay\rangle+\langle\overline{V}_1(y),Ax\rangle\right)\langle\overline{V}_1(x),y\rangle\cdot\phi''\left(\langle Ax,y\rangle\right)\\
&+\left(2\langle\overline{V}_1(x),\overline{V}_1(y)\rangle-\langle x,y\rangle(\|\overline{V}_1(x)\|_2^2+\|\overline{V}_1(y)\|_2^2)\right)\cdot\phi'\left(\langle Ax,y\rangle\right)\Big]\,\mathrm{d}\mu_t(x)\,\mathrm{d}\mu_t(y)\\
&+24(1+\varepsilon_\phi)^3\mu_t\left(\mathbb{S}^{d-1}\setminus S_\alpha^+(t)\right).
\end{aligned}
$$

We can further split the above integral over $\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$ into integrals over $S_\alpha^+(t) \times S_\alpha^+(t)$ and $(\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}) \backslash (S_\alpha^+(t) \times S_\alpha^+(t))$, and obtain that

$$\begin{aligned}\partial_t I_t \leqslant &\int_{S_\alpha^+(t)} \int_{S_\alpha^+(t)} \Big[2 \left(\langle \overline{V}_1(x), Ay \rangle + \langle \overline{V}_1(y), Ax \rangle \right) \langle \overline{V}_1(x), y \rangle \cdot \phi'' \left(\langle Ax, y \rangle \right) \\ &+ \left(2 \langle \overline{V}_1(x), \overline{V}_1(y) \rangle - \langle x, y \rangle (\|\overline{V}_1(x)\|_2^2 + \|\overline{V}_1(y)\|_2^2) \right) \cdot \phi' \left(\langle Ax, y \rangle \right) \Big] \mathrm{d}\mu_t(x) \, \mathrm{d}\mu_t(y) \\ &+ 48(1+\varepsilon_\phi)^3 \mu_t \left(\mathbb{S}^{d-1} \backslash S_\alpha^+(t) \right).\end{aligned}$$

Together with Lemma 4.1, the above inequality yields

$$\begin{aligned}\partial_t I_t &\leqslant -\nu_1 \left(\mathbb{S}^{d-1} \right) \int_{\mathbb{S}^{d-1}} \|\overline{V}_1(x)\|_2^2 \, \mathrm{d}\nu_1(x) + 48(1+\varepsilon_\phi)^3 \mu_t \left(\mathbb{S}^{d-1} \backslash S_\alpha^+(t) \right) \\ &= -\left(1 - \nu_2 \left(\mathbb{S}^{d-1} \right) \right) \int_{\mathbb{S}^{d-1}} \|\overline{V}_1(x)\|_2^2 \, \mathrm{d}\nu_1(x) + 48(1+\varepsilon_\phi)^3 \mu_t \left(\mathbb{S}^{d-1} \backslash S_\alpha^+(t) \right) \\ &\leqslant -\int_{\mathbb{S}^{d-1}} \|\overline{V}_1(x)\|_2^2 \, \mathrm{d}\nu_1(x) + 49(1+\varepsilon_\phi)^3 \mu_t \left(\mathbb{S}^{d-1} \backslash S_\alpha^+(t) \right).\end{aligned}$$

Note that by (4.16), and the estimates that $\|\overline{V}_1(x)\|_2 \leqslant (1+\varepsilon_\phi)$ and $\nu_2 \left(\mathbb{S}^{d-1} \right) = \mu_t \left(\mathbb{S}^{d-1} \backslash S_\alpha^+(t) \right)$,

$$\begin{aligned}&\int_{\mathbb{S}^{d-1}} \|\overline{V}_1(x)\|_2^2 \, \mathrm{d}\nu_1(x) = \int_{\mathbb{S}^{d-1}} \|\overline{V}_1(x)\|_2^2 \, \mathrm{d}\mu_t(x) - \int_{\mathbb{S}^{d-1}} \|\overline{V}_1(x)\|_2^2 \, \mathrm{d}\nu_2(x) \\ &\geqslant \int_{\mathbb{S}^{d-1}} \|\mathcal{Y}_t(x)\|_2^2 \, \mathrm{d}\mu_t(x) - 3(1+\varepsilon_\phi)^2 \mu_t \left(\mathbb{S}^{d-1} \backslash S_\alpha^+(t) \right) - (1+\varepsilon_\phi)^2 \mu_t \left(\mathbb{S}^{d-1} \backslash S_\alpha^+(t) \right).\end{aligned}$$

Hence, because $49(1+\varepsilon_\phi)^3 + 4(1+\varepsilon_\phi)^2 \leqslant 100$, we obtain that

$$\partial_t I_t \leqslant -\int_{\mathbb{S}^{d-1}} \|\mathcal{Y}_t(x)\|_2^2 \, \mathrm{d}\mu_t(x) + 100 \mu_t \left(\mathbb{S}^{d-1} \backslash S_\alpha^+(t) \right).$$

This completes the proof for Theorem 3.5.

□

To conclude this section, we complete the proof of Theorem 2.3 as a corollary of Lemma 4.1 and Remark 4.2.

Proof of Theorem 2.3. Recall first that $\mathsf{E}_\beta[\delta_u] = \max_{\mu \in \mathcal{P}(\mathbb{S}^{d-1})} \mathsf{E}_\beta[\mu]$ by Proposition 2.1 (or Theorem 3.1). Let μ_t be the Wasserstein gradient flow initialized at $\mu_0 = \mu$. For any $t_1 \geqslant 0$, we define the diffeomorphisms $\{\phi_{t_1 \to t}(x)\}_{t \geqslant t_1}$ on $\mathbb{S}^{d-1}$ by solving the ODE

$$\partial_t \phi_{t_1 \to t}(x) = \mathcal{X}_{\mu_t, \beta}(\phi_{t_1 \to t}(x)), \text{ with } \phi_{t_1 \to t_1}(x) = x, \ \forall x \in \mathbb{S}^{d-1}.$$

We first show that $\operatorname{supp}(\mu_t) \subseteq S_\alpha^+(u)$ for any $t \geqslant 0$. Fix an arbitrary $t_1 \geqslant 0$, and we assume that $\operatorname{supp}(\mu_{t_1}) \subseteq S_\alpha^+(u)$. Let $x_{t_1} \in \operatorname{supp}(\mu_{t_1})$ achieve $\min_{x \in \operatorname{supp}(\mu_{t_1})} \langle x, u \rangle$, then

$$\begin{aligned}&\frac{\mathrm{d}}{\mathrm{d}t}\Big|_{t=t_1} \langle \phi_{t_1 \to t}(x_1), u \rangle = \langle \mathcal{X}_{\mu_{t_1}, \beta}(x_{t_1}), u \rangle = \int_{\mathbb{S}^{d-1}} \langle \mathbf{P}_{x_{t_1}}[y], u \rangle e^{\beta \langle x_{t_1}, y \rangle} \, \mathrm{d}\mu_{t_1}(y) \\ &= \int_{\mathbb{S}^{d-1}} (\langle y, u \rangle - \langle x_{t_1}, y \rangle \langle x_{t_1}, u \rangle) e^{\beta \langle x_{t_1}, y \rangle} \, \mathrm{d}\mu_{t_1}(y) \\ &\geqslant \int_{\mathbb{S}^{d-1}} \left(\langle x_{t_1}, u \rangle - \langle x_{t_1}, y \rangle \langle x_{t_1}, u \rangle \right) e^{\beta \langle x_{t_1}, y \rangle} \, \mathrm{d}\mu_{t_1}(y) \geqslant 0,\end{aligned}$$

where the last inequality is because $\langle x_{t_1}, u\rangle > 0$ and $1 \geqslant \langle x_{t_1}, y\rangle > 0$ when $y \in \mathrm{supp}(\mu_{t_1}) \subseteq S_\alpha^+(u)$. Hence, $\min_{x\in \mathrm{supp}(\mu_t)}\langle x, u\rangle$ is nondecreasing in t, and then $\mathrm{supp}(\mu_t) \subseteq S_\alpha^+(u)$ for any $t \geqslant 0$.

Define $I_t = \int_{\mathbb{S}^{d-1}} \|\mathcal{X}_{\mu_t,\beta}(x)\|_2^2 \,\mathrm{d}\mu_t(x)$ so that $I_t = \partial_t \mathsf{E}_\beta[\mu_t]$. Then, combine (3.27), Lemma 4.1, and Remark 4.2, together giving that $\partial_t I_t \leqslant -\frac{e^\beta}{10} I_t$. As a consequence, we see that $I_t \leqslant e^{-\frac{e^\beta}{10}t} I_0$. Using similar arguments as in the proof of Theorem 3.4 in Section 3.2 we can show that there exists $x_\infty \in S_\alpha^+(u)$, such that $W_2(\mu_t, \delta_{x_\infty}) \leqslant \int_t^{+\infty} I_r^{\frac12}\,\mathrm{d}r \leqslant 20 e^{-\beta} e^{-\frac{e^\beta}{20}t} I_0^{\frac12}$, which goes to 0 exponentially fast. Then, we integrate $\partial_t I_t \leqslant -\frac{e^\beta}{10} I_t$ from 0 to $+\infty$, and find that

$$-I_0 = I_\infty - I_0 \leqslant \frac{e^\beta}{10}\left(-\mathsf{E}_\beta[\delta_{x_\infty}] + \mathsf{E}_\beta[\mu]\right).$$

$\square$

5. Some Basic Derivatives and Estimates for the Proof of Theorem 3.6

If $\phi' \equiv 1$ in (3.21) so that $\mathcal{Y}_t = V_t$ from (3.23), then (3.20) coincides with the classical Kuramoto model. Our main strategy is to study f_t as a perturbation of the Wasserstein gradient flow driven by $V_t(x)$. In this section, we gather various perturbative results in this direction. We first define the perturbation

$$W_t(x) \coloneqq \mathcal{Y}_t(x) - V_t(x).$$

Recall that the size of this perturbation is controlled by the parameter ε_ϕ defined as

$$\varepsilon_\phi = (\|A\|_2 + 2)\cdot \|\phi' - 1\|_{C^1(S)}. \tag{5.1}$$

Observe that $\mathcal{Y}_t(x), V_t(x)$ can be viewed as vector fields defined on $\mathbb{R}^d$ although we mainly care about $x \in \mathbb{S}^{d-1}$. The following three kinds of terms appear in our arguments.

Lemma 5.1. *For any $x \in \mathbb{S}^{d-1}$, we have that*

$$\|W_t(x)\|_2 \leqslant \varepsilon_\phi, \quad \|\nabla W_t(x)\|_2 \leqslant \varepsilon_\phi$$

where ∇ is the standard gradient on $\mathbb{R}^d$. Also,

$$\left|\int_{\mathbb{S}^{d-1}} \langle \partial_t W_t(x), \mathcal{Y}_t(x)\rangle \,\mathrm{d}\mu_t(x)\right| \leqslant \varepsilon_\phi \cdot I_t.$$

Proof. Because $x, y \in \mathbb{S}^{d-1}$ in (3.3), we see that

$$\begin{aligned}\|W_t(x)\|_2 &= \|\mathcal{Y}_t(x) - V_t(x)\|_2 = \Big\|\int_{\mathbb{S}^{d-1}} \mathbf{P}_x[y](\phi'(\langle Ax, y\rangle) - 1)\,\mathrm{d}\mu_t(y)\Big\|_2 \\ &\leqslant \|\phi' - 1\|_{C^1(S)} \int_{\mathbb{S}^{d-1}} \mathrm{d}\mu_t(y) \leqslant \varepsilon_\phi.\end{aligned}$$

Similarly, we see that

$$
\begin{aligned}
&\|\nabla W_t(x)\|_2 \\
&= \Big\| \int_{\mathbb{S}^{d-1}} \big[(\phi'(\langle Ax, y\rangle) - 1)(-y \otimes x - \langle x, y\rangle \mathrm{Id}) \\
&+ \phi''(\langle Ax, y\rangle)(Ay \otimes \mathbf{P}_x[y])\big] \,\mathrm{d}\mu_t(y) \Big\|_2 \\
&\leqslant \|\phi' - 1\|_{C^1(S)} (\|A\|_2 + 2) \int_{\mathbb{S}^{d-1}} \mathrm{d}\mu_t(y) = \varepsilon_\phi.
\end{aligned}
$$

Finally, direct computations show that

$$
\begin{aligned}
\partial_t W_t(x) &= \int_{\mathbb{S}^{d-1}} \nabla_{\mathcal{Y}_t(y)} [\mathbf{P}_x[y](\phi'(\langle Ax, y\rangle) - 1)] \,\mathrm{d}\mu_t(y) \\
&= \int_{\mathbb{S}^{d-1}} \mathbf{P}_x \big[\big(\phi'(\langle Ax, y\rangle) - 1\big) \mathcal{Y}_t(y) + \phi''(\langle Ax, y\rangle) \langle Ax, \mathcal{Y}_t(y)\rangle y\big] \,\mathrm{d}\mu_t(y).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\Big| \int_{\mathbb{S}^{d-1}} \langle \partial_t W_t(x), \mathcal{Y}_t(x)\rangle \,\mathrm{d}\mu_t(x) \Big| \\
&= \Big| \iint \big[(\phi'(\langle Ax, y\rangle) - 1)\langle \mathcal{Y}_t(y), \mathcal{Y}_t(x)\rangle \\
&+ \phi''(\langle Ax, y\rangle) \langle Ax, \mathcal{Y}_t(y)\rangle \langle y, \mathcal{Y}_t(x)\rangle\big] \,\mathrm{d}\mu_t(x) \,\mathrm{d}\mu_t(y) \Big| \\
&\leqslant \frac{1}{2} \iint (1 + \|A\|_2) \|\phi' - 1\|_{C^1(S)} \big(\|\mathcal{Y}_t(y)\|_2^2 + \|\mathcal{Y}_t(x)\|_2^2\big) \,\mathrm{d}\mu_t(x) \,\mathrm{d}\mu_t(y) \\
&= \varepsilon_\phi \int_{\mathbb{S}^{d-1}} \|\mathcal{Y}_t(x)\|_2^2 \,\mathrm{d}\mu_t(x).
\end{aligned}
$$

□

Lemma 5.2. *For the derivatives of M_t and R_t, we have the following formulas:*

$$
\partial_t M_t = \int_{\mathbb{S}^{d-1}} \mathcal{Y}_t(y) \,\mathrm{d}\mu_t(y),
$$

and

$$
\partial_t(R_t^2) = 2 \int_{\mathbb{S}^{d-1}} \big(\|\mathcal{Y}_t(y)\|_2^2 - \langle \mathcal{Y}_t(y), W_t(y)\rangle\big) \,\mathrm{d}\mu_t(y).
$$

As a corollary, for any $\varepsilon_\phi > 0$ we see that

$$
I_t - \varepsilon_\phi^2 \leqslant \partial_t(R_t^2) \leqslant 3 I_t + \varepsilon_\phi^2.
$$

Proof. The equations for $\partial_t M_t$ and $\partial_t(R_t^2)$ follow from direct computations, and we can apply Lemma 5.1 to obtain the inequalities for $\partial_t(R_t^2)$. □

Lemma 5.3. *For the derivative of I_t, we also have the following formula:*

$$
\begin{aligned}
\partial_t I_t = \iint \big[&2 \langle \mathcal{Y}_t(x), \mathcal{Y}_t(y)\rangle - \langle x, y\rangle \big(\|\mathcal{Y}_t(x)\|_2^2 + \|\mathcal{Y}_t(y)\|_2^2\big) \\
&+ 2 \langle \mathcal{Y}_t(x), \partial_t W_t(x)\rangle + \nabla W_t(x) (\mathcal{Y}_t(x), \mathcal{Y}_t(x))\big] \,\mathrm{d}\mu_t(x) \,\mathrm{d}\mu_t(y).
\end{aligned} \tag{5.2}
$$

As a corollary, if $\varepsilon_\phi \in \big(0, \frac{1}{10}\big)$, we have that

$$
\partial_t I_t \geqslant -3 I_t.
$$

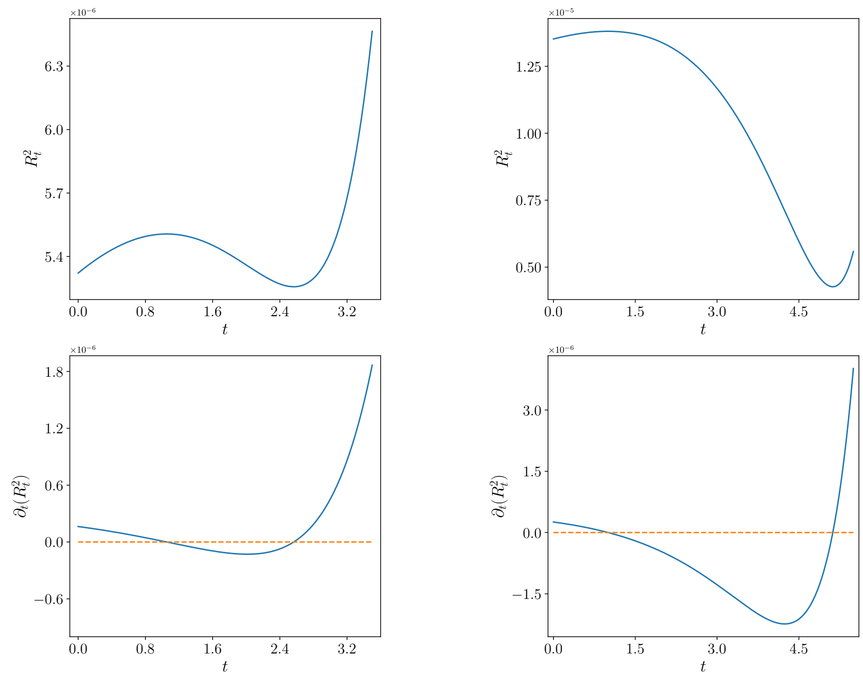

Figure 4. Examples of non-monotonic evolution of R_t. The top row shows R_t^2, and the bottom row shows $\partial_t(R_t^2)$, for two different initial profiles. Each column corresponds to a different initial profile. These plots illustrate that R_t^2 can exhibit non-monotonic behavior, with $\partial_t(R_t^2)$ taking both positive and negative values over time. $\phi(\langle Ax, y\rangle) = e^{0.1\langle x,y\rangle}$ in the plots.

In particular, for any $t_2 \geqslant t_1 \geqslant 0$,

$$I_{t_2} \geqslant I_{t_1} e^{-3(t_2-t_1)}.$$

Proof. (5.2) follows from direct computations. Similar computations also appear in the proof of Lemma 3.3, so we omit the details here. We then apply Lemma 5.1 and obtain that $\partial_t I_t \geqslant -3I_t$. □

6. Almost Kuramoto Model: Proof of Theorem 3.6

In this section, we analyze the dynamics (3.20) under the assumptions of Theorem 3.4. Before moving on to the proofs, we first explain some basic schemes. For the Kuramoto model, that is, $\phi' \equiv 1$, an important fact is that R_t is nondecreasing. We can also deduce that $\partial_t R_t \geqslant 0$ directly from Lemma 5.2. On the other hand, the form of Lemma 5.2 cannot give $\partial_t R_t \geqslant 0$ for our more general dynamics. In fact, one does not expect that $\partial_t R_t \geqslant 0$ in general, as illustrated in Figure 4. For this reason, we treat the case where $\partial_t R_t$ is large and the case where $\partial_t R_t$ is small (and even negative) separately. In Section 6.1, we show that R_t is almost increasing, in the sense that it cannot decrease by more than a factor of R_0; in Section 6.2, we show that when $\partial_t R_t$ is small, U_t is almost static, and the density around the antipodal point $-U_t$ decreases exponentially fast.

6.1. **R_t is almost increasing.**

Lemma 6.1. *Fix a constant $\lambda \in (0,1)$ and an angle $\alpha \in (0, \pi/2)$. Assume that at some time $t_1 \geqslant 0$,*

$$R_{t_1} \geqslant \lambda R_0, \quad \partial_t R_t \Big|_{t=t_1} \geqslant \frac{1}{8}(\sin^4 \alpha)\lambda^3 R_0^3.$$

If $\varepsilon_\phi \leqslant 10^{-3} \sin^2 \alpha \lambda^2 R_0^2$, then $\partial_t(R_t^2) > 0$ on $[t_1, t_1+1]$, and

$$R_{t_1+1}^2 - R_{t_1}^2 \geqslant \frac{\sin^4 \alpha}{100} \lambda^4 R_0^4.$$

Proof. Combine Lemma 5.2 and Lemma 5.3, we see that for any $t \geqslant t_1$,

$$\partial_t(R_t^2) \geqslant I_t - \varepsilon_\phi^2 \geqslant I_{t_1} e^{-3(t-t_1)} - \varepsilon_\phi^2 \geqslant \frac{1}{3} e^{-3(t-t_1)} (\partial_t(R_t^2)|_{t=t_1}) - 2\varepsilon_\phi^2.$$

So, combining this inequality and the assumptions on $R_{t_1}, \partial_t R_t|_{t=t_1}, \varepsilon_\phi$, we see that

$$\partial_t(R_t^2) \geqslant \left(\frac{1}{12} e^{-3(t-t_1)} - 5 \cdot 10^{-5} \right) (\sin^4 \alpha) \lambda^4 R_0^4,$$

which is positive when $t \in [t_1, t_1+1]$. In particular, integrate the above inequality on $[t_1, t_1+1]$, we see that

$$R_{t_1+1}^2 - R_{t_1}^2 \geqslant \frac{\sin^4 \alpha}{100} \lambda^4 R_0^4.$$

□

Lemma 6.2. *For the derivative of U_t, we have the following formula:*

$$\partial_t U_t = \frac{1}{R_t} \mathbf{P}_{U_t} [\partial_t M_t].$$

As a corollary,

$$\|\partial_t U_t\|_2 \leqslant \frac{1}{R_t} I_t^{\frac{1}{2}} \leqslant \frac{1}{R_t} \sqrt{\partial_t(R_t^2) + \varepsilon_\phi^2}.$$

Proof. Direct computations. □

We then define a smooth auxiliary function $\xi_{\alpha_1,\alpha_2}(a)$ on $\mathbb{R}$, such that $\xi_{\alpha_1,\alpha_2}(a) = 1$ when $a \geqslant \cos\alpha_1$ and $\xi_{\alpha_1,\alpha_2}(a) = 0$ when $a \leqslant \cos\alpha_2$, where $0 \leqslant \alpha_1 < \alpha_2 \leqslant \pi$. It is possible to construct such a cutoff function by mollifying indicator functions on $\mathbb{R}$. We denote the derivative of $\xi_{\alpha_1,\alpha_2}(a)$ with respect to a as $\xi'_{\alpha_1,\alpha_2}(a)$. A trivial fact is that we can also assume that $0 \leqslant \xi'_{\alpha_1,\alpha_2}(a) \leqslant 2/(\cos\alpha_1 - \cos\alpha_2)$.

Lemma 6.3. *For the derivative of the measure on the negative spherical cap, we have the following formula:*

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{S}^{d-1}} \xi_{\alpha_1,\alpha_2}(-\langle y, U_t \rangle) \, \mathrm{d}\mu_t(y) \\
&= \int_{\mathbb{S}^{d-1}} \xi'_{\alpha_1,\alpha_2}(-\langle y, U_t \rangle) (-\langle y, \partial_t U_t \rangle - \langle U_t, \mathcal{Y}_t(y) \rangle) \, \mathrm{d}\mu_t(y) \\
&\leqslant \frac{2}{\cos\alpha_1 - \cos\alpha_2} \left(\|\partial_t U_t\|_2 - R_t \sin^2 \alpha_1 + \varepsilon_\phi \right)_+ \\
&\leqslant \frac{2}{\cos\alpha_1 - \cos\alpha_2} \left(\frac{1}{R_t} \sqrt{\partial_t(R_t^2) + \varepsilon_\phi^2} - R_t \sin^2 \alpha_1 + \varepsilon_\phi \right)_+,
\end{aligned} \tag{6.1}$$

where $u_+ \coloneqq \max\{u, 0\}$ for $u \in \mathbb{R}$.

Proof. The derivative equation in (6.1) is by direct computations. For the first inequality, we notice that for those $y \in \mathbb{S}^{d-1}$ such that $-\cos\alpha_1 \leqslant \langle y, U_t\rangle \leqslant -\cos\alpha_2$, we have that

$$\begin{aligned}-\langle U_t, \mathcal{Y}_t(y)\rangle &= -\langle U_t, V_t(y)\rangle - \langle U_t, W_t(y)\rangle = -R_t\|\mathbf{P}_y[U_t]\|_2^2 - \langle U_t, W_t(y)\rangle \\ &\leqslant -R_t \sin^2\alpha_1 + \varepsilon_\phi.\end{aligned}$$

The second inequality follows from Lemma 6.2. □

Lemma 6.4. *For any $\beta \in (0, \pi/2)$, we have that*

$$\partial_t(R_t^2) \geqslant \frac{\sin^2\beta}{2} R_t^2 \left(1 - \frac{R_t + (1+\cos\beta)\mu_t\left(S_\beta^-(t)\right)}{\cos\beta}\right) - \varepsilon_\phi^2.$$

Proof. According to Lemma 5.2, we see that

$$\begin{aligned}\frac{1}{2}\partial_t R_t^2 &= \int_{\mathbb{S}^{d-1}} \left(\|V_t(y) + W_t(y)\|_2^2 - \langle V_t(y) + W_t(y), W_t(y)\rangle\right) \mathrm{d}\mu_t(y) \\ &= \int_{\mathbb{S}^{d-1}} \left(\|V_t(y)\|_2^2 + \langle V_t(y), W_t(y)\rangle\right) \mathrm{d}\mu_t(y) \\ &\geqslant \frac{1}{2}\int_{\mathbb{S}^{d-1}} \left(\|V_t(y)\|_2^2 - \|W_t(y)\|_2^2\right) \mathrm{d}\mu_t(y)\end{aligned}$$

We notice that for $y \in \mathbb{S}^{d-1}\backslash\left(S_\beta^+(t) \cup S_\beta^-(t)\right)$, we have that $\|V_t(y)\|_2^2 \geqslant R_t^2 \sin^2\beta$. Hence,

$$\partial_t R_t^2 \geqslant R_t^2 \sin^2\beta \ \mu_t\left(\mathbb{S}^{d-1}\backslash(S_\beta^+(t) \cup S_\beta^-(t))\right) - \varepsilon_\phi^2. \tag{6.2}$$

On the other hand, by the definition of R_t, we see that

$$\begin{aligned}R_t &= \int_{\mathbb{S}^{d-1}} \langle y, U_t\rangle \mathrm{d}\mu_t(y) \\ &\geqslant \cos\beta\mu_t\left(S_\beta^+(t)\right) - \cos\beta\mu_t\left(\mathbb{S}^{d-1}\backslash(S_\beta^+(t) \cup S_\beta^-(t))\right) - \mu_t\left(S_\beta^-(t)\right) \\ &= \cos\beta - 2\cos\beta\mu_t\left(\mathbb{S}^{d-1}\backslash(S_\beta^+(t) \cup S_\beta^-(t))\right) - (1+\cos\beta)\mu_t\left(S_\beta^-(t)\right),\end{aligned}$$

and so,

$$\mu_t\left(\mathbb{S}^{d-1}\backslash(S_\beta^+(t) \cup S_\beta^-(t))\right) \geqslant \frac{1}{2} - \frac{R_t + (1+\cos\beta)\mu_t\left(S_\beta^-(t)\right)}{2\cos\beta}. \tag{6.3}$$

Combine the above inequality and (6.2), we can obtain the formula in Lemma 6.4. □

Lemma 6.5. *Fix a constant $\lambda \in (1 - 10^{-3}, 1)$ and an angle $\alpha_1 \in [\pi/100, \pi/2)$. Assume that at time t_1, we have that $R_{t_1} \geqslant R_0$, and there is a time window $[t_1, t_2]$, such that when $t \in [t_1, t_2]$,*

$$\partial_t R_t \leqslant \frac{1}{8}(\sin^4\alpha_1)\lambda^3 R_0^3.$$

Then, if

$$\varepsilon_\phi \leqslant 10^{-3}(1-\lambda)\lambda^2 R_0^2,$$

we have that for any $t \in [t_1, t_2]$,

$$\partial_t(R_t^2) \geqslant -\frac{\sin^2\beta}{2\cos\beta}(R_t - \lambda R_0)\left(R_t^2 - \frac{3}{5}(1-\lambda)R_0(R_t + \lambda R_0)\right), \tag{6.4}$$

where β is an angle in $(0, \pi/2)$ such that $\sin^2\beta = \frac{1-\lambda}{5}R_0$. As a corollary, we have that for any $t \in [t_1, t_2]$,

$$R_t \geqslant \lambda R_0.$$

Proof. We use Lemma 6.4 to give a lower bound for $\partial_t R_t^2$ on $[t_1, t_2]$, for which we actually need an upper bound for $\mu_t\left(S_\beta^-(t)\right)$ when $t \in [t_1, t_2]$.

First, take $\delta > 0$ such that for $t \in [t_1, t_1 + \delta]$, we have that $R_t \geqslant \lambda R_0$. This is possible for some small $\delta > 0$ first by the fact that $R_{t_1} \geqslant R_0$ and the continuity of the ODE solution R_t. We show that we can extend the interval $[t_1, t_1 + \delta]$ a little bit longer to an interval $[t_1, t_1 + \delta + \delta']$ for some $\delta' > 0$ small, such that $R_t \geqslant \lambda R_0$ on $[t_1, t_1 + \delta + \delta']$.

Because $\varepsilon_\phi \leqslant 10^{-3}(1-\lambda)\lambda^2 R_0^2$ and $\partial_t R_t \leqslant \frac{1}{8}(\sin^4\alpha_1)\lambda^3 R_0^3$, we find that the right hand side of (6.1) for any $t \in [t_1, t_1 + \delta]$ satisfies that,

$$\begin{aligned}
&\frac{1}{R_t}\sqrt{\partial_t(R_t^2) + \varepsilon_\phi^2} - R_t \sin^2\alpha_1 + \varepsilon_\phi \leqslant \sqrt{2\frac{\partial_t R_t}{R_t}} - R_t \sin^2\alpha_1 + 2\frac{\varepsilon_\phi}{R_t} \\
&\leqslant -\frac{1}{2}\lambda R_0 \sin^2\alpha_1 + \frac{2}{10^3}(1-\lambda)\lambda R_0 \\
&\leqslant \lambda R_0\left(-\frac{1}{5000} + \frac{2}{10^3}\frac{1}{10^3}\right) < 0,
\end{aligned}$$

where in the last inequality, we used the fact that $\sin\alpha_1 \geqslant \sin\frac{\pi}{100} \geqslant \frac{\pi}{100}\cdot\frac{2}{\pi}$ and $(1-\lambda) < 10^{-3}$. Hence, by the continuity of the ODE flow again, there is a small $\delta' > 0$ such that for $t \in [t_1, t_1 + \delta + \delta']$, we have that the right hand side of (6.1) is 0, that is

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{S}^{d-1}} \xi_{\alpha_1,\alpha_2}(-\langle y, U_t\rangle)\,\mathrm{d}\mu_t(y) \leqslant 0 \tag{6.5}$$

for any $t \in [t_1, t_1 + \delta + \delta']$ and any $\alpha_2 \in (\alpha_1, \pi/2)$.

Now, for any $\beta < \alpha_1 < \alpha_2 < \pi/2$ and $t \in [t_1, t_1 + \delta + \delta']$, we have that

$$\begin{aligned}
\mu_t\left(S_\beta^-(t)\right) &\leqslant \int_{\mathbb{S}^{d-1}} \xi_{\alpha_1,\alpha_2}(-\langle y, U_t\rangle)\,\mathrm{d}\mu_t(y) \leqslant \int_{\mathbb{S}^{d-1}} \xi_{\alpha_1,\alpha_2}(-\langle y, U_{t_1}\rangle)\,\mathrm{d}\mu_{t_1}(y) \\
&\leqslant \mu_{t_1}\left(\mathbb{S}^{d-1}\backslash(S_\beta^+(t_1))\right) = \mu_{t_1}\left(S_\beta^-(t_1)\right) + \mu_{t_1}\left(\mathbb{S}^{d-1}\backslash(S_\beta^+(t_1) \cup S_\beta^-(t_1))\right).
\end{aligned}$$

On the other hand, we see that for any $s \geqslant 0$,

$$\begin{aligned}
R_s &= \int_{\mathbb{S}^{d-1}} \langle y, U_s\rangle\,\mathrm{d}\mu_s(y) \\
&\leqslant \mu_s\left(S_\beta^+(s)\right) + \cos\beta\,\mu_s\left(\mathbb{S}^{d-1}\backslash(S_\beta^+(s) \cup S_\beta^-(s))\right) - \cos\beta\,\mu_s\left(S_\beta^-(s)\right) \\
&= \left(1 - \mu_s\left(\mathbb{S}^{d-1}\backslash(S_\beta^+(s) \cup S_\beta^-(s))\right) - \mu_t\left(S_\beta^-(s)\right)\right) \\
&\quad + \cos\beta\,\mu_s\left(\mathbb{S}^{d-1}\backslash(S_\beta^+(s) \cup S_\beta^-(s))\right) - \cos\beta\,\mu_s\left(S_\beta^-(s)\right) \\
&\leqslant 1 - (1 + \cos\beta)\mu_s\left(S_\beta^-(s)\right).
\end{aligned} \tag{6.6}$$

Combine (6.2) and (6.6) for $s = t_1$, we see that when $t \in [t_1, t_1 + \delta + \delta']$,

$$\begin{aligned}\mu_t\left(S_\beta^-(t)\right) &\leqslant \mu_{t_1}\left(S_\beta^-(t_1)\right) + \mu_{t_1}\left(\mathbb{S}^{d-1}\backslash(S_\beta^+(t_1) \cup S_\beta^-(t_1))\right)\\ &\leqslant \frac{1-R_{t_1}}{1+\cos\beta} + \frac{\partial_{t_1} R_{t_1}^2 + \varepsilon_\phi^2}{R_{t_1}^2 \sin^2\beta} \leqslant \frac{1-R_{t_1}}{1+\cos\beta} + \frac{\varepsilon_\phi^2}{R_{t_1}^2\sin^2\beta}.\end{aligned}$$

Hence, Lemma 6.4 gives that when $t \in [t_1, t_1 + \delta + \delta']$,

$$\partial_t(R_t^2) \geqslant \frac{\sin^2\beta}{2\cos\beta}\left(-R_t^3 + b(t_1)R_t^2 + c(t_1)\right),$$

where

$$b(t_1) = R_{t_1} + \cos\beta - 1 - \frac{1+\cos\beta}{R_{t_1}^2\sin^2\beta}\varepsilon_\phi^2, \quad c(t_1) = -\frac{2\cos\beta}{\sin^2\beta}\varepsilon_\phi^2.$$

To proceed, we first remark that if $\varepsilon_\phi = 0$, we see that $b(t_1) = R_{t_1} + \cos\beta - 1 < R_{t_1}$ and $c(t_1) = 0$. In this case, we see that $\partial_t(R_t^2) \geqslant \frac{\sin^2\beta}{2\cos\beta}R_t^2(b(t_1) - R_t)$. The right hand side is nonnegative once R_t reaches $b(t_1)$. Hence, for $t \in [t_1, t_1 + \delta + \delta']$, we have that $R_t \geqslant b(t_1) \geqslant R_0 - \frac{1-\lambda}{5}R_0 > \lambda R_0$. This strict inequality gives a little room when $\varepsilon_\phi \neq 0$. As in the assumption, we have that $\sin^2\beta = \frac{1-\lambda}{5}R_0$, $\varepsilon_\phi \leqslant 10^{-3}(1-\lambda)\lambda^2 R_0^2$, and $R_{t_1} \geqslant R_0$. So,

$$\begin{aligned}b(t_1) &\geqslant R_0 - \frac{1-\lambda}{5}R_0 - \frac{10}{R_0^2(1-\lambda)R_0}\left(10^{-3}(1-\lambda)\lambda^2R_0^2\right)^2\\ &\geqslant R_0 - \frac{1-\lambda}{5}R_0 - \frac{1-\lambda}{5}R_0 = \lambda R_0 + \frac{3}{5}(1-\lambda)R_0,\end{aligned}$$

and

$$c(t_1) \geqslant -\frac{10}{(1-\lambda)R_0}\left(10^{-3}(1-\lambda)\lambda^2R_0^2\right)^2 \geqslant -\frac{3}{5}(1-\lambda)\lambda^2R_0^3.$$

Hence, when $t \in [t_1, t_1 + \delta + \delta']$,

$$\begin{aligned}\partial_t(R_t^2) &\geqslant \frac{\sin^2\beta}{2\cos\beta}\left(-R_t^3 + \left(\lambda R_0 + \frac{3}{5}(1-\lambda)R_0\right)R_t^2 - \frac{3}{5}(1-\lambda)\lambda^2R_0^3\right)\\ &= -\frac{\sin^2\beta}{2\cos\beta}\left(R_t - \lambda R_0\right)\left(R_t^2 - \frac{3}{5}(1-\lambda)R_0(R_t + \lambda R_0)\right),\end{aligned}$$

which is exactly (6.4). We also see that λR_0 is strictly larger than the roots of the quadratic polynomial $x^2 - \frac{3}{5}(1-\lambda)R_0(x + \lambda R_0)$ because $\lambda \in (\frac{2}{3}, 1)$. Because $R_{t_1} \geqslant R_0$, as in the argument for $\varepsilon_\phi = 0$, we see that when $t \in [t_1, t_1 + \delta + \delta']$, $R_t \geqslant \lambda R_0$.

We finally remark that the only assumption we made in the proof is that for some $\delta > 0$ and for $t \in [t_1, t_1 + \delta]$ we have that $R_t \geqslant \lambda R_0$. Under this assumption, we obtained a $\delta' > 0$ such that for $t \in [t_1, t_1 + \delta + \delta']$, we have that (6.4) holds true and also $R_t \geqslant \lambda R_0$. By taking $[t_1, t_1 + \delta]$ as the supremum interval on which $R_t \geqslant \lambda R_0$, we get that for any $t \in [t_1, t_2]$, (6.4) holds true, and $R_t \geqslant \lambda R_0$.

$\square$

Lemma 6.6. *Fix a constant* $\lambda \in (1 - 10^{-10}R_0^2, 1)$. *Then, if*

$$\varepsilon_\phi \leqslant 10^{-3}(1-\lambda)\lambda^2R_0^2,$$

we have that

$$R_t \geqslant \lambda R_0,$$

for any $t \geqslant 0$.

Proof. Fix an angle $\alpha \in [\pi/100, \pi/2)$. We use Lemma 6.1 and Lemma 6.5. We first remark that the assumption on ε_ϕ satisfies the assumptions we made in Lemma 6.1 and Lemma 6.5.

If for any $t \geqslant 0$, we have that $\partial_t R_t \leqslant \frac{1}{8}(\sin^4 \alpha)\lambda^3 R_0^3$, we see that $R_t \geqslant \lambda R_0$ for any $t \geqslant 0$ by Lemma 6.5. Otherwise, assume that $t_1 \geqslant 0$ is the first time such that $\partial_t R_t|_{t=t_1} > \frac{1}{8}(\sin^4 \alpha)\lambda^3 R_0^3$. We have that $R_t \geqslant \lambda R_0$ on $[0, t_1]$, in particular, $R_{t_1} \geqslant \lambda R_0$. By Lemma 6.1, R_t is increasing on $[t_1, t_1 + 1]$. By denoting $\eta = 1 - \lambda^2$, we have that

$$\begin{aligned}
&R_{t_1+1}^2 \geqslant R_{t_1}^2 + \frac{\sin^4 \alpha}{100}\lambda^4 R_0^4 \geqslant \lambda^2 R_0^2 + \frac{16}{10^{10}}\lambda^4 R_0^4 \\
&= R_0^2\left(1 - \eta + \frac{16}{10^{10}}(1-\eta)^2 R_0^2\right) \geqslant R_0^2\left(1 + \frac{16 R_0^2}{10^{10}} - 2\eta\right) \geqslant R_0^2\left(1 + \frac{12 R_0^2}{10^{10}}\right),
\end{aligned}$$

where we also used the fact that $\sin\alpha \geqslant \sin\frac{\pi}{100} \geqslant \frac{\pi}{100}\cdot\frac{2}{\pi}$, and the last inequality is because $\eta = 1 - \lambda^2 \leqslant 2(1-\lambda) \leqslant \frac{2R_0^2}{10^{10}}$, by the assumption of λ. Hence, we see that $R_t \geqslant \lambda R_0$ and $R_{t_1+1} \geqslant R_0$ for $t \in [0, t_1+1]$. We can run this argument again starting from $t = t_1 + 1$ and extend the interval at least by 1. So, $R_t \geqslant \lambda R_0$ for any $t \geqslant 0$. □

6.2. **Measures on Negative Caps Decrease Exponentially Fast.** In the following, we show an instability result for the negative spherical cap. Recall that μ_t has a density $f_t \in L^2(\mathbb{S}^{d-1})$.

Lemma 6.7. *For any* α_1, α_2 *with* $0 \leqslant \alpha_1 < \alpha_2 \leqslant \pi$, *we have the following formula:*

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{S}^{d-1}} \xi_{\alpha_1,\alpha_2}\left(-\langle y, U_t\rangle\right) f_t^2(y)\,\mathrm{d}y \\
&= \int_{\mathbb{S}^{d-1}} \xi'_{\alpha_1,\alpha_2}\left(-\langle y, U_t\rangle\right)\left(-\langle y, \partial_t U_t\rangle - \langle U_t, \mathcal{Y}_t(y)\rangle\right) f_t^2(y)\,\mathrm{d}y \\
&\quad + \int_{\mathbb{S}^{d-1}} \xi_{\alpha_1,\alpha_2}\left(-\langle y, U_t\rangle\right)\Big((d-1)\langle M_t, y\rangle - \mathring{\operatorname{div}}_{\mathbb{S}^{d-1}} W_t(y)\Big) f_t^2(y)\,\mathrm{d}y.
\end{aligned} \tag{6.7}$$

Also,

$$\frac{\mathrm{d}}{\mathrm{d}t}\|f_t\|_{L^2(\mathbb{S}^{d-1})}^2 \leqslant (d-1)(R_t + \varepsilon_\phi)\|f_t\|_{L^2(\mathbb{S}^{d-1})}^2. \tag{6.8}$$

Proof. (6.7) is by direct computations. For the inequality, if we let $\alpha_1 \to \pi^-$ (or just replace ξ_{α_1,α_2} with the constant function 1), we see that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\|f_t\|_{L^2(\mathbb{S}^{d-1})}^2 &= \int_{\mathbb{S}^{d-1}} \Big((d-1)\langle M_t, y\rangle - \mathring{\operatorname{div}}\, W_t(y)\Big) f_t^2(y)\,\mathrm{d}y \\
&\leqslant (d-1)(R_t + \varepsilon_\phi)\|f_t\|_{L^2(\mathbb{S}^{d-1})}^2.
\end{aligned}$$

□

Lemma 6.8. *Fix a constant $\lambda \in (2/3, 1)$ and an angle $\alpha_1 \in [\pi/100, \pi/2)$. If there is a time window $[t_1, t_2]$, such that when $t \in [t_1, t_2]$,*

$$R_t \geqslant \lambda R_0, \quad \partial_t R_t \leqslant \frac{1}{8}(\sin^4 \alpha_1)\lambda^3 R_0^3,$$

and if

$$\varepsilon_\phi \leqslant \frac{1}{10^4}\lambda^2 R_0^2 \cos\alpha_1,$$

then, we have that for any $t \in [t_1, t_2]$,

$$\frac{\mathrm{d}}{\mathrm{d}t} f_t^2\left(S_{\alpha_1}^-(t)\right) \leqslant -\frac{(d-1)\lambda R_0 \cos\alpha_1}{2} f_t^2\left(S_{\alpha_1}^-(t)\right).$$

Here, we define $f_t^2(A) = \int_A f_t^2(x)\,\mathrm{d}x$ for any measurable set $A \subseteq \mathbb{S}^{d-1}$.

Proof. By Lemma 6.7, we see that for any $\alpha_2 \in (\alpha_1, \pi/2)$, we have that

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{S}^{d-1}} \xi_{\alpha_1,\alpha_2}\left(-\langle y, U_t\rangle\right) f_t^2(y)\,\mathrm{d}y \\
&\leqslant \int_{\mathbb{S}^{d-1}} \xi'_{\alpha_1,\alpha_2}\left(-\langle y, U_t\rangle\right)\left(\|\partial_t U_t\|_2 - R_t\sin^2\alpha_1 + \varepsilon_\phi\right) f_t^2(y)\,\mathrm{d}y \\
&\quad + \int_{\mathbb{S}^{d-1}} \xi_{\alpha_1,\alpha_2}\left(-\langle y, U_t\rangle\right)\left(-(d-1)R_t\cos\alpha_2 + (d-1)\varepsilon_\phi\right) f_t^2(y)\,\mathrm{d}y.
\end{aligned}$$

Notice that we used the fact that $\langle U_t, V_t(y)\rangle = R_t\|\mathbf{P}_y[U_t]\|_2^2 \geqslant R_t\sin^2\alpha_1$ when $\cos\alpha_2 \leqslant -\langle y, U_t\rangle \leqslant \cos\alpha_1$, and the fact that $\langle M_t, y\rangle \leqslant -R_t\cos\alpha_2$ when $\cos\alpha_2 \leqslant -\langle y, U_t\rangle \leqslant 1$.

Combine Lemma 6.2, we see that, similar to the proof of Lemma 6.5, by the assumptions on $R_t, \partial_t R_t, \varepsilon_\phi$,

$$\begin{aligned}
\|\partial_t U_t\|_2 - R_t\sin^2\alpha_1 + \varepsilon_\phi &\leqslant \sqrt{2\frac{\partial_t R_t}{R_t}} - R_t\sin^2\alpha_1 + 2\frac{\varepsilon_\phi}{R_t} \\
&\leqslant \lambda R_0\left[-\frac{1}{2}\sin^2\alpha_1 + \frac{\cos\alpha_1}{5000}\right] \leqslant \lambda R_0\left[-\frac{1}{2}\left(\frac{1}{50}\right)^2 + \frac{\cos\alpha_1}{5000}\right] < 0,
\end{aligned}$$

for any $t \in [t_1, t_2]$, where we used the fact that $\sin\left(\frac{\pi}{100}\right) \geqslant \frac{2}{\pi}\cdot\frac{\pi}{100}$. Also,

$$-(d-1)R_t\cos\alpha_2 + (d-1)\varepsilon_\phi \leqslant (d-1)\lambda R_0\left(-\cos\alpha_2 + \frac{1}{10}\cos\alpha_1\right).$$

Hence, combine the above two parts, we have that for any $t \in [t_1, t_2]$ and any $\alpha_2 \in (\alpha_1, \pi/2)$,

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{S}^{d-1}} \xi_{\alpha_1,\alpha_2}\left(-\langle y, U_t\rangle\right) f_t^2(y)\,\mathrm{d}y \\
&\leqslant (d-1)\lambda R_0\left(-\cos\alpha_2 + \frac{1}{10}\cos\alpha_1\right)\int_{\mathbb{S}^{d-1}} \xi_{\alpha_1,\alpha_2}\left(-\langle y, U_t\rangle\right) f_t^2(y)\,\mathrm{d}y.
\end{aligned}$$

We can then obtain the conclusion by sending $\alpha_2 \to \alpha_1^+$. $\square$

Next, for any $t_1 \geqslant 0$, we define the diffeomorphisms $\{\phi_{t_1\to t}(x)\}_{t\geqslant t_1}$ on $\mathbb{S}^{d-1}$ by solving the ODE

$$\partial_t\phi_{t_1\to t}(x) = \mathcal{Y}_t(\phi_{t_1\to t}(x)), \text{ with } \phi_{t_1\to t_1}(x) = x, \ \forall x \in \mathbb{S}^{d-1}.$$

Lemma 6.9. *Fix a constant $\lambda \in (2/3, 1)$ and an angle $\alpha \in [\pi/100, \pi/2)$. If there is a time window $[t_1, t_2]$, such that when $t \in [t_1, t_2]$,*

$$R_t \geqslant \lambda R_0, \quad \partial_t R_t \leqslant \frac{1}{8}(\sin^4 \alpha)\lambda^3 R_0^3,$$

and if

$$\varepsilon_\phi \leqslant \frac{1}{10^3}\lambda^2 R_0^2 \sin\alpha,$$

then, for any $t_3, t_4 \in [t_1, t_2]$, $t_3 \leqslant t_4$, and $x \in \mathbb{S}^{d-1}$ such that

$$\phi_{t_3\to t_4}(x) \in \mathbb{S}^{d-1} \backslash \left(S_\alpha^-(t_4) \cup S_\alpha^+(t_4)\right),$$

we have that

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=t_4} \langle \phi_{t_3\to t}(x), U_t\rangle \geqslant \lambda R_0 \frac{\sin^2\alpha}{4}. \tag{6.9}$$

As a corollary, if we define,

$$\delta = \delta(\lambda, R_0, \alpha) := \frac{4}{\lambda R_0 \sin^2\alpha}, \tag{6.10}$$

then if $t_2 - t_1 \geqslant \delta$, we have that

$$\phi_{t_1\to t_2}\left(\mathbb{S}^{d-1}\backslash S_\alpha^-(t_1)\right) \subseteq S_\alpha^+(t_2), \quad \mathbb{S}^{d-1}\backslash S_\alpha^+(t_2) \subseteq \phi_{t_1\to t_2}\left(S_\alpha^-(t_1)\right).$$

Proof. By Lemma 6.2, and the assumptions on $R_t, \partial_t R_t, \varepsilon_\phi$, we see that

$$\begin{aligned}
&\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=t_4} \langle \phi_{t_3\to t}(x), U_t\rangle = \langle \mathcal{Y}_{t_4}\left(\phi_{t_3\to t_4}(x)\right), U_t\rangle + \langle \phi_{t_3\to t_4}(x), \partial_t U_t\rangle \\
&= \langle V_{t_4}\left(\phi_{t_3\to t_4}(x)\right), U_t\rangle + \langle W_{t_4}\left(\phi_{t_3\to t_4}(x)\right), U_t\rangle + \langle \phi_{t_3\to t_4}(x), \partial_t U_t\rangle \\
&\geqslant R_{t_4}\|\mathbf{P}_{\phi_{t_3\to t_4}(x)}[U_t]\|_2^2 - \varepsilon_\phi - \|\mathbf{P}_{\phi_{t_3\to t_4}(x)}[U_t]\|_2 \frac{1}{R_{t_4}}\left(\sqrt{\partial_t(R_t^2)|_{t=t_4}} + \varepsilon_\phi\right) \\
&\geqslant \lambda R_0 \|\mathbf{P}_{\phi_{t_3\to t_4}(x)}[U_t]\|_2 \left(\|\mathbf{P}_{\phi_{t_3\to t_4}(x)}[U_t]\|_2 - \frac{1}{2}\sin^2\alpha\right) - \frac{2}{\lambda R_0}\varepsilon_\phi \\
&\geqslant \lambda R_0(\sin\alpha)\left(\sin\alpha - \frac{1}{2}\sin^2\alpha\right) - \frac{2}{\lambda R_0}\varepsilon_\phi \\
&\geqslant \lambda R_0 \frac{\sin^2\alpha}{2} - \lambda R_0 \frac{\sin\alpha}{500} \geqslant \lambda R_0 \frac{\sin^2\alpha}{4},
\end{aligned}$$

where in the first inequality, we used the fact that $\partial_t U_t$ is in the tangent plane of U_t and $\|\mathbf{P}_{U_t}[\phi_{t_3\to t_4}(x)]\|_2 = \|\mathbf{P}_{\phi_{t_3\to t_4}(x)}[U_t]\|_2$, and in the last inequality, we used that fact that $\frac{\pi}{100}\cdot\frac{2}{\pi} \leqslant \sin\frac{\pi}{100} \leqslant \sin\alpha$. □

Theorem 6.10. *Fix a constant $\lambda \in (1 - 10^{-10}R_0^2, 1)$ and the angle $\alpha = \pi/100$. If there is a time window $[t_1, t_2]$, such that when $t \in [t_1, t_2]$,*

$$R_t \geqslant \lambda R_0, \quad \partial_t R_t \leqslant \frac{1}{8}(\sin^4\alpha)\lambda^3 R_0^3,$$

and if

$$\varepsilon_\phi \leqslant \frac{1}{10^4}\lambda^2 R_0^2 \sin\alpha\cos\alpha,$$

then there is a T of the form

$$T = C_u t_1 + C_0 \tag{6.11}$$

where C_u is a universal constant and C_0 is a constant depending on R_0 and $\|f_0\|_{L^2(\mathbb{S}^{d-1})}$, such that either $t_2 - t_1 \leqslant T$ or $t_2 = +\infty$.

Proof. Assume that $t_2 - t_1 > T > \delta(\lambda, R_0, \alpha)$, where $\delta(\lambda, R_0, \alpha)$ is defined in (6.10). For any $r \in [t_1 + \delta, t_2]$, by Lemma 6.9, we see that

$$\mu_r\left(\mathbb{S}^{d-1} \backslash S_\alpha^+(r)\right) \leqslant \mu_r\left(\phi_{r-\delta,r}\left(S_\alpha^-(r-\delta)\right)\right) = \mu_{r-\delta}\left(S_\alpha^-(r-\delta)\right).$$

By Lemma 6.8, we see that

$$f_{r-\delta}^2\left(S_\alpha^-(r-\delta)\right) \leqslant e^{-\frac{(d-1)\lambda R_0 \cos\alpha}{2}(r-\delta-t_1)} f_{t_1}^2\left(S_\alpha^-(t_1)\right).$$

Hence, by Hölder's inequality, we have that

$$\mu_r\left(\mathbb{S}^{d-1} \backslash S_\alpha^+(r)\right) \leqslant C_d \cdot e^{-\frac{(d-1)\lambda R_0 \cos\alpha}{4}(r-\delta-t_1)} \left[f_{t_1}^2\left(S_\alpha^-(t_1)\right)\right]^{\frac{1}{2}}, \tag{6.12}$$

where $C_d > 0$ is a constant depending on d (actually the square root of the surface measure of $\mathbb{S}^{d-1}$, which is less than 10 for any d). By Lemma 6.7, we see that

$$\left[f_{t_1}^2\left(S_\alpha^-(t_1)\right)\right]^{\frac{1}{2}} \leqslant \|f_{t_1}\|_{L^2(\mathbb{S}^{d-1})} \leqslant e^{(d-1)t_1} \|f_0\|_{L^2(\mathbb{S}^{d-1})}$$

Also, by the choice of δ in (6.10), and $\alpha \geqslant \frac{\pi}{100}$ we see that

$$\frac{(d-1)\lambda R_0 \cos\alpha}{4} \delta \leqslant 10^4(d-1).$$

Hence,

$$\mu_r\left(\mathbb{S}^{d-1} \backslash S_\alpha^+(r)\right) \leqslant 10 \cdot e^{-\frac{(d-1)\lambda R_0 \cos\alpha}{4} r} \cdot e^{(d-1)(2t_1 + 10^4)} \cdot \|f_0\|_{L^2(\mathbb{S}^{d-1})}. \tag{6.13}$$

On the other hand, by Theorem 3.5, we have that $I_t + \partial_t I_t \leqslant 10^2 \mu_t\left(\mathbb{S}^{d-1} \backslash S_\alpha^+(t)\right)$. Multiply e^t on both sides and integrate from t_1 to r, we obtain that

$$I_r \leqslant I_{t_1} e^{-r+t_1} + \frac{10^3 \cdot e^{(d-1)(3t_1+10^4)} \cdot \|f_0\|_{L^2(\mathbb{S}^{d-1})}}{1 - \frac{(d-1)\lambda R_0 \cos\alpha}{4}} e^{-\frac{(d-1)\lambda R_0 \cos\alpha}{4} r}.$$

We notice that, if $\frac{\lambda R_0 \cos\alpha}{4} r$ is much larger than $3t_1 + 10^4$, the right hand side of the above inequality can be very small, and hence I_r is quantitatively small. Also, notice that, by Lemma 5.2 and the assumption that $R_t \geqslant \lambda R_0$, we have that for any $r \in [t_1 + \delta, t_2]$,

$$\partial_t R_t\Big|_{t=r} \leqslant \frac{3I_r}{2\lambda R_0} + \frac{\varepsilon_\phi^2}{2\lambda R_0} \leqslant \frac{3I_r}{2\lambda R_0} + 10^{-8}\lambda^3 R_0^3 \sin^2\alpha.$$

Combine the above two inequalities, and the fact that $I_t \leqslant 2$ by its definition, we can simplify the above expression by writing

$$\partial_t R_t\Big|_{t=r} \leqslant \frac{C_0}{\lambda R_0} e^{-\frac{(d-1)}{8}\left(r - 30t_1 - 10^5\right)} + 10^{-8}\lambda^3 R_0^3 \sin^2\alpha,$$

where C_0 is a constant depending on R_0 and $\|f_0\|_{L^2(\mathbb{S}^{d-1})}$. Hence, for

$$T = \frac{8}{d-1} \log\left(\frac{C_1}{R_0^4}\right) + 30t_1 + 10^5,$$

where C_1 is a constant depending on R_0 and $\|f_0\|_{L^2(\mathbb{S}^{d-1})}$, we see that if $t_2 - t_1 \geqslant T$, then $\partial_t R_t$ is upper-bounded by a positive function which is smaller than $\frac{1}{8}(\sin^4\alpha)\lambda^3 R_0^3$ for any $t \geqslant t_1 + T$, and we can repeat the above arguments until $t = +\infty$. □

Proof of Theorem 3.6. Recall that $\alpha = \frac{\pi}{100}$ now by our assumption. Fix the $\lambda = 1 - 10^{-10}R_0^4 \geqslant 1 - 10^{-10}R_0^2$. We divide $\mathbb{R}_{\geqslant 0}$ into pieces $0 = s_{-1} \leqslant t_0 < s_0 \leqslant t_1 < s_1 \leqslant t_2 < s_2 \cdots$, where for any $k \geqslant 0$

$$t_k \coloneqq \inf \left\{ t \geqslant s_{k-1} \;\middle|\; \partial_t R_t \geqslant \frac{1}{8}(\sin^4 \alpha)\lambda^3 R_0^3 \right\}, \quad s_k \coloneqq t_k + 1.$$

We first show that this construction must stop at some k_*-th step. By Lemma 6.6, we have that $R_t \geqslant \lambda R_0$ for all $t \geqslant 0$. Actually, by Lemma 6.1 and the proof of Lemma 6.6, we see that for any $k > 0$,

$$R_{s_k}^2 \geqslant R_{t_k}^2 + \frac{\sin^4 \alpha}{100}\lambda^4 R_0^4 \geqslant \lambda^2 R_0^2 + \frac{\sin^4 \alpha}{100}\lambda^4 R_0^4 \geqslant R_0^2 \left(1 + \frac{12R_0^2}{10^{10}}\right).$$

Hence, in Lemma 6.5, if we replace R_0 with $R_{s_{k-1}}$, we see that on $[s_{k-1}, t_k]$, $\partial_t R_t \leqslant \frac{1}{8}(\sin^4 \alpha)\lambda^3 R_0^3 \leqslant \frac{1}{8}(\sin^4 \alpha)\lambda^3 R_{s_{k-1}}^3$, and the assumption of Lemma 6.5 is satisfied, and hence $R_{t_k} \geqslant \lambda R_{s_{k-1}}$. Hence, we can use Lemma 6.1 and the proof of Lemma 6.6 again, and see that for any $k > 0$,

$$\begin{aligned} R_{s_k}^2 &\geqslant R_{t_k}^2 + \frac{\sin^4 \alpha}{100}\lambda^4 R_0^4 \geqslant \lambda^2 R_{s_{k-1}}^2 + \frac{16}{10^{10}}\lambda^4 R_0^4 \\ &\geqslant R_{s_{k-1}}^2 + \lambda^2 - 1 + \frac{16}{10^{10}}\lambda^4 R_0^4 \geqslant R_{s_{k-1}}^2 + \frac{12}{10^{10}} R_0^6. \end{aligned}$$

Because $R_{s_k}^2 \leqslant 1$, we must have that $k_* \leqslant 10^9 R_0^{-6}$. Together with Theorem 6.10, we see that for any $k \in [\![0, k_*]\!]$, $t_k - s_{k-1} \leqslant C_u s_{k-1} + C_0$, with C_u and C_0 obtained in Theorem 6.10, and $t_{k_*+1} = +\infty$. Hence, we may set $\widetilde{T}_0 = s_{k_*}$, which depends on R_0 and $\|f_0\|_{L^2(\mathbb{S}^{d-1})}$. By (6.13) and its proof, the following inequality holds true

$$\mu_t \left(\mathbb{S}^{d-1} \backslash S_\alpha^+(U_t)\right) \leqslant \widetilde{C}_0 e^{-(d-1)\tilde{c}_1 R_0 t}, \quad \forall t \geqslant \widetilde{T}_0\,, \tag{6.14}$$

where $\widetilde{C}_0$ is a constant depending on R_0 and $\|f_0\|_{L^2(\mathbb{S}^{d-1})}$, and $\tilde{c}_1$ is a universal constant. If $R_0 > \frac{1}{2}$, the desired form of result in Theorem 3.6 follows directly by identifying $T_0 = \widetilde{T}_0$, $C_0 = \widetilde{C}_0$ and $c_1 = \tilde{c}_1$. Otherwise, we define $\widetilde{T}_1 = \widetilde{T}_0 \vee \frac{\log(10\widetilde{C}_0)}{(d-1)\tilde{c}_1 R_0}$ such that the mass outside the cap $S_\alpha^+(U_{\widetilde{T}_1})$ is small: $\mu_{\widetilde{T}_1}(\mathbb{S}^{d-1} \backslash S_\alpha^+(U_{\widetilde{T}_1})) \leqslant \frac{1}{10}$. At time $\widetilde{T}_1$, we estimate the lower bound for $R_{\widetilde{T}_1}$ as follows:

$$\begin{aligned} R_{\widetilde{T}_1} &= \int_{\mathbb{S}^{d-1}} \langle y, U_{\widetilde{T}_1} \rangle \, \mathrm{d}\mu_{\widetilde{T}_1}(y) \\ &\geqslant \cos\alpha \; \mu_{\widetilde{T}_1}\left(S_\alpha^+(\widetilde{T}_1)\right) - \mu_{\widetilde{T}_1}\left(\mathbb{S}^{d-1} \backslash S_\alpha^+(\widetilde{T}_1)\right) \\ &= \cos\alpha - (1 + \cos\alpha)\; \mu_{\widetilde{T}_1}\left(\mathbb{S}^{d-1} \backslash S_\alpha^+(\widetilde{T}_1)\right) \\ &\geqslant \frac{9}{10}\cos\alpha - \frac{1}{10} \geqslant \frac{1}{2}. \end{aligned} \tag{6.15}$$

Thus, we can reset the starting time of the estimate (6.14) to $\widetilde{T}_1$. There exists $\widetilde{T}_2 \geqslant 0$, depending on $\|f_{\widetilde{T}_1}\|_{L^2(\mathbb{S}^{d-1})}$, and hence on $\|f_0\|_{L^2(\mathbb{S}^{d-1})}$ and R_0 via (6.8), such that

$$\mu_t \left(\mathbb{S}^{d-1} \backslash S_\alpha^+(U_t)\right) \leqslant \widetilde{C}_0 e^{-\frac{1}{2}(d-1)\tilde{c}_1 (t - \widetilde{T}_1)}, \quad \forall t \geqslant \widetilde{T}_2 + \widetilde{T}_1\,.$$

The desired result in Theorem 3.6 then follows by identifying $T_0 = \widetilde{T}_2 + \widetilde{T}_1$, $C_0 = \widetilde{C}_0 e^{\frac{1}{2}(d-1)\tilde{c}_1 \widetilde{T}_1}$ and $c_1 = \tilde{c}_1/2$. $\square$

7. More accurate C_0, T_0 in Theorem 3.4: Proof of Theorem 3.8

The assumptions in this section are the same as in Theorem 3.4, that is, we have a family of $L^2(\mathbb{S}^{d-1})$ probability densities, $\{f_t(x)\}_{t\geqslant 0}$ satisfying (3.20), and $\varepsilon_\phi \leqslant c_u R_0^6$. We notice that C_0, T_0 in Theorem 3.4 (and Theorem 3.6) comes from the proof for Theorem 6.10. In particular, the term T in (6.11) of Theorem 6.10 depends on t_1. Because of this dependence, when t is in the interval where $\partial_t R_t \geqslant \frac{1}{8}(\sin^4\alpha)\lambda^3 R_0^3$, we lose control of the growth of $f_t^2\left(\mathbb{S}^{d-1}\backslash S_\alpha^+(t)\right)$. The following arguments are mainly for fixing this issue. For this purpose, we carefully investigate the characteristic flow associated with the dynamics (3.20). Our analysis is inspired by the problems for the Kuramoto model considered in [HKMP20, MP22] where $d=2$ and $\beta=0$, but our more general dynamics (3.20) and the geometry of $\mathbb{S}^{d-1}$ make the arguments more involved than the circle case.

We adopt the diffeomorphism notation we used in proving Lemma 6.9. That is, for any $t_1 \geqslant 0$, we define the diffeomorphisms $\{\phi_{t_1\to t}(x)\}_{t\geqslant t_1}$ on $\mathbb{S}^{d-1}$ by solving the ODE

$$\partial_t \phi_{t_1\to t}(x) = \mathcal{Y}_t(\phi_{t_1\to t}(x)), \text{ with } \phi_{t_1\to t_1}(x) = x, \ \forall x \in \mathbb{S}^{d-1}.$$

After exploiting more properties of $\phi_{t_1\to t}$, we will be able to modify our Theorem 6.10.

Lemma 7.1. *For any $t \geqslant t_1 \geqslant 0$, and any measurable set $B \subseteq \mathbb{S}^{d-1}$, we have that*

$$\frac{\mathrm{d}}{\mathrm{d}t} f_t^2\left(\phi_{t_1\to t}(B)\right) \leqslant 2(d-1)\cdot f_t^2\left(\phi_{t_1\to t}(B)\right).$$

Proof. For simplicity, we prove the case where $\phi_{t_1\to t}(B)$ has a smooth topological boundary $\partial\phi_{t_1\to t}(B)$ in $\mathbb{S}^{d-1}$. The general cases for B can be done using the area formula (change of variables), and similar computations were used to prove Lemma A.1 in [HHL25].

When $\partial\phi_{t_1\to t}(B)$ is smooth, we notice that

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} f_t^2\left(\phi_{t_1\to t}(B)\right) = \frac{\mathrm{d}}{\mathrm{d}t}\int_{\phi_{t_1\to t}(B)} f_t^2(x)\,\mathrm{d}x \\
&= \int_{\partial\phi_{t_1\to t}(B)} f_t^2(x)\langle \nu(x), \partial_t\phi_{t_1\to t}(\phi_{t_1\to t}^{-1}(x))\rangle\,\mathrm{d}\mathcal{H}^{n-2}(x) + \int_{\phi_{t_1\to t}(B)} \partial_t(f_t^2(x))\,\mathrm{d}x \\
&= \int_{\phi_{t_1\to t}(B)} \mathring{\operatorname{div}}\left(f_t^2(x)\mathcal{Y}_t(x)\right)\mathrm{d}x + \int_{\phi_{t_1\to t}(B)} \partial_t(f_t^2(x))\,\mathrm{d}x.
\end{aligned}$$

where $n(x)$ is the outer unit normal vector of $\partial\phi_{t_1\to t}(B)$ in $\mathbb{S}^{d-1}$, and $\mathcal{H}^{n-2}(x)$ is the Hausdorff measure. In the last line, we used the divergence theorem and $\partial_t\phi_{t_1\to t}(\phi_{t_1\to t}^{-1}(x)) = \mathcal{Y}_t(x)$. Because f_t satisfies (3.20), we have that

$$\frac{\mathrm{d}}{\mathrm{d}t} f_t^2\left(\phi_{t_1\to t}(B)\right) = -\int_{\phi_{t_1\to t}(B)} f_t^2(x)\,\mathring{\operatorname{div}}\left(\mathcal{Y}_t(x)\right)\mathrm{d}x.$$

By (3.21), we see that

$$\begin{aligned}
-\mathring{\operatorname{div}}\left(\mathcal{Y}_t(x)\right) &= \int_{\mathbb{S}^{d-1}} \Big[(d-1)\langle x, y\rangle - \mathring{\operatorname{div}}\left(W_t(x)\right)\Big] f_t(y)\,\mathrm{d}y \\
&\leqslant (d-1)(\langle x, M_t\rangle + \varepsilon_\phi).
\end{aligned}$$

Combine the above inequalities and the fact that $\langle x, M_t\rangle \leqslant 1$, we get Lemma 7.1. $\square$

In the following, we use $\mathrm{Conv}[S]$ to denote the geodesically convex hull of a set $S \subseteq \mathbb{S}^{d-1}$, that is, the intersection of all those closed geodesically convex subsets of $\mathbb{S}^{d-1}$ that contain S. The definition implies that $\mathrm{Conv}(S)$ is unique and closed. We first need the following geometric fact:

Lemma 7.2. *Let $S \subseteq \mathbb{S}^{d-1}$ be a closed subset. If $\inf_{x,y\in S}\langle x,y\rangle > \frac{\sqrt{2}}{2}$, then, $\inf_{x,y\in S}\langle x,y\rangle = \inf_{x,y\in \mathrm{Conv}[S]}\langle x,y\rangle$.*

Proof. Because $S \subseteq \mathrm{Conv}[S]$, we have that $\inf_{x,y\in S}\langle x,y\rangle \geqslant \inf_{x,y\in \mathrm{Conv}[S]}\langle x,y\rangle$. Assume that $\inf_{x,y\in S}\langle x,y\rangle > \inf_{x,y\in \mathrm{Conv}[S]}\langle x,y\rangle$, and $\inf_{x,y\in \mathrm{Conv}[S]}\langle x,y\rangle$ is achieved by $Z_1, Z_2 \in \mathrm{Conv}[S]$, and we use θ to denote the angle between Z_1, Z_2, that is, $\cos\theta = \langle Z_1, Z_2\rangle$ and $\theta \in [0, \frac{\pi}{2})$. Notice that we can get $\theta < \frac{\pi}{2}$, because the maximal angle of points in S does not exceed $\frac{\pi}{4}$. Assume that $Z_1 \notin S$, then we consider the spherical cap $S^+_\theta(Z_2)$, where we used the definition (4.1). Because $\cos\theta = \inf_{x,y\in \mathrm{Conv}[S]}\langle x,y\rangle$, we have that $\mathrm{Conv}[S] \subseteq S^+_\theta(Z_2)$. Hence, $S \subseteq S^+_\theta(Z_2)$. We extend the geodesic from Z_1 to Z_2 further to a point Z_3, such that $\langle Z_1, Z_3\rangle = 0$. Recall the triangle inequality on sphere: for any X_1, X_2, X_3 such that $X_1, X_3 \in S^+_{\frac{\pi}{2}}(X_2)$, we have that $\theta_{21} + \theta_{23} \geqslant \theta_{13}$, where θ_{21} is the angle between X_2 and X_1, θ_{23} is the angle between X_2 and X_3, and θ_{13} is the angle between X_1 and X_3. Hence, $S^+_\theta(Z_2) \subseteq S^+_{\frac{\pi}{2}}(Z_3)$, and the boundaries of these two sets are only tangent at Z_1. Apparently, $S \subseteq \mathrm{Conv}[S] \subseteq S^+_\theta(Z_2) \subseteq S^+_{\frac{\pi}{2}}(Z_3)$. Because $Z_1 \notin S$, and the boundaries of $S^+_\theta(Z_2)$ and $S^+_{\frac{\pi}{2}}(Z_3)$ only intersect at Z_1, we see that the boundary of $S^+_{\frac{\pi}{2}}(Z_3)$ does not contain any point in S. Hence, we can take a very $\epsilon > 0$, such that $S \subseteq S^+_{\frac{\pi}{2}-\epsilon}(Z_3)$. Because $S^+_{\frac{\pi}{2}-\epsilon}(Z_3)$ is also a closed geodesically convex set, by the definition of $\mathrm{Conv}[S]$, we must have that $\mathrm{Conv}[S] \subseteq S^+_{\frac{\pi}{2}-\epsilon}(Z_3)$. This is a contradiction, because $Z_1 \in \mathrm{Conv}[S]$ but $Z_1 \notin S^+_{\frac{\pi}{2}-\epsilon}(Z_3)$. ☐

Lemma 7.3. *Fix a time $t_1 \geqslant 0$ and a closed subset $B \subseteq \mathbb{S}^{d-1}$ which is properly contained in a hemisphere of $\mathbb{S}^{d-1}$. Define*

$$D_t(B) := \inf_{x,y\in \mathrm{Conv}[\phi_{t_1\to t}(B)]}\langle x,y\rangle,$$

and $\Gamma(B) := \mu_{t_1}(B)(1 + D_{t_1}(B)) - 1$. If $\Gamma(B) > 0$, $D_{t_1}(B) > 0$, and if we have that $\varepsilon_\phi^2 \leqslant \frac{1}{4}(1 - D_{t_1}(B))\Gamma(B)^2$, then for any $t \geqslant t_1$, we have that $D_t(B) \geqslant D_{t_1}(B)$, and

$$\inf_{x\in \mathrm{Conv}[\phi_{t_1\to t}(B)]}\langle x, M_t\rangle \geqslant \mu_{t_1}(B)\left(1 + D_t(B)\right) - 1 \geqslant \Gamma(B), \tag{7.1}$$

and

$$1 - D_t(B) \leqslant \max\left\{(1 - D_{t_1}(B))e^{\frac{-\Gamma(B)}{4}(t-t_1)},\ \frac{4}{\Gamma(B)^2}\varepsilon_\phi^2\right\}. \tag{7.2}$$

Proof. The first inequality (7.1) basically follows from the fact that for any set $A \subset \mathbb{S}^{d-1}$, $\mu_t(\phi_{t_1\to t}(A))$ is a constant because f_t satisfies (3.20) and $\phi_{t_1\to t}$ is its characteristic flow. More precisely,

$$
\begin{aligned}
&\inf_{x\in \mathrm{Conv}[\phi_{t_1\to t}(B)]}\langle x, M_t\rangle = \inf_{x\in \mathrm{Conv}[\phi_{t_1\to t}(B)]}\int_{\mathbb{S}^{d-1}}\langle x,y\rangle f_t(y)\,\mathrm{d}y\\
&\geqslant \inf_{x\in \mathrm{Conv}[\phi_{t_1\to t}(B)]}\int_{\mathrm{Conv}[\phi_{t_1\to t}(B)]}\langle x,y\rangle f_t(y)\,\mathrm{d}y - \int_{\mathbb{S}^{d-1}\setminus \mathrm{Conv}[\phi_{t_1,t}(B)]} f_t(y)\,\mathrm{d}y\\
&\geqslant D_t(B)\mu_t\left(\mathrm{Conv}[\phi_{t_1\to t}(B)]\right) - \left(1-f_t\left(\mathrm{Conv}[\phi_{t_1\to t}(B)]\right)\right)\\
&= f_t\left(\mathrm{Conv}[\phi_{t_1\to t}(B)]\right)(1+D_t(B))-1\\
&\geqslant \mu_t\left(\phi_{t_1\to t}(B)\right)(1+D_t(B))-1 = \mu_{t_1}(B)\,(1+D_t(B))-1.
\end{aligned}
$$

To prove the second inequality (7.2), we need to compute the derivatives of $D_t(B)$ in t. Let $\phi_{t_1\to t}(x), \phi_{t_1\to t}(y)$ be two points in $\mathrm{Conv}[\phi_{t_1\to t}(B)]$, we have that

(7.3)

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\langle \phi_{t_1\to t}(x), \phi_{t_1\to t}(y)\rangle = \langle \mathcal{Y}_t(\phi_{t_1\to t}(x)), \phi_{t_1\to t}(y)\rangle + \langle \phi_{t_1\to t}(x), \mathcal{Y}_t(\phi_{t_1\to t}(y))\rangle\\
&= \langle V_t(\phi_{t_1\to t}(x)), \phi_{t_1\to t}(y)\rangle + \langle \phi_{t_1\to t}(x), V_t(\phi_{t_1\to t}(y))\rangle\\
&\quad + \langle W_t(\phi_{t_1\to t}(x)), \phi_{t_1\to t}(y)\rangle + \langle \phi_{t_1\to t}(x), W_t(\phi_{t_1\to t}(y))\rangle\\
&= \Big\langle M_t, \mathbf{P}_{\phi_{t_1\to t}(x)}[\phi_{t_1\to t}(y)]\Big\rangle + \Big\langle M_t, \mathbf{P}_{\phi_{t_1\to t}(y)}[\phi_{t_1\to t}(x)]\Big\rangle\\
&\quad + \Big\langle W_t(\phi_{t_1\to t}(x)), \mathbf{P}_{\phi_{t_1\to t}(x)}[\phi_{t_1\to t}(y)]\Big\rangle + \Big\langle \mathbf{P}_{\phi_{t_1\to t}(y)}[\phi_{t_1\to t}(x)], W_t(\phi_{t_1\to t}(y))\Big\rangle.
\end{aligned}
$$

If we let $\theta\in[0,\frac{\pi}{2}]$ such that $\cos(2\theta) = \langle \phi_{t_1\to t}(x), \phi_{t_1\to t}(y)\rangle$, then we see that $\|\mathbf{P}_{\phi_{t_1\to t}(x)}[\phi_{t_1\to t}(y)]\|_2 = \sin(2\theta)$. Also, there is a $Z\in \mathrm{Conv}[\phi_{t_1\to t}(B)]\subseteq \mathbb{S}^{d-1}$ such that $\mathbf{P}_{\phi_{t_1\to t}(x)}[\phi_{t_1\to t}(y)] + \mathbf{P}_{\phi_{t_1\to t}(y)}[\phi_{t_1\to t}(x)] = 2\sin(\theta)\sin(2\theta)\cdot Z$. This Z is actually the middle point on the shortest great circle connecting $\phi_{t_1\to t}(x), \phi_{t_1\to t}(y)$. Hence,

$$
\frac{\mathrm{d}}{\mathrm{d}t}\langle \phi_{t_1\to t}(x), \phi_{t_1\to t}(y)\rangle \geqslant \inf_{z\in \mathrm{Conv}[\phi_{t_1\to t}(B)]}\langle z, M_t\rangle - 2\varepsilon_\phi\sin(2\theta),
$$

where we also used Lemma 5.1. Because $\phi_{t_1\to t}(x), \phi_{t_1\to t}(y)$ were chosen arbitrarily, by writing $\sin(2\theta) = \sqrt{1-\cos^2(2\theta)}$ and $\sin(\theta) = \sqrt{\frac{1-\cos(2\theta)}{2}}$, we obtain that

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}D_t(B) &\geqslant 2\sqrt{1-D_t(B)^2}\left(\sqrt{\frac{1-D_t(B)}{2}}\inf_{z\in \mathrm{Conv}[\phi_{t_1\to t}(B)]}\langle z, M_t\rangle - \varepsilon_\phi\right)\\
&\geqslant 2\sqrt{1-D_t(B)^2}\left(\sqrt{\frac{1-D_t(B)}{2}}\left[\mu_{t_1}(B)\,(1+D_t(B))-1\right] - \varepsilon_\phi\right),
\end{aligned}
\tag{7.4}
$$

where in the last step, we used the first inequality (7.1) which we just proved. Assume that $[t_1,t_2]$ is the maximal interval such that (7.2) holds true for any $t\in[t_1,t_2]$, then we want to show that $t_2 = +\infty$. First, because we have (7.2) on $[t_1,t_2]$, we obtain that $D_t(B)\geqslant D_{t_1}(B)>0$ for any $t\in[t_1,t_2]$, where we also used the assumption that $\varepsilon_\phi^2\leqslant \frac{1}{4}(1-D_{t_1}(B))\Gamma(B)^2$. If t_2 is a finite number, then we let

$t = t_2$ in (7.4), and obtain that

$$\begin{aligned}
&\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=t_2}(1-D_t(B)) \leqslant -2\sqrt{1-D_{t_2}(B)^2}\left(\sqrt{\frac{1-D_{t_2}(B)}{2}}\left[\mu_{t_1}(B)\left(1+D_{t_2}(B)\right)-1\right]-\varepsilon_\phi\right) \\
&\leqslant -2\sqrt{1-D_{t_2}(B)^2}\left(\sqrt{\frac{1-D_{t_2}(B)}{2}}\left[\mu_{t_1}(B)\left(1+D_{t_1}(B)\right)-1\right]-\varepsilon_\phi\right) \\
&= -2\sqrt{1-D_{t_2}(B)^2}\left(\sqrt{\frac{1-D_{t_2}(B)}{2}}\Gamma(B)-\varepsilon_\phi\right).
\end{aligned}$$

Now, by the assumption of t_2, we also have that

$$1-D_{t_2}(B) = \max\left\{(1-D_{t_1}(B))e^{\frac{-\Gamma(B)}{4}(t_2-t_1)},\ \frac{4}{\Gamma(B)^2}\varepsilon_\phi^2\right\} \geqslant \frac{4}{\Gamma(B)^2}\varepsilon_\phi^2.$$

Hence,

$$\begin{aligned}
&\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=t_2}(1-D_t(B)) \leqslant -\sqrt{1+D_{t_2}(B)}\cdot[1-D_{t_2}(B)]\Gamma(B)\left(\sqrt{2}-1\right) \\
&\leqslant -\frac{2}{5}[1-D_{t_2}(B)]\Gamma(B).
\end{aligned}$$

Because $D_{t_2}(B) < 1$ as B is an open set in $\mathbb{S}^{d-1}$, by the continuity of the solution, there is a small time interval $[t_2, t_2+\delta]$ for some $\delta > 0$, such that for $t \in [t_2, t_2+\delta]$, we have that

$$\frac{\mathrm{d}}{\mathrm{d}t}(1-D_t(B)) < -\frac{1}{4}[1-D_t(B)]\Gamma(B).$$

Hence, for any $t \in [t_2, t_2+\delta]$, (7.2) also holds true, because

$$\begin{aligned}
&1-D_t(B) < (1-D_{t_2}(B))e^{-\frac{\Gamma(B)}{4}(t-t_2)} \\
&\leqslant \begin{cases} (1-D_{t_1}(B))e^{\frac{-\Gamma(B)}{4}(t-t_1)}, \text{ if } 1-D_{t_2}(B) = (1-D_{t_1}(B))e^{\frac{-\Gamma(B)}{4}(t_2-t_1)}, \\ \frac{4}{\Gamma(B)^2}\varepsilon_\phi^2, \text{ if } 1-D_{t_2}(B) = \frac{4}{\Gamma(B)^2}\varepsilon_\phi^2, \end{cases} \\
&\leqslant \max\left\{(1-D_{t_1}(B))e^{\frac{-\Gamma(B)}{4}(t-t_1)},\ \frac{4}{\Gamma(B)^2}\varepsilon_\phi^2\right\},
\end{aligned}$$

which contradicts to the assumption that $[t_1, t_2]$ is the maximal interval on which (7.2) holds true for any $t \in [t_1, t_2]$. $\square$

Lemma 7.4. *There is a $T_* \leqslant 10^4 R_0^{-3}$, such that if we let $\alpha_* \in (0, \frac{\pi}{2})$ satisfy $\sin^2(\alpha_*) = 10^{-2}R_0$, then the set $B_* := S^+_{\alpha_*}(T_*)$ satisfies the assumptions for the set B for $t_1 = T_*$ in Lemma 7.3. Furthermore, if $\varepsilon_\phi \leqslant 10^{-2}R_0^2$, then for any $t \geqslant T_*$,*

$$\mu_t\left(\mathrm{Conv}[\phi_{T_*\to t}(B_*)]\right) \geqslant \mu_{T_*}(B_*) \geqslant \frac{1}{2}\left(1+\frac{9}{10}R_0\right),$$

and

$$\inf_{x,y\in\mathrm{Conv}[\phi_{T_*\to t}(B_*)]}\langle x, y\rangle = D_t(B_*) \geqslant D_{T_*}(B_*) = 1-\frac{1}{50}R_0, \tag{7.5}$$

and

$$\inf_{x\in\mathrm{Conv}[\phi_{T_*\to t}(B_*)]}\langle x, M_t\rangle \geqslant \mu_{T_*}(B_*)\left(1+D_t(B_*)\right)-1 \geqslant \frac{4}{5}R_0. \tag{7.6}$$

Proof. We define

$$T_* \coloneqq \inf\left\{t\geqslant 0 \mid \partial_t R_t \leqslant 10^{-4}R_0^3\right\}.$$

Because, $1 \geqslant R_{T_*}-R_0 \geqslant T_* 10^{-4}R_0^3$, we have that $T_* \leqslant 10^4 R_0^{-3}$. Also, by the definition of T_*, we have that R_t is strictly increasing on $[0,T_*]$, in particular, $R_t \geqslant R_0$ for $t\in[0,T_*]$.

In order to verify the assumptions in Lemma 7.3 for $B_* = S^+_{\alpha_*}(T_*)$, we need to get the corresponding $D_{T_*}(B_*)$ and $f_{T_*}(B_*)$. First, it is easy to see that

$$D_{T_*}(B_*) = \cos(2\alpha_*) = 1-2\sin^2(\alpha_*) = 1-\frac{1}{50}R_0.$$

Then, we are going to estimate $\mu_{T_*}(B_*)$. By the same reason as in proving the first inequality in (6.6), that is, divide the integral $R_{T_*} = \int_{\mathbb{S}^{d-1}}\langle y, U_{T_*}\rangle f_{T_*}(y)\,\mathrm{d}y$ into integrals over $S^+_{\alpha_*}(T_*)$, $\mathbb{S}^{d-1}\backslash(S^+_{\alpha_*}(T_*)\cup S^-_{\alpha_*}(T_*))$, and $S^-_{\alpha_*}(T_*)$, we have that

$$\begin{aligned}
R_0 \leqslant R_{T_*} &\leqslant \mu_{T_*}\left(S^+_{\alpha_*}(T_*)\right) + \cos(\alpha_*)\mu_{T_*}\left(\mathbb{S}^{d-1}\backslash(S^+_{\alpha_*}(T_*)\cup S^-_{\alpha_*}(T_*))\right)\\
&\quad - \cos(\alpha_*)\mu_{T_*}\left(S^-_{\alpha_*}(T_*)\right)\\
&= (1+\cos(\alpha_*))\mu_{T_*}\left(S^+_{\alpha_*}(T_*)\right) + 2\cos(\alpha_*)\mu_{T_*}\left(\mathbb{S}^{d-1}\backslash(S^+_{\alpha_*}(T_*)\cup S^-_{\alpha_*}(T_*))\right)\\
&\quad - \cos(\alpha_*)\\
&\leqslant 2\mu_{T_*}\left(S^+_{\alpha_*}(T_*)\right) + 2\mu_{T_*}\left(\mathbb{S}^{d-1}\backslash(S^+_{\alpha_*}(T_*)\cup S^-_{\alpha_*}(T_*))\right) - \cos(\alpha_*).
\end{aligned}$$

Next, by the same reason as in proving (6.2), we have that

$$\begin{aligned}
&\mu_{T_*}\left(\mathbb{S}^{d-1}\backslash(S^+_{\alpha_*}(T_*)\cup S^-_{\alpha_*}(T_*))\right) \leqslant \frac{\partial_t R_t^2|_{t=T_*}+\varepsilon_\phi^2}{R^2_{T_*}(\sin^2(\alpha_*))}\\
&= \frac{2\partial_t R_t|_{t=T_*}}{R_{T_*}(\sin^2(\alpha_*))} + \frac{\varepsilon_\phi^2}{R^2_{T_*}(\sin^2(\alpha_*))} \leqslant \frac{2\cdot 10^{-4}R_0^3}{R_0\cdot 10^{-2}R_0} + \frac{10^{-4}R_0^2}{10^{-2}R_0} = 3\cdot 10^{-2}R_0.
\end{aligned}$$

Combine the above two inequalities, and $\cos(\alpha_*) \geqslant \cos^2(\alpha_*) = 1-\sin^2(\alpha_*) = 1-10^{-2}R_0$, we see that

$$\mu_{T_*}\left(S^+_{\alpha_*}(T_*)\right) \geqslant \frac{1}{2}\left(1+R_0-7\cdot 10^{-2}R_0\right) \geqslant \frac{1}{2}\left(1+\frac{9}{10}R_0\right).$$

Hence,

$$\begin{aligned}
&\Gamma\left(S^+_{\alpha_*}(T_*)\right) = \mu_{T_*}\left(S^+_{\alpha_*}(T_*)\right)\left(1+D_{T_*}(S^+_{\alpha_*}(T_*)\right)-1\\
&\geqslant \frac{1}{2}\left(1+\frac{9}{10}R_0\right)\left(2-\frac{1}{50}R_0\right)-1\\
&= \frac{9}{10}R_0 - \frac{1}{100}R_0 - \frac{9}{1000}R_0^2 \geqslant \frac{4}{5}R_0.
\end{aligned}$$

Then, our assumption on ε_ϕ in Lemma 7.4 also implies the assumption on ε_ϕ in Lemma 7.3, because

$$\frac{1}{4}\left(1 - D_{T_*}(S^+_{\alpha_*}(T_*)\right)\Gamma\left(S^+_{\alpha_*}(T_*)\right)^2 \geqslant \frac{16}{5000}R_0^3 \geqslant 10^{-4}R_0^4 \geqslant \varepsilon_\phi^2.$$

By our Lemma 7.3, we can finish the proof for Lemma 7.4. □

Before we proceed, we need to give some further definitions. Let T_*, α_* be the time and the angle obtained in Lemma 7.4, and recall that $B_* := S^+_{\alpha_*}(T_*)$. Now we define the set $B_t := \phi_{T_*\to t}(B_*)$, and its $\frac{R_0^2}{10^4}$-neighborhood set $\widetilde{B}_t$

$$\widetilde{B}_t := \left\{ x \in \mathbb{S}^{d-1} \;\middle|\; \sup_{y\in B_t}\langle x, y\rangle \geqslant 1 - \frac{R_0^2}{10^4} \right\}.$$

The following lemma is a further step after Lemma 6.9.

Lemma 7.5. *Fix a constant $\lambda \in (2/3, 1)$ and an angle $\alpha \in [\pi/100, \pi/2)$. Let T_*, α_* be the time and the angle obtained in Lemma 7.4. If there is a time window $[t_1, t_2]$, such that when $t \in [t_1, t_2]$,*

$$R_t \geqslant \lambda R_0, \quad \partial_t R_t \leqslant \frac{1}{8}(\sin^4\alpha)\lambda^3 R_0^3,$$

and if

$$\varepsilon_\phi \leqslant \frac{1}{10^4}\lambda^2 R_0^2 \sin\alpha\cos\alpha,$$

then, for any $t_3 \to t_4 \in [t_1, t_2]$, $t_3 \leqslant t_4$, and any $x, y \in \mathbb{S}^{d-1}$ such that

$$\phi_{t_3\to t_4}(x) \in S^+_\alpha(t_4), \quad y \in B_*, \text{ and } \left\langle \phi_{t_3\to t_4}(x), \phi_{T_*\to t_4}(y)\right\rangle \leqslant 1 - \frac{R_0^2}{10^4},$$

we have that

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=t_4} \left\langle \phi_{t_3\to t}(x), \phi_{T_*\to t}(y)\right\rangle \geqslant \frac{\lambda R_0^{\frac{3}{2}}(\cos^2\alpha)}{4}\left(1 - \left\langle \phi_{t_3\to t}(x), \phi_{T_*\to t}(y)\right\rangle\right). \tag{7.7}$$

As a corollary, if we define,

$$\widetilde{\delta} = \widetilde{\delta}(\lambda, R_0, \alpha) := \frac{4\log(10^4 R_0^{-2})}{\lambda R_0^{\frac{3}{2}}(\cos^2\alpha)}, \tag{7.8}$$

then if $t_2 - t_1 \geqslant \widetilde{\delta}$, we have that

$$\phi_{t_1\to t_2}\left(S^+_\alpha(t_1)\right) \subseteq \widetilde{B}_{t_2}.$$

Proof. Take any $x \in \mathbb{S}^{d-1}$ and any $y \in B_*$ such that $\left\langle \phi_{t_3\to t_4}(x), \phi_{T_*\to t_4}(y)\right\rangle \leqslant 1 - \frac{R_0^2}{10^4}$ and $\phi_{t_3\to t_4}(x) \in S^+_\alpha(t_4)$. Notice that $\phi_{t_3\to t_4}(x)$ and $\phi_{T_*\to t_4}(y)$ are in the same hemisphere, because (7.6) means that $\phi_{T_*\to t_4}(y) \in S^+_{\frac{\pi}{2}}(t_4)$. By the same computation as (7.3), we see that

$$\begin{aligned}
&\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=t_4} \left\langle \phi_{t_3\to t}(x), \phi_{T_*\to t}(y)\right\rangle = 2\sin(\theta)\sin(2\theta)\langle Z, M_t\rangle \\
&\quad + \left\langle W_{t_4}(\phi_{t_3\to t_4}(x)), \mathbf{P}_{\phi_{t_3\to t_4}(x)}[\phi_{T_*\to t_4}(y)]\right\rangle + \left\langle \mathbf{P}_{\phi_{T_*\to t_4}(y)}[\phi_{t_3\to t_4}(x)], W_t(\phi_{T_*\to t_4}(y))\right\rangle \\
&\geqslant 2\sin(\theta)\sin(2\theta)\langle Z, M_t\rangle - 2\varepsilon_\phi\sin(2\theta),
\end{aligned}$$

where $\theta \in [0, \frac{\pi}{2}]$ such that $\cos(2\theta) = \langle \phi_{t_3\to t_4}(x), \phi_{T_*\to t_4}(y)\rangle$, and $Z \in \mathbb{S}^{d-1}$ is the middle point on the shortest great circle connecting $\phi_{t_3\to t_4}(x), \phi_{T_*\to t_4}(y)$. Notice that because (7.6) implies $\langle \phi_{T_*\to t_4}(y), M_t\rangle > 0$, we have that

$$\langle Z, M_t\rangle = \frac{\langle \phi_{t_3\to t_4}(x) + \phi_{T_*\to t_4}(y), M_t\rangle}{\|\phi_{t_3\to t_4}(x) + \phi_{T_*\to t_4}(y)\|_2} \geqslant \frac{\langle \phi_{t_3\to t_4}(x), M_t\rangle}{2} \geqslant \frac{\cos\alpha}{2} R_t \geqslant \frac{\cos\alpha}{2}\lambda R_0.$$

Also, we can write $\sin(2\theta) = \sqrt{1-\cos^2(2\theta)}$ and $\sin(\theta) = \sqrt{\frac{1-\cos(2\theta)}{2}}$. By the assumption, $\cos(2\theta) = \left\langle \phi_{t_3\to t_4}(x), \phi_{T_*\to t_4}(y)\right\rangle \leqslant 1 - \frac{R_0^2}{10^4}$, which implies that

$$\sin(\theta)\langle Z, M_t\rangle \geqslant \sqrt{\frac{R_0^2}{2\cdot 10^4}}\frac{\cos\alpha}{2}\lambda R_0 \geqslant \frac{\lambda R_0^2\cos\alpha}{300} \geqslant 10\varepsilon_\phi.$$

So,

$$\begin{aligned}\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=t_4} \left\langle \phi_{t_3\to t}(x), \phi_{T_*\to t}(y)\right\rangle &\geqslant \frac{\lambda R_0 \cos\alpha}{2}\sin(\theta)\sin(2\theta) \\ &= \frac{\lambda R_0\cos\alpha}{2\sqrt{2}}(1-\cos(2\theta))\sqrt{1+\cos(2\theta)}.\end{aligned}$$

Because we cannot rule out the case when $\cos(2\theta) < 0$, we need to get a lower bound for $1+\cos(2\theta)$. (7.6) implies $\langle \phi_{T_*\to t_4}(y), M_{t_4}\rangle > \frac{4}{5}R_0$. Because $M_{t_4} = R_{t_4}U_{t_4}$ and $R_{t_4} \leqslant 1$, we see that $\langle \phi_{T_*\to t_4}(y), U_{t_4}\rangle > \frac{4}{5}R_0$. Because $\phi_{t_3\to t_4}(x) \in S_\alpha^+(t_4)$, we have that $\langle \phi_{t_3\to t_4}(x), U_{t_4}\rangle \geqslant \cos\alpha$. We use the following fact: for any $Z_1, Z_2, Z_3 \in \mathbb{S}^{d-1}$,

$$\begin{aligned}\langle Z_1, Z_2\rangle &= \langle Z_1, Z_3\rangle\langle Z_2, Z_3\rangle + \langle \mathbf{P}_{Z_3}[Z_1], \mathbf{P}_{Z_3}[Z_2]\rangle \\ &\geqslant \langle Z_1, Z_3\rangle\langle Z_2, Z_3\rangle - \|\mathbf{P}_{Z_3}[Z_1]\|_2\|\mathbf{P}_{Z_3}[Z_2]\|_2.\end{aligned} \tag{7.9}$$

Hence,

$$\begin{aligned}\cos(2\theta) &= \langle \phi_{t_3\to t_4}(x), \phi_{T_*\to t_4}(y)\rangle \\ &\geqslant \langle \phi_{T_*\to t_4}(y), U_{t_4}\rangle\langle \phi_{t_3\to t_4}(x), U_{t_4}\rangle - 1 \geqslant \frac{4}{5}R_0\cos\alpha - 1.\end{aligned}$$

Combine the above arguments, we obtain (7.7):

$$\left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=t_4} \left\langle \phi_{t_3\to t}(x), \phi_{T_*,t}(y)\right\rangle \geqslant \frac{\lambda R_0^{\frac{3}{2}}(\cos^2\alpha)}{4}(1-\cos(2\theta)).$$

Next, we show that if $t_2 - t_1 \geqslant \widetilde{\delta}$ for the $\widetilde{\delta}$ defined in (7.8), we have $\phi_{t_1\to t_2}\left(S_\alpha^+(t_1)\right) \subseteq \widetilde{B}_{t_2}$. Because the assumptions on R_t, $\partial_t R_t$, and ε_ϕ in Lemma 6.9 are also satisfied here, by (6.9) in Lemma 6.9, we first know that $\phi_{t_1\to t}\left(S_\alpha^+(t_1)\right) \subseteq S_\alpha^+(t)$ for any time $t \in [t_1, t_2]$, because (6.9) means that for points already in $S_\alpha^+(t_1)$, those points along the characteristic flow, that is, $\phi_{t_1\to t}$, cannot escape the cap $S_\alpha^+(t)$ for any time $t \in [t_1, t_2]$. By (7.7), we have that for any $x \in S_\alpha^+(t_1)$, $y \in B_*$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(1 - \left\langle \phi_{t_1\to t}(x), \phi_{T_*\to t}(y)\right\rangle\right) \leqslant -\frac{\lambda R_0^{\frac{3}{2}}(\cos^2\alpha)}{4}\left(1 - \left\langle \phi_{t_1\to t}(x), \phi_{T_*\to t}(y)\right\rangle\right),$$

as long as $1 - \left\langle \phi_{t_1\to t}(x), \phi_{T_*\to t}(y)\right\rangle \geqslant \frac{R_0^2}{10^4}$. Hence, after at most $\widetilde{\delta}$ time, we have that $1 - \left\langle \phi_{t_1\to t}(x), \phi_{T_*\to t}(y)\right\rangle \leqslant \frac{R_0^2}{10^4}$, which implies that for any $t \geqslant t_1 + \widetilde{\delta}$, $\phi_{t_1\to t_2}\left(S_\alpha^+(t_1)\right)$ is contained in $\widetilde{B}_t$, the $\frac{R_0^2}{10^4}$-neighborhood of B_t. $\square$

Lemma 7.6. *Fix a constant $\lambda \in (2/3, 1)$ and an angle $\alpha \in [\pi/100, \pi/2)$. Let T_*, α_* be the time and the angle obtained in Lemma 7.4. Assume that there is a time window $[t_1, t_2]$, such that when $t \in [t_1, t_2]$,*

$$R_t \geqslant \lambda R_0, \quad \partial_t R_t \leqslant \frac{1}{8}(\sin^4 \alpha)\lambda^3 R_0^3.$$

If

$$\varepsilon_\phi \leqslant \frac{1}{10^4}\lambda^2 R_0^2 \sin\alpha\cos\alpha,$$

and if $t_2 - t_1 \geqslant \delta + \widetilde{\delta}$ for δ defined in (6.10) *in Lemma 6.9, and $\widetilde{\delta}$ defined in* (7.8) *in Lemma 7.5, we have that*

$$f_{t_2}^2\left(\mathbb{S}^{d-1}\backslash\widetilde{B}_{t_2}\right) \leqslant f_{t_1}^2\left(S_\alpha^-(t_1)\right)\cdot e^{3(d-1)(\widetilde{\delta}+\delta)}\cdot e^{-\frac{(d-1)\lambda R_0\cos\alpha}{2}(t_2-t_1)}.$$

Proof. By Lemma 7.5, we have that $\phi_{t_2-\widetilde{\delta}\to t_2}\left(S_\alpha^+(t_2-\widetilde{\delta})\right) \subseteq \widetilde{B}_{t_2}$. Hence, because $\phi_{t_2-\widetilde{\delta}\to t_2}$ is a diffeomorphism on $\mathbb{S}^{d-1}$, we see that

$$\begin{aligned}
&f_{t_2}^2\left(\mathbb{S}^{d-1}\backslash\widetilde{B}_{t_2}\right) \leqslant f_{t_2}^2\left(\mathbb{S}^{d-1}\backslash\phi_{t_2-\widetilde{\delta}\to t_2}\left(S_\alpha^+(t_2-\widetilde{\delta})\right)\right) = f_{t_2}^2\left(\phi_{t_2-\widetilde{\delta}\to t_2}\left(\mathbb{S}^{d-1}\backslash S_\alpha^+(t_2-\widetilde{\delta})\right)\right)\\
&\leqslant e^{2(d-1)\widetilde{\delta}}\cdot f_{t_2-\widetilde{\delta}}^2\left(\mathbb{S}^{d-1}\backslash S_\alpha^+(t_2-\widetilde{\delta})\right),
\end{aligned}$$

where in the second inequality, we used Lemma 7.1. Next, by Lemma 6.9, we have that $\mathbb{S}^{d-1}\backslash S_\alpha^+(t_2-\widetilde{\delta}) \subseteq \phi_{t_2-\widetilde{\delta}-\delta\to t_2-\widetilde{\delta}}\left(S_\alpha^-(t_2-\widetilde{\delta}-\delta)\right)$. Using Lemma 7.1 again, we have that

$$\begin{aligned}
f_{t_2}^2\left(\mathbb{S}^{d-1}\backslash\widetilde{B}_{t_2}\right) &\leqslant e^{2(d-1)\widetilde{\delta}}\cdot f_{t_2-\widetilde{\delta}}^2\left(\phi_{t_2-\widetilde{\delta}-\delta\to t_2-\widetilde{\delta}}\left(S_\alpha^-(t_2-\widetilde{\delta}-\delta)\right)\right)\\
&\leqslant e^{2(d-1)(\widetilde{\delta}+\delta)}\cdot f_{t_2-\widetilde{\delta}-\delta}^2\left(S_\alpha^-(t_2-\widetilde{\delta}-\delta)\right).
\end{aligned}$$

By Lemma 6.8, we finally obtain that

$$f_{t_2}^2\left(\mathbb{S}^{d-1}\backslash\widetilde{B}_{t_2}\right) \leqslant f_{t_1}^2\left(S_\alpha^-(t_1)\right)\cdot e^{2(d-1)(\widetilde{\delta}+\delta)}\cdot e^{-\frac{(d-1)\lambda R_0\cos\alpha}{2}(t_2-t_1-\delta-\widetilde{\delta})}.$$

□

Before we proceed, we need another auxiliary lemma similar to Lemma 7.4.

Lemma 7.7. *Let T_* be the time obtained in Lemma 7.4 and take two times t_1, t such that $T_* \leqslant t_1 \leqslant t$. If $\varepsilon_\phi \leqslant 10^{-3}R_0^2$, then*

$$\mu_t\left(\phi_{t_1\to t}\left(\widetilde{B}_{t_1}\right)\right) \geqslant \frac{1}{2}\left(1+\frac{9}{10}R_0\right),$$

and

$$\inf_{x,y\in\mathrm{Conv}\left[\phi_{t_1\to t}\left(\widetilde{B}_{t_1}\right)\right]}\langle x,y\rangle \geqslant D_{t_1}(\widetilde{B}_{t_1}) = 1-\frac{1}{10}R_0,$$

and

$$\inf_{x\in\mathrm{Conv}\left[\phi_{t_1\to t}\left(\widetilde{B}_{t_1}\right)\right]}\langle x, M_t\rangle \geqslant \mu_{t_1}\left(\widetilde{B}_{t_1}\right)\left(1+D_{t_1}(\widetilde{B}_{t_1})\right)-1 \geqslant \frac{4}{5}R_0. \tag{7.10}$$

Proof. The proof is similar to the proof for Lemma 7.4. We need to verify the assumptions in Lemma 7.3 for the set $\widetilde{B}_{t_1}$, for which we need to get the corresponding $D_{t_1}(\widetilde{B}_{t_1})$ and $f_{t_1}(\widetilde{B}_{t_1})$. First, take any $x, y \in \widetilde{B}_{t_1}$, then by the definition of $\widetilde{B}_{t_1}$, there are corresponding $x', y' \in B_{t_1} = \phi_{T_*, t_1}(B_*)$, such that $\langle x, x'\rangle \geqslant 1 - \frac{R_0^2}{10^4}$ and $\langle y, y'\rangle \geqslant 1 - \frac{R_0^2}{10^4}$. Using the inequality (7.9), we have that

$$\begin{aligned}\langle x, y\rangle &\geqslant \langle x, y'\rangle\langle y, y'\rangle - \sqrt{1-\langle y, y'\rangle^2}\\ &\geqslant \langle x', y'\rangle\langle x, x'\rangle\langle y, y'\rangle - \sqrt{1-\langle y, y'\rangle^2} - \sqrt{1-\langle x, x'\rangle^2}.\end{aligned}$$

By (7.5) in Lemma 7.4, we have that $\langle x', y'\rangle \geqslant 1 - \frac{R_0}{50}$. Hence, we have that

$$\begin{aligned}\langle x, y\rangle &\geqslant \left(1-\frac{R_0}{50}\right)\left(1-\frac{R_0^2}{10^4}\right)\left(1-\frac{R_0^2}{10^4}\right) - 2\sqrt{\frac{2R_0^2}{10^4}-\frac{R_0^4}{10^8}}\\ &\geqslant 1 - \frac{3R_0}{100} - \frac{3R_0}{100} \geqslant 1 - \frac{R_0}{10} > \frac{\sqrt{2}}{2}.\end{aligned}$$

Because $x, y \in \widetilde{B}_{t_1}$ are arbitrary, we have that $\inf_{x,y\in\widetilde{B}_{t_1}}\langle x, y\rangle \geqslant 1 - \frac{R_0}{10} > \frac{\sqrt{2}}{2}$. By Lemma 7.2, we have that $D_{t_1}(\widetilde{B}_{t_1}) = \inf_{x,y\in\mathrm{Conv}[\widetilde{B}_{t_1}]}\langle x, y\rangle \geqslant 1 - \frac{R_0}{10}$. Also, by Lemma 7.4, we have that

$$\mu_{t_1}\left(\widetilde{B}_{t_1}\right) \geqslant \mu_{t_1}(B_{t_1}) = \mu_{T_*}(B_*) \geqslant \frac{1}{2}\left(1 + \frac{9}{10}R_0\right).$$

Hence, to check the assumptions in Lemma 7.3, we see that

$$\begin{aligned}\Gamma\left(\widetilde{B}_{t_1}\right) &= \mu_{t_1}\left(\widetilde{B}_{t_1}\right)\left(1 + D_{t_1}(\widetilde{B}_{t_1})\right) - 1\\ &\geqslant \frac{1}{2}\left(1+\frac{9}{10}R_0\right)\left(2-\frac{R_0}{10}\right) - 1\\ &= \frac{18}{20}R_0 - \frac{1}{20}R_0 - \frac{9}{200}R_0^2 \geqslant \frac{4}{5}R_0.\end{aligned}$$

Also, by the definition of $\widetilde{B}_{t_1}$, we see that $D_{t_1}(\widetilde{B}_{t_1}) = \inf_{x,y\in\mathrm{Conv}[\widetilde{B}_{t_1}]}\langle x, y\rangle \leqslant 1 - \frac{R_0^2}{10^4}$. We get that

$$\frac{1}{4}\left(1 - D_{t_1}(\widetilde{B}_{t_1})\right)\Gamma\left(\widetilde{B}_{t_1}\right)^2 \geqslant \frac{1}{4}\cdot\frac{R_0^2}{10^4}\cdot\frac{16R_0^2}{25} \geqslant \varepsilon_\phi^2.$$

We then finish the proof by Lemma 7.3. $\square$

Proof of Theorem 3.8. Similar to the proof for Theorem 3.6, we fix the $\lambda = 1 - 10^{-10}R_0^4 \geqslant 1 - 10^{-10}R_0^2$, $\alpha = \frac{\pi}{100}$, and divide $\mathbb{R}_{\geqslant 0}$ into pieces $0 = s_{-1} \leqslant t_0 < s_0 \leqslant t_1 < s_1 \leqslant t_2 < s_2 \cdots$, where for any $k \geqslant 0$

$$t_k := \inf\left\{t \geqslant s_{k-1} \;\middle|\; \partial_t R_t \geqslant \frac{1}{8}(\sin^4\alpha)\lambda^3 R_0^3\right\}, \quad s_k := t_k + 1.$$

As in the proof for Theorem 3.6, we showed that this construction must stop at some k_*-th step and $k_* \leqslant 10^9 R_0^{-6}$. After the time s_{k_*}, we already saw in the proof of Theorem 3.6 that $f_t\left(\mathbb{S}^{d-1}\backslash S_\alpha^+(t)\right)$ starts to decay exponentially fast. To go further, we estimate how large s_{k_*} can be without using Theorem 6.10 directly.

We first prove that there is an upper bound C_* depending on $\|f_0\|_{L^2(\mathbb{S}^{d-1})}$ and R_0, such that for any $k \in [-1, k_*]$, $f^2_{s_k}\left(S^-_\alpha(s_k)\right) \leqslant C_*$. We fix a $k \in [-1, k_*]$. First, if $s_k \leqslant T_*$ for the T_* obtained in Lemma 7.4, by Lemma 7.1, we have that

$$f^2_{s_k}\left(S^-_\alpha(s_k)\right) \leqslant f^2_{s_k}\left(\mathbb{S}^{d-1}\right) \leqslant \|f_0\|^2_{L^2(\mathbb{S}^{d-1})} \cdot e^{2(d-1)s_k} \leqslant \|f_0\|^2_{L^2(\mathbb{S}^{d-1})} \cdot e^{2(d-1)T_*}.$$

Now, if $s_k > T_*$, take the l such that the following two conditions are satisfied:

(1) For any $p \in [l+1, k]$ ($\varnothing$ if $l = k$), $t_p - s_{p-1} \leqslant \delta + \widetilde{\delta}$ for δ defined in (6.10) in Lemma 6.9, and $\widetilde{\delta}$ defined in (7.8) in Lemma 7.5.

(2) $t_l - (\delta + \widetilde{\delta}) > s_{l-1} > T_*$ or $s_{l-1} \leqslant T_* \leqslant s_l$.

If the first case in the condition (2) above holds true, we apply the (7.10) in Lemma 7.7, which implies that $\phi_{t_l \to s_k}\left(\widetilde{B}_{t_l}\right) \subseteq S^+_{\frac{\pi}{2}}(s_k)$. Hence, because in this case, $s_k - t_l \leqslant (k-l)(\delta + \widetilde{\delta} + 1) + 1$, by Lemma 7.1, we get that

$$f^2_{s_k}\left(S^-_\alpha(s_k)\right) \leqslant f^2_{s_k}\left(\mathbb{S}^{d-1} \backslash \phi_{t_l \to s_k}\left(\widetilde{B}_{t_l}\right)\right) \leqslant e^{2(d-1)[(k-l)(\delta+\widetilde{\delta}+1)+1]} f^2_{t_l}\left(\mathbb{S}^{d-1} \backslash \widetilde{B}_{t_l}\right).$$

Now, combine this inequality with Lemma 7.6 for the interval $[s_{l-1}, t_l]$, we get that

$$f^2_{s_k}\left(S^-_\alpha(s_k)\right) \leqslant e^{3(d-1)[(k-l+1)(\delta+\widetilde{\delta}+1)]} f^2_{s_{l-1}}\left(S^-_\alpha(s_{l-1})\right).$$

Using this inequality, we can iteratively pull s_k back to the time when the second case in the condition (2) happens, and this iteration does not exceed k_*-times. In this case, we have that $s_{k-1} \leqslant T_* \leqslant s_k$. If $T_* < s_k - \delta - \widetilde{\delta} - 1$, we have that

$$\begin{aligned} f^2_{s_k}\left(S^-_\alpha(s_k)\right) &\leqslant f^2_{s_k}\left(\mathbb{S}^{d-1} \backslash \phi_{t_k \to s_k}\left(\widetilde{B}_{t_k}\right)\right) \leqslant e^{2(d-1)} f^2_{t_k}\left(\mathbb{S}^{d-1} \backslash \widetilde{B}_{t_k}\right) \\ &\leqslant e^{3(d-1)(\delta+\widetilde{\delta}+1)} f^2_{T_*}\left(S^-_\alpha(T_*)\right) \leqslant e^{3(d-1)(\delta+\widetilde{\delta}+1+T_*)} \|f_0\|^2_{L^2(\mathbb{S}^{d-1})}, \end{aligned}$$

where the last inequality follows from Lemma 7.1 on $[0, T_*]$. If $T_* \geqslant s_k - \delta - \widetilde{\delta} - 1$, then $s_k - T_* \leqslant \delta + \widetilde{\delta} + 1$. Using Lemma 7.1 again, we have that

$$\begin{aligned} f^2_{s_k}\left(S^-_\alpha(s_k)\right) &\leqslant f^2_{s_k}\left(\mathbb{S}^{d-1} \backslash B_{s_k}\right) \leqslant e^{2(d-1)(s_k - T_*)} f^2_{T_*}\left(\mathbb{S}^{d-1} \backslash B_*\right) \\ &\leqslant e^{2(d-1)(\delta+\widetilde{\delta}+1+T_*)} \|f_0\|^2_{L^2(\mathbb{S}^{d-1})}. \end{aligned}$$

Combine the above arguments in all possibilities, we obtain that for any $k \in [-1, k_*]$, we have that

$$f^2_{s_k}\left(S^-_\alpha(s_k)\right) \leqslant e^{4(d-1)[k_*(\delta+\widetilde{\delta})+T_*]} \|f_0\|^2_{L^2(\mathbb{S}^{d-1})}.$$

Next, we need to estimate each $t_k - s_{k-1}$ for $k \in [0, k_*]$. Assume that $t_k - s_{k-1} > \delta$ for δ defined in (6.10) in Lemma 6.9. Because for $t \in [s_{k-1}, t_k]$, by definition of t_k, we have that $\partial_t R_t \leqslant \frac{1}{8}(\sin^4 \alpha)\lambda^3 R_0^3$, we can then apply Lemma 6.9 to obtain that for any $r \in [s_{k-1} + \delta, t_k]$, $\mathbb{S}^{d-1} \backslash S^+_\alpha(r) \subseteq \phi_{r-\delta \to r}\left(S^-_\alpha(r - \delta)\right)$. By the same reason we obtained (6.12) in Theorem 6.10, we can obtain that

$$\mu_r\left(\mathbb{S}^{d-1} \backslash S^+_\alpha(r)\right) \leqslant 10 \cdot e^{-\frac{(d-1)\lambda R_0 \cos\alpha}{4}(r - \delta - s_{k-1})} \left[f^2_{s_{k-1}}\left(S^-_\alpha(s_{k-1})\right)\right]^{\frac{1}{2}}.$$

Applying the upper bound for $f^2_{s_{k-1}}\left(S^-_\alpha(s_{k-1})\right)$ we got earlier, we see that

$$\mu_r\left(\mathbb{S}^{d-1} \backslash S^+_\alpha(r)\right) \leqslant e^{-\frac{(d-1)\lambda R_0 \cos\alpha}{4}(r - s_{k-1})} \cdot e^{3(d-1)[k_*(\delta+\widetilde{\delta})+T_*]} \|f_0\|_{L^2(\mathbb{S}^{d-1})}, \tag{7.11}$$

for any $r \in [s_{k-1} + \delta, t_k]$. To simplify the notation in the proof, we let $\eta \coloneqq \frac{(d-1)\lambda R_0 \cos\alpha}{4}$ and $A \coloneqq e^{3(d-1)[k_*(\delta+\widetilde{\delta})+T_*]}\left[\|f_0\|_{L^2((\mathbb{S}^{d-1}))}\right]^{\frac{1}{2}}$. By Theorem 3.5, we have that $I_t + \partial_t I_t \leqslant 10^2 \mu_t \left(\mathbb{S}^{d-1} \backslash S_\alpha^+(t)\right) \leqslant 10^2 A e^{-\eta(t-s_{k-1})}$. Multiply e^t on both sides and integrate from s_{k-1} to r, we obtain that

$$I_r \leqslant I_{s_{k-1}} e^{-r+s_{k-1}} + 10^3 A e^{-\xi(t-s_{k-1})},$$

where $\xi = \eta$ if $\eta < 1$ and $\xi = \frac{1}{2}$ if $\eta \geqslant 1$. Also, by the fact that $I_t \leqslant 2$ using its definition directly, we can simplify the above inequality and obtain that for any $r \in [s_{k-1} + \delta, t_k]$,

$$I_r \leqslant 10^4 A e^{-\xi(r-s_{k-1})}.$$

By Lemma 6.6, we have that $R_t \geqslant \lambda R_0$. Using Lemma 5.2, we have that for any $r \in [s_{k-1} + \delta, t_k]$,

$$\partial_t R_t\big|_{t=r} \leqslant \frac{3 I_r}{2\lambda R_0} + \frac{\varepsilon_\phi^2}{2\lambda R_0} \leqslant \frac{3 I_r}{2\lambda R_0} + 10^{-8} \lambda^3 R_0^3 \sin^2 \alpha.$$

In particular, we can pick $r = t_k$, and by the construction of t_k, we must have that

$$\frac{1}{8} (\sin^4 \alpha) \lambda^3 R_0^3 \leqslant \partial_t R_t\big|_{t=t_k}.$$

Recall that we already fixed $\alpha = \frac{\pi}{100}$ in the assumption, and λ is very close to 1 by our choice at the beginning. So, combine the two inequalities for $\partial_t R_t\big|_{t=t_k}$ and I_{t_k}, we have that

$$\frac{\sin^4 \alpha}{24} \lambda^4 R_0^4 \leqslant I_{t_k} \leqslant 10^4 A e^{-\xi(t_k - s_{k-1})}.$$

Hence,

$$\xi(t_k - s_{k-1}) \leqslant 3(d-1)[k_*(\delta + \widetilde{\delta}) + T_*] + \log\left[10^{17} R_0^{-4} \|f_0\|_{L^2((\mathbb{S}^{d-1}))}\right]. \tag{7.12}$$

Using (7.12), we sum both sides from $k = 1$ to $k = k_*$, and obtain that

$$\begin{aligned} \xi s_* &\leqslant 1 + 3(d-1)k_*[k_*(\delta + \widetilde{\delta}) + T_*] + k_* \log\left[10^{17} R_0^{-4} \|f_0\|_{L^2((\mathbb{S}^{d-1}))}\right]. \\ &\leqslant 10^{23}(d-1) R_0^{-14} + 10^9 R_0^{-6} \log\left[\|f_0\|_{L^2((\mathbb{S}^{d-1}))}\right], \end{aligned}$$

where we also used the fact that $k_* \leqslant 10^9 R_0^{-6}$, $T_* \leqslant 10^4 R_0^{-3}$, $\delta \leqslant 10^4 R_0^{-1}$, and $\widetilde{\delta} \leqslant 10^2 R_0^{-2}$. Hence, apply (7.11) for $r \in [s_{k_*} + \delta, +\infty]$ we see that

$$\mu_r \left(\mathbb{S}^{d-1} \backslash S_\alpha^+(r)\right) \leqslant e^{-\frac{(d-1)\lambda R_0 \cos \alpha}{4}(r - s_{k_*})} \cdot e^{3(d-1)[k_*(\delta+\widetilde{\delta}) + T_*]} \|f_0\|_{L^2((\mathbb{S}^{d-1}))}.$$

Hence, if we set

$$S_0 := \xi^{-1} \left[10^{24}(d-1) R_0^{-14} + 10^9 R_0^{-6} \log\left(\|f_0\|_{L^2(\mathbb{S}^{d-1})}\right)\right],$$

we have that when $r \geqslant S_0$,

$$\mu_r \left(\mathbb{S}^{d-1} \backslash S_\alpha^+(r)\right) \leqslant e^{-\frac{(d-1)R_0}{8}(r - S_0)} \cdot \|f_0\|_{L^2(\mathbb{S}^{d-1})}.$$

We now eliminate the dependence on the initial radius R_0 in the exponent by further evolving the flow. Specifically, we define

$$S_1 = \begin{cases} S_0, & \text{if } \mu_{S_0}(\mathbb{S} \backslash S_\alpha^+(S_0)) \leqslant 0.1, \\ S_0 + \dfrac{8}{R_0} \log\left(10 \|f_0\|_{L^2(\mathbb{S})}\right), & \text{otherwise.} \end{cases}$$

Then, as established in (6.15), we have $R_{S_1} \geqslant \frac{1}{2}$. Restart the flow at time S_1, and we define

$$S_2 := \xi_1^{-1} \left[10^{24} 2^{14}(d-1) + 10^9 2^6 \log\left(\|f_{S_1}\|_{L^2(\mathbb{S}^{d-1})}\right)\right],$$

with $\xi_1^{-1} = \frac{8}{(d-1)\lambda_1 \cos\alpha} \vee 1$ and $\lambda_1 = 1 - 10^{-10}2^{-4}$. For all $r \geqslant S_1 + S_2$, we then have the estimate

$$\mu_r\left(\mathbb{S}^{d-1}\backslash S_\alpha^+(r)\right) \leqslant e^{-\frac{(d-1)}{16}(r-S_1-S_2)} \cdot \|f_{S_1+S_2}\|_{L^2(\mathbb{S}^{d-1})}$$
$$\leqslant e^{-\frac{(d-1)}{16}(r-33(S_1+S_2))} \cdot \|f_0\|_{L^2(\mathbb{S}^{d-1})},$$

where we used the bound $\|f_{S_1+S_2}\|_{L^2(\mathbb{S}^{d-1})} \leqslant e^{2(d-1)(S_1+S_2)}\|f_0\|_{L^2(\mathbb{S}^{d-1})}$ from Lemma 7.1. To estimate S_1+S_2, note from Lemma 7.1 that $\log\left(\|f_{S_1}\|_{L^2(\mathbb{S}^{d-1})}\right) \leqslant 2(d-1)S_1 + \log\left(\|f_0\|_{L^2(\mathbb{S}^{d-1})}\right)$, which yields the upper bound

$$S_2 \leqslant 10^{31}(d-1) + 10^{14}(d-1)S_1 + 10^{13}\log\left(\|f_0\|_{L^2(\mathbb{S}^{d-1})}\right)$$

Combining with the definition of S_1 and noting that $R_0 \leqslant 1$, we arrive at

$$S_1 + S_2 \leqslant \left[\frac{16}{(d-1)R_0} \vee 1\right](d-1)\left[10^{39}(d-1)R_0^{-14} + 10^{24}R_0^{-6}\log\left(\|f_0\|_{L^2(\mathbb{S}^{d-1})}\right)\right].$$

Finally, if we set

$$T_0 \coloneqq \left[\frac{16}{(d-1)R_0} \vee 1\right](d-1)\left[10^{41}(d-1)R_0^{-14} + 10^{26}R_0^{-6}\log\left(\|f_0\|_{L^2(\mathbb{S}^{d-1})}\right)\right],$$

we have that when $r \geqslant T_0$

$$\mu_r\left(\mathbb{S}^{d-1}\backslash S_\alpha^+(r)\right) \leqslant e^{-\frac{(d-1)}{16}(r-T_0)} \cdot \|f_0\|_{L^2(\mathbb{S}^{d-1})},$$

and the result stated in Theorem 3.8 then follows. □

Acknowledgments Y.P. is supported in part by NSF under CCF-2131115. P.R. is supported by NSF grants DMS-2022448 and CCF2106377. The authors thank Tim Roith for carefully reading our preprint and pointing out a helpful example in Remark 3.5 of [BKK+25], which complements our Theorem 3.1.

References

[ABK+22] Pedro Abdalla, Afonso S Bandeira, Martin Kassabov, Victor Souza, Steven H Strogatz, and Alex Townsend. Expander graphs are globally synchronising. *arXiv preprint arXiv:2210.12788*, 2022.

[ABPV+05] Juan A Acebrón, Luis L Bonilla, Conrad J Pérez Vicente, Félix Ritort, and Renato Spigler. The Kuramoto model: A simple paradigm for synchronization phenomena. *Reviews of modern physics*, 77(1):137–185, 2005.

[AGS05] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[AST24] Álvaro Rodríguez Abella, João Pedro Silvestre, and Paulo Tabuada. The asymptotic behavior of attention in transformers. *arXiv preprint arXiv:2412.02682*, 2024.

[BCB15] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[BCM14] Dario Benedetto, Emanuele Caglioti, and Umberto Montemagno. On the complete phase synchronization for the kuramoto model in the mean-field limit. *arXiv preprint arXiv:1407.6551*, 2014.

[BKK+25] 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. *arXiv preprint arXiv:2501.03096*, 2025.

[BPA25] Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. Emergence of metastable clustering in mean-field transformer models. In *International Conference on Learning Representations*, 2025.

[CACP25] 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.

[CHJK12] Young-Pil Choi, Seung-Yeal Ha, Sungeun Jung, and Yongduck Kim. Asymptotic formation and orbital stability of phase-locked states for the Kuramoto model. *Physica D: Nonlinear Phenomena*, 241(7):735–754, 2012.

[CM14] Tobias Holck Colding and William P Minicozzi. Łojasiewicz inequalities and applications. *Surveys in Differential Geometry*, 19(1):63–82, 2014.

[CNWR24] Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*, 2024.

[CRMB24] Christopher Criscitiello, Quentin Rebjock, Andrew D McRae, and Nicolas Boumal. Synchronization on circles and spheres with nonlinear interactions. *arXiv preprint arXiv:2405.18273*, 2024.

[DCLT19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. Association for Computational Linguistics, 2019.

[DV05] Laurent Desvillettes and Cédric Villani. On the trend to global equilibrium for spatially inhomogeneous kinetic systems: the boltzmann equation. *Inventiones mathematicae*, 159(2):245–316, 2005.

[FL19] Amic Frouvelle and Jian-Guo Liu. Long-time dynamics for a simple aggregation equation on the sphere. In *Stochastic Dynamics Out of Equilibrium: Institut Henri Poincaré, Paris, France, 2017*, pages 457–479. Springer, 2019.

[GKPR24] Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model. *arXiv preprint arXiv:2410.06833*, 2024.

[GLPR24] 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.

[GLPR25] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bull. Amer. Math. Soc.*, 2025.

[Goo24] Gemini Team Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

[HHK10] Seung-Yeal Ha, Taeyoung Ha, and Jong-Ho Kim. On the complete synchronization of the kuramoto phase model. *Physica D: Nonlinear Phenomena*, 239(17):1692–1700, 2010.

[HHL25] Daniel Zhengyu Huang, Jiaoyang Huang, and Zhengjiang Lin. Convergence analysis of probability flow ode for score-based generative models. *IEEE Transactions on Information Theory*, 2025.

[HKMP20] Seung-Yeal Ha, Young-Heon Kim, Javier Morales, and Jinyeong Park. Emergence of phase concentration for the kuramoto–sakaguchi equation. *Physica D: Nonlinear Phenomena*, 401:132154, 2020.

[HKPZ16] Seung-Yeal Ha, Dongnam Ko, Jinyeong Park, and Xiongtao Zhang. Collective synchronization of classical and quantum oscillators. *EMS Surveys in Mathematical Sciences*, 3(2):209–267, 2016.

[KBH24] Hugo Koubbi, Matthieu Boussard, and Louis Hernandez. The impact of LoRA on the emergence of clusters in transformers. *arXiv preprint arXiv:2402.15415*, 2024.

[KK84] Yoshiki Kuramoto and Yoshiki Kuramoto. *Chemical turbulence*. Springer, 1984.

[KPR24] Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking. *arXiv preprint arXiv:2411.04990*, 2024.

[Kur75] Yoshiki Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In *International Symposium on Mathematical Problems in Theoretical Physics: January 23–29, 1975, Kyoto University, Kyoto/Japan*, pages 420–422. Springer, 1975.

[LCG+20] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*, 2020.

[Łoj63] Stanislaw Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

[MP22] Javier Morales and David Poyato. On the trend to global equilibrium for kuramoto oscillators. *Annales de l'Institut Henri Poincaré C*, 40(3):631–716, 2022.

[MTG17] Johan Markdahl, Johan Thunberg, and Jorge Gonçalves. Almost global consensus on the n-sphere. *IEEE Transactions on Automatic Control*, 63(6):1664–1675, 2017.

[PRY25] Yury Polyanskiy, Philippe Rigollet, and Andrew Yao. Synchronization of self-attention dynamics on the circle, 2025.

[SABP22] 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.

[Sim83] Leon Simon. Asymptotics for a class of non-linear evolution equations, with applications to geometric problems. *Annals of Mathematics*, 118(3):525–571, 1983.

[SS24] 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.

[Vas17] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[Vil09] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[ZBB+23] Yi Zhang, Arturs Backurs, Sebastien Bubeck, Ronen Eldan, Suriya Gunasekar, and Tal Wagner. Unveiling transformers with LEGO: A synthetic reasoning task, 2023.

(SC) Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA
Email address: `schen636@mit.edu`

(ZL) Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA
Email address: `linzj@mit.edu`

(YP) Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA
Email address: `yp@mit.edu`

(PR) Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA
Email address: `rigollet@math.mit.edu`