Synchronization of mean-field models on the circle

Y. Polyanskiy*, P. Rigollet†, A. Yao‡

January 2025

Abstract

This paper considers a mean-field model of n interacting particles whose state space is the unit circle, a generalization of the classical Kuramoto model. Global synchronization is said to occur if after starting from almost any initial state, all particles coalesce to a common point on the circle. We propose a general synchronization criterion in terms of L_1-norm of the third derivative of the particle interaction function. As an application we resolve a conjecture for the so-called self-attention dynamics (stylized model of transformers), by showing synchronization for all $\beta \geq -0.16$, which significantly extends the previous bound of $0 \leq \beta \leq 1$ from Criscitiello, Rebjock, McRae, and Boumal [11]. We also show that global synchronization does not occur when $\beta < -2/3$.

1 Introduction

In this paper, we consider a simple dynamical system consisting of n interacting particles $x_1, \ldots, x_n$ situated on the unit sphere $\mathbb{S}^1$, which we identify with the standard 1-torus $\mathbb{S}^1 \simeq \mathbb{T} \triangleq \mathbb{R}/2\pi\mathbb{Z}$. The dynamics are given by

$$\dot{x}_i(t) = -\sum_{j=1}^{n} f(x_i(t) - x_j(t)), \forall i \in [n]. \tag{S1}$$

Interest in such systems originated with the work of Kuramoto [24], who analyzed the special case where $f(x) = \sin(x)$.[1] Kuramoto discovered a fascinating synchronization phenomenon in this system, which we define below.

Definition 1.1. In (S1), *synchronization* occurs for a starting point $(x_i(0))_{1 \leq i \leq n}$ if there exists $x^* \in [0, 2\pi)$ such that $\lim_{t\to\infty} x_i(t) \equiv x^*$ for all $i \in [n]$. We say that a pair (A, f) exhibits *global synchronization* if synchronization occurs for almost every starting point (with respect to the volume measure).

*Department of EECS, MIT. Email: yp@mit.edu

†Department of Mathematics, MIT. Email: rigollet@math.mit.edu

‡Departments of EECS and Mathematics, MIT. Email: ajyao@mit.edu

[1] We consider only the so-called *homogeneous* Kuramoto model, without particle-dependent drift terms.

Mathematical models of synchronization, such as Kuramoto's, are simplified abstractions of the ubiquitous self-organization phenomena observed in nature [36]. While many other models have been proposed in physics [35], biology [27], and engineering [4, 19], Kuramoto's model remains the simplest to exhibit this effect.

The seminal paper of Kuramoto [24] has ushered a rich line of work sharpening, generalizing, and applying his original framework; see [32, 3]. After successive advances, global synchronization for the mean-field Kuramoto model was eventually established in [33], which also expanded the original Kuramoto model beyond the mean-field case by allowing each particle to interact with only a subset of the others; see [1, 20] for the most recent advances on this front.

The dynamical system (S1) is an example of a *mean-field* model, since its motion can be rewritten as:

$$\dot{x}_i(t) = \mathcal{X}[\mu_n(t)](x_i)\,, \qquad \mu_n(t) \triangleq \frac{1}{n}\sum_{j=1}^{n} \delta_{x_j(t)}\,, \quad \mathcal{X}[\mu](x) \triangleq -\int_{\mathbb{T}} f(x-y)\mu(dy)\,,$$

where $\mu_n(t)$ denotes the empirical measure (distribution) of the particles and $\mathcal{X}[\mu]$ is the measure-dependent vector field driving each particle. Since particles are indistinguishable in mean-field models, it suffices to study the evolution of the empirical measure $\mu_n(t)$, which satisfies the continuity equation:

$$\dot{\mu}(t) + \operatorname{div}(\mu\mathcal{X}[\mu]) = 0\,. \tag{1}$$

When $\mu(0) = \mu_n(0) = \frac{1}{n}\sum_{j=1}^{n}\delta_{x_j(0)}$, studying (1) is equivalent to studying (S1), but one can also study solutions of (1) starting from non-discrete measures $\mu(0)$.

For the Kuramoto mean-field model, i.e. $f(x) = \sin(x)$, [14] showed explicit (exponential) estimates of speed of convergence to synchronization for the dynamical system (S1) and, subsequently, [28] extend the exponential convergence to the more general case of evolution of measures solving continuity equation (1). Both works in fact consider a more general case of the Kuramoto mean-field model with state space $\mathbb{S}^d$ with $d \geq 1$.

A recent resurgence of interest in mean-field models on the sphere and torus arose from a discovery of [17] that with $f(x) = f_\beta(x) \triangleq \sin(x)e^{\beta\cos(x)}, \beta \in \mathbb{R}$ the resulting interacting particle system is intimately related to evolution of internal representations in transformers [34], which are modern neural networks forming the backbone of large language models (LLMs). When $f = f_\beta$, we call (S1) *self-attention dynamics*, on which there is a fast growing body of work [29, 15, 21, 16, 11, 16, 6, 2, 7, 8, 9]. Despite the complexity of practical transformers, the simple model of self-attention dynamics is remarkably effective at predicting how signals propagate through internal layers. From the practical point of view, global synchronization is an abstraction of a complex phenomenon in LLMs known as clustering or oversmoothing, e.g. [13, 30, 12, 23].

The work [17] establishes global synchronization whenever $\beta = O(1/n)$ or $\beta = \Omega(n^2)$ (in both cases $\beta \geq 0$ is also required) and for all state spaces $\mathbb{S}^d$, $d \geq 1$. Shortly afterwards, [11] made an important observation that an earlier work of [26] in fact shows global synchronization for all $\beta \geq 0$ and $d \geq 2$. For $d = 1$, the authors of [11] improved the argument of [17] and showed synchronization for $\beta \leq 1$. In [21], the results on global

synchronization are further generalized but under the additional assumption that the summation defining $\dot{x}_i$ in (S1) only extends to $j < i$, which corresponds to a simple model of the ubiquitous "decoder-only" LLMs implementing next-token prediction.

Meta-stability. The novel aspect of self-attention dynamics compared to Kuramoto's model is the emergence of meta-stability above a certain critical value of $\beta > 0$. Specifically, [17] observed empirically that once β is sufficiently large, then particles initialized i.i.d. uniformly evolve in two phases: first, they quickly group into $\asymp \sqrt{\beta}$ tight clusters and then over a much slower time-horizon the clusters progressively merge until only one remains, thus attaining global synchronization. In [16] it is confirmed that localized groups of particles contract exponentially quickly to their common center. Bruno, Pasqualotto, and Agazzi in [6] showed that after initializing the particles $x_i(0)$ i.i.d. from the uniform measure on the circle, at time $t \asymp \log n$ the empirical measure $\mu_n(t)$ develops periodic lumps with high probability. More explicitly, they show that for any $0 < \delta \ll 1$, there exists a $\frac{2\pi}{k}$-periodic probability distribution $\nu_{per}(t)$, with $k \asymp \sqrt{\beta}$, which is δ-away from uniform (as measured, for example, by the Wasserstein distance W_1) and $W_1(\mu_n(t), \nu_{per}(t)) \to 0$ in probability as $n \to \infty$ for $t = t(n, \delta, \beta) \asymp \log n$.

In the present work, we show that with probability 1, for any fixed n and $\beta \geq 0$, we must have that $\mu_n(t) \to \delta_{x_\infty}$ as $t \to \infty$. In turn, this implies that although the $\frac{2\pi}{k}$-periodic phase is rather long-lived, it will eventually collapse, which implies that it is meta-stable.

Contributions. In addition to (S1), in the context of Transformers a so-called "normalized" version of this dynamics is also important. This generalization of (S1) can be stated in the following form:

$$\dot{x}_i(t) = -\frac{1}{g_i(x_1(t), \dots, x_n(t))} \sum_{j=1}^{n} f(x_i(t) - x_j(t)),\ 1 \leq i \leq n, \tag{S2}$$

where $g : \mathbb{T}^n \to \mathbb{R}^n_{>0}$ is some smooth function. In this work, we propose a general criterion for systems (S1) and (S2) with state-space $\mathbb{S}^1 = \mathbb{T}$ to be globally synchronizing. As an application, we prove (a) global synchronization for transformers (i.e. $f = f_\beta$) for all $\beta \geq 0$, thus completing the study of this class of interaction functions; (b) initiate the study of $\beta < 0$ and show global synchronization for $\beta \geq -0.16$ and non-synchronization for $\beta < -2/3$. Finally, in Section 6 we extend the criterion to a certain class of non-mean-field systems.

Organization. Section 2 states all of our results formally. Section 3 contains proof of the main criterion for stability of system (S1), i.e. Theorem 2.1. Section 4 extends the results to the normalized system (S2). Section 5 verifies that the general criterion in Theorem 2.1 applies to Transformer dynamics on the circle ($f = f_\beta$ for all $\beta > -0.16$). Finally, in Section 6 we further generalize the results to the dynamics where particles are aggregated with unequal weights.

Acknowledgments. PR is supported by NSF grants DMS-2022448 and CCF-2106377.

2 Main results

Given a smooth vector field $\mathcal{F} : \mathcal{M} \to T\mathcal{M}$ on a manifold $\mathcal{M}$, a dynamical system solves the ordinary differential equation (ODE):

$$\dot{\mathbf{x}} = \mathcal{F}(\mathbf{x}) \,.$$

Point $\mathbf{x}$ is called stationary if $\mathcal{F}(\mathbf{x}) = 0$, since started from this point the system does not move: $\dot{\mathbf{x}} = 0$. If the trajectory of a dynamical system converges, then the limiting point should necessarily be a stationary point. We call a stationary point $\mathbf{x}$ *locally unstable* if the Jacobian of $\mathcal{F}$ at $\mathbf{x}$ has an eigenvalue with positive real part. Note that this implies that for a small neighborhood U around $\mathbf{x}$, almost all initializations $\mathbf{x}_0 \in U$ result in trajectories that escape from U.

The dynamical system (S1) that we consider here corresponds to taking $\mathcal{M} = \mathbb{T}^n$ and the vector field with i-th component being

$$\mathcal{F}(\mathbf{x})_i = \mathcal{X}[\mu_n](x_i) = -\int f(x_i - y) d\mu_n(y) \,.$$

As we discussed, mean-field models can also be thought of as n exchangeable particles (each with state space of $\mathbb{T}$) each driven by a time-dependent vector field $\mathcal{X}[\mu_n(t)] : \mathbb{T} \to T\mathbb{T}$, which is a function of the empirical measure μ_n, cf. (1).

Theorem 2.1. *Consider the mean-field model* (S1) *on* $\mathbb{T}$. *Let* $\tau \in (0, \pi]$ *satisfy* $f'(x) < 0$ *for all* $x \notin [-\tau, \tau]$. *If*

$$\tau \int_{-\pi}^{\pi} |f'''(x)|_+ dx \le 4 \left(1 + \frac{\tau}{2\pi}\right) f'(0),$$

then every stationary point $(x_1, \dots, x_n)$ *of the system* (S1) *on* $\mathbb{T}^n$ *is either locally unstable or synchronized (i.e.* $x_1 = \cdots = x_n$*).*

This characterization constitutes the main result of our work. However, to establish global synchronization in systems (S1) and (S2), we require two additional (though now standard) ingredients: Łojasiewicz's theorem and the center-stable manifold theorem.

2.1 Gradient ascent dynamics

The first obstacle to synchronization could be the emergence of limit cycles. It turns out, however, that Transformer dynamics is special since it can be written as a gradient ascent for a certain energy function $\mathsf{E}(\mathbf{x})$, see [17, (3.5)] and (3) below. Consequently, as explained in [17, Appendix A] (also [18, Corollary 5.1]), classical Łojasiewicz's theorem [25] guarantees that as long as E is real analytic, the gradient ascent dynamics $\dot{\mathbf{x}} = \nabla_{\mathcal{M}} \mathsf{E}(\mathbf{x})$ over a compact Riemannian manifold must converge to some stationary point $\mathbf{x}_\infty$. (We denote $\nabla_{\mathcal{M}}$ the Riemannian gradient on a manifold, see [5] for details.)

The next issue is that even for $f = f_\beta$, the system (S1) has many stationary points other than the synchronized ones (for example, when the particles are placed at the vertices of a regular n-gon). While Theorem 2.1 ascertains those must be locally unstable,

we still need to rule out existence of those serendipitous initial conditions that would lead to those limiting configurations. This is the content of another classical ingredient: the center-stable manifold theorem, which indeed shows that the limiting stationary point is almost always a stable one [31, Theorem III.7 and Exercise III.3].

Putting these two ideas together, we get the following result that together with Theorem 2.1 ascertain convergence to the synchronized states in f_β dynamics with or without normalization.

Lemma 2.2 ([17, Lemma A.1])**.** *Let $\mathcal{M}$ be a compact Riemannian manifold and let $\mathsf{E} : \mathcal{M} \to \mathbb{R}$ be a smooth function. The set of initial conditions $X_0 \in \mathcal{M}$ for which the gradient ascent system*

$$\begin{cases} \dot{\mathbf{x}}(t) = \nabla_{\mathcal{M}} \mathsf{E}(\mathbf{x}(t)), \\ \mathbf{x}(0) = X_0 \end{cases}$$

converges to a critical point of E at which the Hessian of E has a positive eigenvalue is of volume zero.

With these preparations, we are ready to derive our main global synchronization results by applying Lemma 2.2 and Theorem 2.1 to systems (S1) with $f(x) = h(\cos(x))\sin(x)$. Indeed, for such systems we can see that dynamics becomes a gradient ascent on the potential

$$E(\mathbf{x}) = \sum_{i,j} \phi(\cos(x_i - x_j)), \qquad \phi(t) = \int_0^t h(s)ds\,. \tag{2}$$

Note that a critical point $\mathbf{x}$ is locally unstable if and only if the Hessian at $\mathbf{x}$ has positive eigenvalue since the Hessian is symmetric, thus enabling application of Theorem 2.1. See Theorem 2.3 shortly, for the full statement.

2.2 Adjusted gradient ascent

It turns out that the method discussed above is applicable not only to systems of the type (S1), but also to more general systems with particle-dependent normalization factors, i.e. systems of the type (S2). We need to consider this extension because the the simplified model of self-attention (see (T2) below) has precisely such form.

Theorem 2.3. *Assume that $f(x) = \sin(x)h(\cos(x))$, where h is a real-analytic function on an open set containing $[-1, 1]$ and $\tau \int_{-\pi}^{\pi} |f'''(x)|_+ dx \leq 4\left(1 + \frac{\tau}{2\pi}\right) f'(0)$, where τ is as in Theorem 2.1. Then, global synchronization occurs in (S2).*

The full proof of this result is given in Section 4 below, but the idea is simple. First, when $f(x) = \sin(x)h(\cos(x))$ and normalization factors $g_i = 1$, then as we have seen in the previous section we are dealing with a gradient ascent on the potential (2), which thus must converge (generically) to a locally stable critical point. In the case of $g_i \neq 1$, we can follow the idea suggested in [17, Section 3.4 and Remark B.1]: by introducing a non-flat Riemannian metric on $(\mathbb{S}^1)^{\otimes n}$ we can make sure that the gradient of the same energy (2)

results in the normalized dynamics (S2). Then, this case also reduces to an application of Theorem 2.1.

We mention that we further generalize the last result to systems where each particle contributes to the RHS in (S2) with its own, particle-dependent weight factor. See Section 6 for more.

2.3 Self-attention dynamics

Next, we discuss an application of the main results to self-attention dynamics. Recall that the latter [17, (USA)] is defined as

$$\dot{x}_i(t) = -\sum_{j=1}^{n} e^{\beta\cos(x_i(t)-x_j(t))}\sin(x_i(t)-x_j(t)),\ 1\le i\le n, \tag{T1}$$

which corresponds to taking $f(x) = f_\beta(x) = \sin(x)e^{\beta\cos(x)}$. The global synchronization conjectured in [17] for all $\beta \ge 0$ was only shown for $\beta \le 1$ and $\beta \ge \Omega(1/n)$, cf. [11]. In this section we resolve the conjecture in full and in fact even extend it to a portion of $\beta < 0$.

Define the number

$$a(\beta) = \inf\{\tau : f'_\beta(x) < 0\,,\ \forall\, x \in (\tau,\pi], \tau \in [0,\pi]\}\,.$$

Theorem 2.3 implies that whenever

$$4\frac{1+\frac{a(\beta)}{2\pi}}{a(\beta)\int_{-\pi}^{\pi}|f'''(x)|_+dx} > 1,$$

global synchronization occurs. In fact, using specific properties of f_β we can strengthen the criterion in Theorem slightly, see Corollary 3.3 below, and guarantee global synchronization under the weaker assumption of

$$4\frac{1+\frac{a(\beta)}{\pi}}{a(\beta)\int_{-\pi}^{\pi}|f'''(x)|_+dx} > 1\,.$$

The quantity on the left-hand side is termed the *synchronization ratio* and we numerically plot it on Fig. 1. As one can see the criterion indeed is verified in the region of $\beta \ge -0.25$. We formally verify the inequality in a slightly narrower region of $\beta \ge -0.16$ below.

Corollary 2.4. *Suppose $\beta \ge -0.16$. Then, global synchronization occurs in (T1).*

The dynamics (T1) is a simplification of the actual self-attention dynamics, which is given by [17, (SA)]:

$$\dot{x}_i(t) = -\frac{1}{\sum_{j=1}^{n} e^{\beta\cos(x_i(t)-x_j(t))}}\sum_{j=1}^{n} e^{\beta\cos(x_i(t)-x_j(t))}\sin(x_i(t)-x_j(t)),\ 1\le i\le n. \tag{T2}$$

As already explained in the previous section (Theorem 2.3), the results about unnormalized system can be easily transported to results about the normalized system, which allows us to conclude with the following:

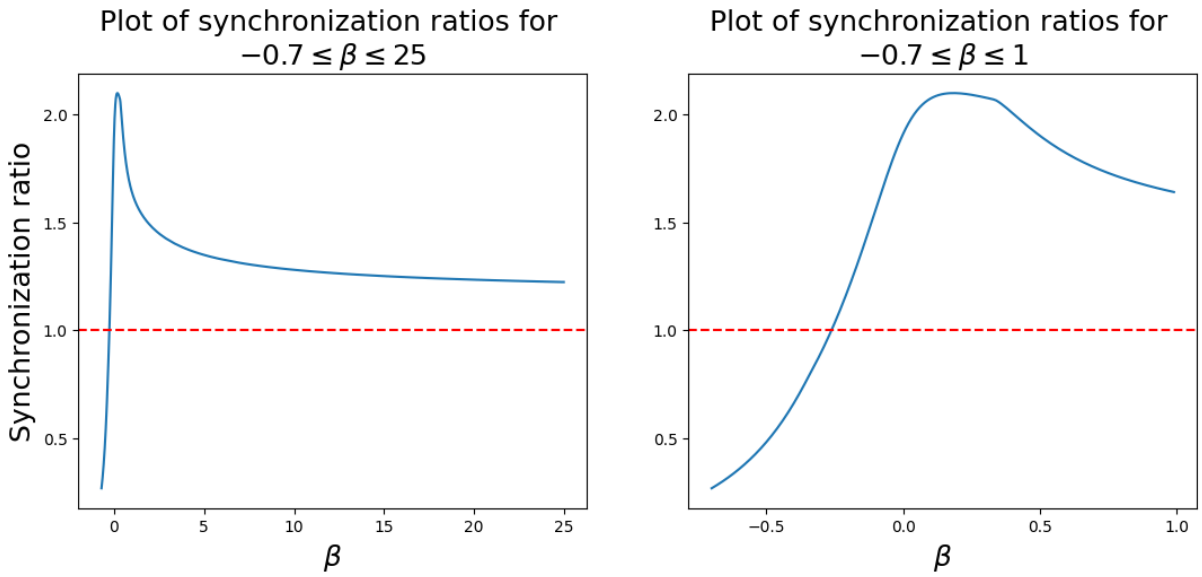

Figure 1: The figure plots the synchronization ratio $\langle |f'''|_+, \tau \rangle_{L_2}^{-1} 4\left(1 + \frac{\tau}{\pi}\right) f'(0)$ from Corollary 3.3 with $f(x)$ set as $\sin(x)e^{\beta(\cos(x)-1)}$ and M set as π. A ratio greater than one indicates that we have determined that global synchronization occurs.

Corollary 2.5. *Suppose $\beta \geq -0.16$. Then, global synchronization occurs in (T2).*

Finally, one might wonder whether global synchronization occurs for even more negative values of β. We show that this is not the case, thus leaving only the region $\beta \in \left(-\frac{2}{3}, -0.25\right)$ in uncertain synchronization status.

Corollary 2.6. *Suppose $\beta < -\frac{2}{3}$. There exists a constant C_β such that if n is divisible by 3 or $n \geq C_\beta$, then global synchronization does not occur in either* (T1) *or* (T2).

The proofs of all results can be found in Section 5.

We remark that for $\beta > 0$ self-attention dynamics (normalized or not) corresponds to gradient ascent on the potential

$$E(\mathbf{x}) = \frac{1}{\beta} \sum_{i,j} e^{\beta \cos(x_i - x_j)} \,. \tag{3}$$

For $\beta < 0$, self-attention dynamics is a *gradient descent* on the potential

$$E(\mathbf{x}) = \frac{1}{|\beta|} \sum_{i,j} e^{-|\beta| \cos(x_i - x_j)} \,.$$

In either case, the global optimizer is clearly the synchronized configuration. However, in the latter case local extrema with non-zero volume basin of attraction may emerge for large $|\beta|$.

Finally, we remark that *gradient descent* on the potential (3) with $\beta > 0$ yields a completely different dynamics, corresponding to taking $f(x) = -\sin(x)e^{\beta \cos(x)}$ in (S1).

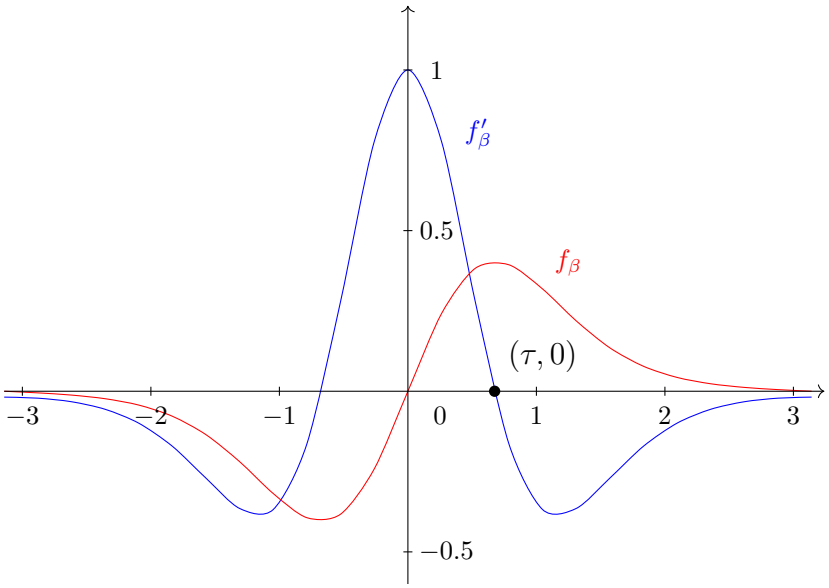

Figure 2: The function f_β (red) and its derivative f'_β (blue) for $\beta = 2$. The parameter $\tau = \tau(\beta)$ is defined as the unique solution to the equation $f'_\beta(\tau) = 0$ over $[0, \pi]$. For $\beta = 2$, $\tau \simeq 0.6749$.

In this case the particles tend to equi-disperse on the circle. Indeed, the unique global minimizer of (3) is the n-gon as shown in [10]. At the level of the evolution of measures, the unique global minimizer of the functions $\mu \mapsto \int\int e^{\beta\cos(x-y)}\mu(dx)\mu(dy)$ can be easily seen to be the uniform measure (e.g. by noticing that Fourier coefficients of $e^{\beta\cos(x)}$ are all positive, cf. [17, Section 7.2]), thus explaining the equidistribution tendency.

3 Proof of Theorem 2.1

Suppose $\mathbf{x} = (x_i)_{1\le i\le n} \in \mathbb{T}^n$. For the system (S1), a point is *stationary* iff

$$\sum_{j=1}^{n} f(x_i - x_j) = 0,\ \forall i \in [n]. \tag{C1}$$

In view of Lemma 2.2, we also need a condition for the point $\mathbf{x}$ to be stable. In fact, we only need a weaker condition that is implied by stability, which has been the basis of proving synchronization in Kuramoto-type dynamics since its introduction in [33, (2.4)]. We say that point $\mathbf{x}$ is *cut-stable* if

$$\sum_{i\in S,\, j\in S^C} f'(x_i - x_j) \ge 0 \tag{C2}$$

for all $S \subset [n]$ such that the value of x_i is the same for all $i \in S$.

Our proof will show that any stationary point satisfying (C2) must be synchronized. In other words, we will show that for any non-synchronized stationary point $\mathbf{x}$ there exists a set S defining an "escape direction" for the linearization of the system (S1), corresponding to moving points in S clockwise and points in S^c counter-clockwise at the same speed.[2]

We start with a review of the proof from [17] which applies to $f(x) = f_\beta(x) = e^{\beta\cos(x)}\sin(x)$, whose derivative f'_β when $\beta = 2$ is shown in Fig. 2, where the crucial parameter τ from Theorem 2.1 is also shown. If we apply (C2) with $S = \{1\}$ then we see that there must be at least one particle, say 2, at distance $\le \tau$ from 1 (otherwise all $f'(x_1 - x_j) < 0$). We can now apply the argument to $S = \{1, 2\}$ to find that 3 must be at distance $\le \tau$, etc. Overall, if $n\tau < \pi$ then all particles must be inside one half-circle. But then if i_0 is the boundary particle, then (C1) with $i = i_0$ implies that all $x_j = x_{i_0}$, because $f(x_{i_0} - x_j) \ge 0$. Unfortunately, this proof only shows synchronization when $\pi > n\tau \asymp \frac{n}{\sqrt{\beta}}$. Our contribution here is a method that extends to arbitrary large n.

To describe our idea, let us define the vector field acting on particles as

$$\chi(x) = \sum_j f(x - x_j)\,.$$

Then from (C1) and (C2) (applied with $S = \{i\}$) we know that

$$\chi(x_i) = 0, \qquad \chi'(x_i) \ge f'(0)\,.$$

Consider a pair of adjacent particles $x_i < x_{i+1}$. Because $\chi(x_i) = \chi(x_{i+1})$, we have from integration by parts that

$$\chi'(x_i) + \chi'(x_{i+1}) = \int_{x_i}^{x_{i+1}} \chi'''(x)\frac{(x - x_i)(x_{i+1} - x)}{x_{i+1} - x_i}dx\,.$$

As we have shown above, all intervals $x_{i+1} - x_i$ except possibly 1 are bounded by τ. In the special case when *all of them* are bounded by τ we can notice that the factor multiplying $\chi'''(x)$ is positive and upper-bounded by $\tau/2$. Thus summing over $i = 1, \dots, n-1$ we obtain

$$\begin{aligned} 2nf'(0) \le 2\sum_i \chi'(x_i) &= \sum_{i=1}^{n-1}\int_{x_i}^{x_{i+1}} \chi'''(x)\frac{(x - x_i)(x_{i+1} - x)}{(x_{i+1} - x_i)}dx \\ &\le \frac{\tau}{2}\int |\chi'''(x)|_+ dx \le \frac{n\tau}{2}\int |f'''|_+ dx\,. \end{aligned}$$

This inequality, however, is not possible if (as is the case for $f = f_\beta$ for large β) we have $\tau\int|f'''|_+ < 4f'(0)$. Consequently, one of the assumptions must be violated. The full proof below will show that in fact this contradiction implies that $x_1 = \ldots = x_n$.

We proceed to the formal proof of Theorem 2.1 and consider $\mathbf{x} = (x_1, \dots, x_n)$ satisfying conditions (C1) and (C2). Denote the distinct values of the x_i as $0 \le \theta_1 < \theta_2 < \cdots <$

[2]Note that when state space is $\mathbb{S}^d$ with $d > 1$, then finding escape directions in mean-field systems can be done by pulling all particles toward the same direction (subject to spherical constraints), cf. [26] and [11]. Our method is more involved.

$\theta_K < 2\pi$ and let $\theta_{K+1} = 2\pi + \theta_1$. For $1 \le j \le K$, we define $\Delta_j = \theta_{j+1} - \theta_j$, where we set $\Delta_K = 2\pi + \theta_1 - \theta_K$ to account for the periodic boundary. We refer to the Δ_j as the *gaps*. Furthermore, we let $\tau_{\max}(\mathbf{x})$ denote the maximum of Δ_j for $1 \le j \le K$.

Lemma 3.1. *Assume that (C2) is satisfied at $(x_1, \dots, x_n)$. There do not exist distinct gaps ω_1 and ω_2 such that $\omega_1, \omega_2 > \tau$.*

Proof. The idea is the same as [17, Appendix B]. Suppose that the line ℓ intersects the interiors of both ω_1 and ω_2. Let S be the set of i such that x_i is on one side of ℓ. Then, for all $i \in S$ and $j \in S^C$, we have that $x_i - x_j \notin (-\min(\omega_1, \omega_2), \min(\omega_1, \omega_2)) \pmod{2\pi}$. Because $\min(\omega_1, \omega_2) > \tau$, $x_i - x_j \notin [-\tau, \tau] \pmod{2\pi}$, which implies that $f'(x_i - x_j) < 0$. This is a contradiction to (C2). ∎

For $i \in [K]$, let N_i be the multiplicity of θ_i, i.e. the number of $j \in [n]$ such that $x_j = \theta_i$. Also, let

$$\langle f, g \rangle_{L_2} \triangleq \int_{-\pi}^{\pi} f(x) g(x) dx \,.$$

Lemma 3.2. *Suppose $(x_1, \dots, x_n) \in \mathbb{T}^n$ is stationary (C1) and cut-stable (C2) for the system* (S1)*. Furthermore, assume that $(x_1, \dots, x_n)$ is not synchronized. Then,*

$$\langle 1, |f'''|_+ \rangle_{L_2} \ge \min\left(\frac{8}{\tau}, \frac{4}{\tau} + \frac{4}{\tau_{max}(x_1, \dots, x_n)} \right) f'(0).$$

Proof. For the sake of contradiction, assume that $K > 1$. Let

$$\varphi(x) = \sum_{j=1}^{n} f'(x - x_j).$$

Then, after using (C2) with S equal to the set of $j \in [n]$ such that $x_j = \theta_i$,

$$\varphi(\theta_i) = N_i f'(0) + \sum_{j \in [n],\, x_j \neq \theta_i} f'(\theta_i - x_j) \ge N_i f'(0). \tag{4}$$

Let

$$\Psi = \sum_{i=1}^{K} \mathbf{1}\{x \in (\theta_i, \theta_{i+1})\} \Psi_i,$$

where $\Psi_i : [\theta_i, \theta_{i+1}] \to \mathbb{R}$ is twice differentiable. Then, for $i \in [K]$,

$$\langle \Psi_i, \varphi'' \rangle_{L_2} = \Psi_i \varphi' \big|_{\theta_i}^{\theta_{i+1}} - \langle \Psi_i', \varphi' \rangle_{L_2} = (\Psi_i \varphi' - \Psi_i' \varphi) \big|_{\theta_i}^{\theta_{i+1}} + \langle \Psi_i'', \varphi \rangle_{L_2} \,.$$

Assume that $\Psi_i'' = -a_i$ for a constant a_i and $\Psi_i(\theta_i) = \Psi_i(\theta_{i+1}) = 0$ over $[\theta_i, \theta_{i+1}]$, or equivalently that $\Psi_i(x) = \frac{a_i}{2}(x - \theta_i)(\theta_{i+1} - x)$. Then, we have that

$$\langle \Psi_i, \varphi'' \rangle_{L_2} = -\Psi_i'(\theta_{i+1}) \varphi(\theta_{i+1}) + \Psi_i'(\theta_i) \varphi(\theta_i) - a_i \langle \mathbf{1}\{(\theta_i, \theta_{i+1})\}, \varphi \rangle_{L_2}$$

$$= \frac{a_i}{2}\Delta_i(\varphi(\theta_i) + \varphi(\theta_{i+1})),$$

since

$$\int_{\theta_i}^{\theta_{i+1}} \varphi(x)dx = \sum_{j=1}^{n} f(\theta_{i+1} - x_j) - \sum_{j=1}^{n} f(\theta_i - x_j) = 0.$$

Hence,

$$\langle \Psi, \varphi'' \rangle_{L_2} = \sum_{i=1}^{K} \frac{a_i}{2}\Delta_i(\varphi(\theta_i) + \varphi(\theta_{i+1})).$$

In particular, assuming that $a_i \geq 0$ for all $i \in [K]$, using (4) gives that

$$\langle \Psi, \varphi'' \rangle_{L_2} \geq \sum_{i=1}^{K} \frac{a_i}{2}\Delta_i(N_i + N_{i+1})f'(0).$$

Furthermore, since $\Psi_i \in [0, \frac{a_i\Delta_i^2}{8}]$, we have that

$$\left\langle \sum_{i=1}^{K} a_i\Delta_i^2 \mathbf{1}\{(\theta_i, \theta_{i+1})\}, |\varphi''|_+ \right\rangle_{L_2} \geq 4\sum_{i=1}^{K} a_i\Delta_i(N_i + N_{i+1})f'(0).$$

Since $\sum_{j=1}^{n} |f'''(x - x_j)|_+ \geq |\varphi''|_+$, we have that

$$\sum_{j=1}^{n} \left\langle \sum_{i=1}^{K} a_i\Delta_i^2 \mathbf{1}\{(\theta_i, \theta_{i+1})\}, |f'''(x - x_j)|_+ \right\rangle_{L_2} \geq 4\sum_{i=1}^{K} a_i\Delta_i(N_i + N_{i+1})f'(0),$$

or equivalently,

$$\sum_{j=1}^{n} \left\langle \sum_{i=1}^{K} a_i\Delta_i \mathbf{1}\{(\theta_i, \theta_{i+1})\}, |f'''(x - x_j)|_+ \right\rangle_{L_2} \geq 4\sum_{i=1}^{K} a_i(N_i + N_{i+1})f'(0). \tag{5}$$

By Lemma 3.1, we can have that $\tau_{\max}(x_1, \ldots, x_n)$ is greater than τ, but all other gaps must be at most τ. For brevity, we use $\tau_{\max}$ to denote $\tau_{\max}(x_1, \ldots, x_n)$ in the remainder of the proof.

In (5), set $a_i = \frac{1}{\Delta_i}$ to obtain

$$\sum_{j=1}^{n} \left\langle \sum_{i=1}^{K} \mathbf{1}\{(\theta_i, \theta_{i+1})\}, |f'''(x - x_j)|_+ \right\rangle_{L_2} \geq 4\sum_{i=1}^{K} \frac{N_i + N_{i+1}}{\Delta_i} f'(0)$$

$$\Leftrightarrow n \langle 1, |f'''|_+ \rangle_{L_2} \geq 4\sum_{i=1}^{K} \frac{N_i + N_{i+1}}{\Delta_i} f'(0).$$

Assume that the gap $\tau_{\max}$ is between θ_ℓ and $\theta_{\ell+1}$. Since $\Delta_i \leq \tau$ for $i \in [K]\backslash\{\ell\}$,

$$n \langle 1, |f'''|_+ \rangle_{L_2} \geq 4 \left(\sum_{i \in [K]\backslash\{\ell\}} \frac{N_i + N_{i+1}}{\tau} + \frac{N_\ell + N_{\ell+1}}{\tau_{\max}} \right) f'(0)$$

$$= 4\left(\frac{2n}{\tau} - (N_\ell + N_{\ell+1}) \cdot \left(\frac{1}{\tau} - \frac{1}{\tau_{\max}}\right)\right) f'(0).$$

The result follows after noting that $N_\ell + N_{\ell+1} \leq n$ because $K \geq 2$. ■

Corollary 3.3. *Let $\tau \in (0, \pi]$ be such that $f'(x) < 0$ for all $x \notin [-\tau, \tau]$. Assume that M is a positive real number such that for all stable, stationary, and non-synchronized points $\mathbf{x}$ of (S1), $\tau_{max}(\mathbf{x}) < M$. If*

$$\tau \int_{-\pi}^{\pi} |f'''(x)|_+ dx \leq 4\left(1 + \frac{\tau}{M}\right) f'(0),$$

then every stable stationary point $(x_1, \ldots, x_n)$ of system (S1) on $\mathbb{T}^n$ is synchronized, i.e. $x_1 = \cdots = x_n$.

Note that Corollary 3.3 is more general than Theorem 2.1 since $M = 2\pi$ is always a valid choice.

Proof. From Lemma 3.2, if $\mathbf{x} = (x_1, \ldots, x_n)$ is not synchronized, then

$$\langle 1, |f'''|_+ \rangle_{L_2} \geq \min\left(\frac{8}{\tau}, \frac{4}{\tau} + \frac{4}{\tau_{\max}(\mathbf{x})}\right) f'(0) > 4\left(\frac{1}{\tau} + \frac{1}{M}\right) f'(0)$$

because $\tau_{\max}(\mathbf{x}) < M$, which is a contradiction. ■

Remark 3.4. In many cases, such as the Kuramoto model and self-attention dynamics, it is straightforward to show that $\tau_{\max}(\mathbf{x}) < \pi$ for $\mathbf{x}$ to be stable, stationary, and non-synchronized, which leads to an improvement upon Theorem 2.1 by setting $M = \pi$ in Corollary 3.3. For examples of such results, see [17, Lemma 6.4] and [11, Lemma 10].

Corollary 3.5. *Assume that $f'(0) \geq 0$ and $\tau \langle 1, |f'''|_+ \rangle_{L_2} \leq 4\left(1 + \frac{\tau}{2\pi}\right) f'(0)$. Suppose $a, b \in \mathbb{R}$ satisfy $ab \geq 0$ and a and b are not both zero. Assume that*

$$\sum_{j=1}^{n} af(x_i - x_j) - bf(x_j - x_i) = 0, \, \forall i \in [n] \tag{6}$$

and (C2) are satisfied at $(x_1, \ldots, x_n)$. Then, $x_i = x_j$ for all $i, j \in [n]$.

Proof. Let $g(x) = |a|f(x) - |b|f(-x)$. Then, $g'(x) = |a|f'(x) + |b|f'(-x)$ and $g'''(x) = |a|f'''(x) + |b|f'''(-x)$. We have that g satisfies (C1) and (C2) at $(x_1, \ldots, x_n)$. Furthermore, $g'(0) = (|a| + |b|)f'(0) \geq 0$ and $|g'''(x)|_+ \leq |a||f'''(x)|_+ + |b||f'''(-x)|_+$ so $\tau \langle 1, |g'''|_+ \rangle_{L_2} \leq \tau(|a| + |b|) \langle 1, |f'''|_+ \rangle_{L_2} \leq 4\left(1 + \frac{\tau}{2\pi}\right) g'(0)$. The result follows from Theorem 2.1. ■

4 Gradient ascent systems

The goal of this section is to prove Theorem 2.3, which characterizes the asymptotic behavior of (adjusted) mean-field gradient ascent systems. Following the statement of the theorem, assume that $f(x) = \sin(x)h(\cos(x))$ for some real-analytic function h on an open set containing $[-1, 1]$; recall that $g : (\mathbb{S}^1)^n \to \mathbb{R}_{>0}$ is smooth.

As explained in Lemma 2.2, gradient ascent almost always converges to a stationary point. The key idea of the proof is to use Lemma 2.2 to argue that the stationary point almost always has negative semi-definite Hessian and then apply Theorem 2.1 to characterize the stationary point as being synchronized. As part of the proof, we show that a stationary point with negative semi-definite Hessian satisfies conditions (C1) and (C2) so that we can apply Theorem 2.1.

We rewrite (S2) in terms of points on the manifold $(\mathbb{S}^1)^n$. Note that this is the setting that [17, 16] originally consider. For $i \in [n]$ and $t \geq 0$, let

$$\mathfrak{p}_i(t) = (\cos(x_i(t)), \sin(x_i(t))) \in \mathbb{S}^1.$$

For a point $\mathfrak{p}_i \in \mathbb{S}^1$ let us introduce $\mathfrak{p}_i^\perp$ to be (the unique) positively oriented unit vector in the tangent space at $\mathfrak{p}_i$. Let us denote by $P_{\mathfrak{p}_i^\perp}(\cdot)$ the linear operator orthogonally projecting $\mathbb{R}^2$ onto the span of $\mathfrak{p}_i^\perp$. With this notation we can rewrite (S2) as

$$\dot{\mathfrak{p}}_i(t) = \frac{1}{g_i(\mathfrak{p}(t))} \sum_{j=1}^n h(\langle \mathfrak{p}_i(t), \mathfrak{p}_j(t) \rangle) P_{\mathfrak{p}_i(t)^\perp}(\mathfrak{p}_j(t)) \, \forall i \in [n], \tag{S2'}$$

where $\mathfrak{p}(t) = (\mathfrak{p}_1(t), \ldots, \mathfrak{p}_n(t)) \in (\mathbb{S}^1)^n$ for $t \geq 0$. To see the equivalence, note that $P_{\mathfrak{p}_i(t)^\perp}\mathfrak{p}_j(t) = -\sin(x_i(t) - x_j(t))\mathfrak{p}_i(t)^\perp$ and $\langle \mathfrak{p}_i(t), \mathfrak{p}_j(t) \rangle = \cos(x_i(t) - x_j(t))$.

Next, we rewrite (S2') as the gradient of a function over a Riemannian manifold. Suppose $\varphi(x) = \int_0^x h(x)dx$, which is also an analytic function. Let

$$E(x_1, \ldots, x_n) = \frac{1}{2} \sum_{i,j=1}^n \varphi(\langle x_i, x_j \rangle).$$

In order to describe the system as gradient ascent, we construct a Riemannian manifold such that the gradient computed with respect to its metric is the dynamics of (S2'). We follow the ideas of [17, Section 3.4] to state and prove the following lemma.

Lemma 4.1. *Suppose the Riemannian manifold M over $(\mathbb{S}^1)^n$ has positive-definite inner product $\langle (a_1, \ldots, a_n), (b_1, \ldots, b_n) \rangle_P = \sum_{i=1}^n \alpha_i(P) a_i b_i$ at $P = (p_1, \ldots, p_n) \in (\mathbb{S}^1)^n$. Then,*

$$(\nabla_M)_i E(P) = \frac{1}{\alpha_i(P)} \sum_{j=1}^n h(\langle p_i, p_j \rangle) P_{p_i^\perp}(p_j) \, \forall i \in [n].$$

Proof. Let Y be a vector field over $(\mathbb{S}^1)^n$ with gradient flow Φ_Y^t. Furthermore, suppose the vector field B satisfies

$$B_i(P) = \frac{1}{\alpha_i(P)} \sum_{j=1}^n h(\langle p_i, p_j \rangle) P_{P_i^\perp}(P_j)$$

for $i \in [n]$. Then, it suffices to prove that

$$\left.\frac{d}{dt}\right|_{t=0} E(\Phi_Y^t(P)) = \langle Y(P), B(P)\rangle_P .$$

By considering a linear basis over $T_P(\mathbb{S}^1)^n$, we only need to show this holds when $Y(P) = (Ap_1, 0, \ldots, 0)$ for a non-zero skew symmetric A. In this case, $\Phi_Y^t(P) = (e^{At}p_1, p_2, \ldots, p_n)$, so

$$E(\Phi_Y^t(P)) = \sum_{j\neq 1} \varphi(\langle e^{At}p_1, p_j\rangle) + \frac{1}{2}\varphi(\langle e^{At}p_1, e^{At}p_1\rangle) + C,$$

where C does not depend on t. Thus,

$$\frac{d}{dt}E(\Phi_Y^t(P)) = \sum_{j\neq 1} h(\langle e^{At}p_1, p_j\rangle)\left\langle Ae^{At}p_1, p_j\right\rangle + h(\langle e^{At}p_1, e^{At}p_1\rangle)\left\langle Ae^{At}p_1, p_1\right\rangle$$

$$\Rightarrow \left.\frac{d}{dt}\right|_{t=0} E(\Phi_Y^t(P)) = \sum_{j=1}^{n} h(\langle p_1, p_j\rangle)\langle Ap_1, p_j\rangle .$$

It suffices to prove that

$$\sum_{j=1}^{n} h(\langle p_1, p_j\rangle)\langle Ap_1, p_j\rangle = \langle Y(P), B(P)\rangle_P = \left\langle Ap_1, \sum_{j=1}^{n} h(\langle p_1, p_j\rangle)P_{p_1^\perp}(p_j)\right\rangle .$$

Hence, it suffices to prove that for all $v \in \mathbb{S}^1$,

$$\langle Ap_1, v\rangle = \left\langle Ap_1, P_{p_1^\perp}(v)\right\rangle = \langle Ap_1, v - \langle v, p_1\rangle p_1\rangle ,$$

which would be implied by $\langle Ap_1, p_1\rangle = 0$. Observe that $\langle Ap_1, p_1\rangle = p_1^\top Ap_1$, which equals 0 because A is skew-symmetric. ■

Lemma 2.2 implies the almost always convergence to a critical point of E with a negative semidefinite Hessian. However, as explained in [17, Remark B.1], when analyzing whether the Hessian is negative semidefinite at a critical point, any two metrics are equivalent. We state this classical idea in the following lemma, and afterwards, we use it to prove the main result.

Lemma 4.2. *Suppose $R_1 = ((\mathbb{S}^{d-1})^n, g_1)$ and $R_2 = ((\mathbb{S}^{d-1})^n, g_2)$ are Riemannian manifolds. Let P be a critical point of the analytic function $\gamma : (\mathbb{S}^{d-1})^n \to \mathbb{R}$. For $1 \leq i \leq 2$, let H_i be the Hessian of γ at P with respect to R_i. Suppose $v \in T_P((\mathbb{S}^{d-1})^n)$. Then, $\langle H_1 v, v\rangle_{R_1} > 0 \Leftrightarrow \langle H_2 v, v\rangle_{R_2} > 0$.*

Remark 4.3. The set of critical points for both metrics are the same. We abuse notation and assume that H_i is a matrix expressing the Hessian in terms of an orthonormal basis for the metric R_i at P, and thus write $v^\top H_i v$ instead of $\langle H_i v, v\rangle_{R_i}$, implying that $v \in T_P$ is itself expressed as a column vector in the orthonormal basis with respect to R_i at P.

Proof of Theorem 2.3. For this proof, we are working in the setting of points on $(\mathbb{S}^1)^n$ that we have introduced in this section. Let R be the Riemannian manifold over $(\mathbb{S}^1)^n$ with positive-definite inner product $\langle (a_1,\dots,a_n),(b_1,\dots,b_n)\rangle_P = \sum_{i=1}^n g_i(P)a_ib_i$ at P. Then, from Lemma 4.1 with $\alpha_i = g_i$ for all $i \in [n]$,

$$\dot{\mathfrak{p}}(t) = \nabla_R E(\mathfrak{p}(t))$$

in (S2'). Applying Lemma 2.2 gives that we have almost-sure convergence to a critical point $P = (p_1,\dots,p_n)$ with negative semi-definite Hessian. Then, if H is the Hessian of E at P with respect to R and expressed in terms of an orthonormal basis for the metric $\langle\cdot,\cdot\rangle_P$, we assume that there does not exist $v \in T_P((\mathbb{S}^1)^n)$ such that $v^\top Hv > 0$.

First, observe that for all $i \in [n]$,

$$\sum_{j=1}^n h(\langle p_i, p_j\rangle) P_{p_i^\perp}(p_j) = 0$$

by the definition of a critical point of (S2'), which can in turn be rewritten as $\sum_j h(\cos(x_i - x_j))\sin(x_j - x_i) = \sum_j f(x_j - x_i) = 0$, thus verifying condition (C1).

Let S be the Riemannian manifold over $(\mathbb{S}^1)^n$ with the standard inner product. Let H_S be the Hessian of E at P with respect to S and expressed in terms of the standard orthonormal basis. By applying Lemma 4.2 with the Riemannian manifolds R and S and the function E as γ, because there does not exist $v \in T_P(\mathbb{S}^1)^n)$ such that $v^\top Hv > 0$, there does not exist $v \in T_P((\mathbb{S}^1)^n)$ such that $v^\top H_S v > 0$.

The next step is to show that P, when written as an element of $\mathbb{T}^n$, is cut-stable, cf. (C2), so that we can apply Theorem 2.1. For this purpose, we follow the method of [17, Appendix A].

For $t \geq 0$, define

$$p(t) = [e^{c_iBt}p_i]_{i\in[n]},$$

where $B = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ is skew-symmetric. First, observe that

$$E(p(t)) = \sum_{i,j\in[n],i\neq j} \varphi(\langle e^{c_iBt}p_i, e^{c_jBt}p_j\rangle) + C,$$

where C does not depend on t.

Suppose we let $v = \frac{d}{dt}\big|_{t=0} p(t)$. Observe that $v_i = c_iBp_i$ for all $i \in [n]$, so $v \in T_P((\mathbb{S}^1)^n)$. Then,

$$\frac{d^2}{dt^2}\Big|_{t=0} E(p(t)) = v^\top H_S v \leq 0. \tag{7}$$

Next, where $h(x) = \frac{d}{dx}\varphi(x)$,

$$\frac{d}{dt}E(p(t)) = \sum_{i,j\in[n],i\neq j} h(\langle e^{c_iBt}p_i, e^{c_jBt}p_j\rangle)\left(\langle c_iBe^{c_iBt}p_i, e^{c_jBt}p_j\rangle + \langle e^{c_iBt}p_i, c_jBe^{c_jBt}p_j\rangle\right).$$

Furthermore,

$$\frac{d^2}{dt^2}E(p(t)) = \sum_{i,j\in[n],i\neq j}$$
$$\Bigg[h'(\langle e^{c_iBt}p_i, e^{c_jBt}p_j\rangle)\left(\langle c_iBe^{c_iBt}p_i, e^{c_jBt}p_j\rangle + \langle e^{c_iBt}p_i, c_jBe^{c_jBt}p_j\rangle\right)^2$$
$$+ h(\langle e^{c_iBt}p_i, e^{c_jBt}p_j\rangle) \times \Bigg(\langle c_i^2B^2e^{c_iBt}p_i, e^{c_jBt}p_j\rangle + \langle c_iBe^{c_iBt}p_i, c_jBe^{c_jBt}p_j\rangle$$
$$+ \langle e^{c_iBt}p_i, c_j^2B^2e^{c_jBt}p_j\rangle\Bigg)\Bigg]$$

For $1 \leq i \leq n$, let x_i be the unique element of $\mathbb{T}$ such that $p_i = (\cos(x_i), \sin(x_i))$. Because $B^2 = -I_2$, $\langle B^2p_i, p_j\rangle = -\cos(x_i - x_j)$. Since B is the rotation by 90° matrix, $\langle Bp_i, p_j\rangle = \cos(x_i + 90 - x_j) = \sin(x_j - x_i)$. Thus,

$$\left.\frac{d^2}{dt^2}\right|_{t=0} E(p(t))$$
$$= \sum_{i,j\in[n],i\neq j} (c_i - c_j)^2(-\cos(x_i - x_j)h(\cos(x_i - x_j)) + \sin(x_i - x_j)^2h'(\cos(x_i - x_j))).$$

Since $\left.\frac{d^2}{dt^2}\right|_{t=0}E(p(t)) \leq 0$ by (7),

$$\sum_{i,j\in[n],i\neq j} (c_i - c_j)^2(\cos(x_i - x_j)h(\cos(x_i - x_j)) - \sin(x_i - x_j)^2h'(\cos(x_i - x_j))) \geq 0.$$

Observe that because $f(x) = \sin(x)h(\cos(x))$, $f'(x) = \cos(x)h(\cos(x)) - \sin(x)^2h'(\cos(x))$. Therefore,

$$\sum_{i,j\in[n]} (c_i - c_j)^2f'(x_i - x_j) \geq 0,$$

so (C2) is satisfied by setting $c_i = \mathbf{1}\{i \in S\}$ for $S \subset [n]$. Then, by Theorem 2.1, the x_i are synchronized, which finishes the proof. ■

Remark 4.4. Observe that we have shown that if the critical point P has negative semi-definite Hessian, then

$$\sum_{i,j\in[n]} (c_i - c_j)^2f'(x_i - x_j) \geq 0$$

for all $c_i, c_j \in \mathbb{R}$. This result is well-known, but we include the computations for completeness. The other direction is true as well, since for any $v \in T_P((\mathbb{S}^1)^n)$, we have that $v_i = c_iBp_i$ for some $c_i \in \mathbb{R}$ for all $i \in [n]$.

5 Application to self-attention dynamics

In this section, we prove Corollary 2.4, Corollary 2.5 and Corollary 2.6.

Proof of Corollary 2.4. For $\beta \geq 0$, system (T1) is equivalent to (S1) with $f(x) = \sin(x)$ $e^{\beta(\cos(x)-1)}$ and $g = 1$. By Theorem 2.3, it suffices to show that $\tau \langle 1, |f'''|_+ \rangle < 4f'(0)$. We prove this by considering cases for β.

First, observe that

$$
\begin{aligned}
f'(x) &= (\cos(x) - \beta \sin(x)^2) e^{\beta(\cos(x)-1)}, \\
f''(x) &= (-\sin(x) - 3\beta \sin(x)\cos(x) + \beta^2 \sin(x)^3) e^{\beta(\cos(x)-1)}, \\
f'''(x) &= -(\cos(x) + 3\beta \cos(x)^2 - 4\beta \sin(x)^2 - 6\beta^2 \cos(x)\sin(x)^2 + \beta^3 \sin(x)^4) e^{\beta(\cos(x)-1)}.
\end{aligned}
$$

We reference these expressions later. Furthermore, it is clear that we can set

$$
\tau = \arccos\left(\frac{\sqrt{1 + 4\beta^2} - 1}{2\beta} \right).
$$

Case of $\beta > \frac{1}{3}$. The positive regions of f''' over $(-\pi, \pi]$ are $(-a, -b) \cup (b, a)$ for some a, b such that $0 < a < b < \pi$, see Lemma A.1. Since f''' is even, it suffices to prove that

$$
\langle \tau, \mathbf{1}\{(-a, -b) \cup (b, a)\} f''' \rangle < 4 \Leftrightarrow \tau(f''(a) - f''(b)) < 2,
$$

which follows from Lemma A.4, Lemma A.5, Lemma A.6, and Lemma A.7.

Case of $0 < \beta \leq \frac{1}{3}$. The positive region of f''' over $[0, 2\pi)$ is $(a, 2\pi - a)$ for some $a \in (0, \frac{\pi}{2})$, see Lemma B.1. Then, it suffices to prove that

$$
\langle \tau, \mathbf{1}\{(a, 2\pi - a)\} f''' \rangle < 4 \Leftrightarrow \tau f''(a) > -2,
$$

which is proved in Lemma B.4.

Case of $\beta = 0$. This corresponds to the Kuramoto model. In this case, $f'(x) = \cos(x)$ and $\tau = \frac{\pi}{2}$, because if $x \in (\frac{\pi}{2}, \frac{3\pi}{2})$ then $f'(x)$ is negative. Then, $f''(x) = -\sin(x)$ and $f'''(x) = -\cos(x)$ has positive region $(\frac{\pi}{2}, \frac{3\pi}{2})$ in $[0, 2\pi)$. Using the notation for the $\beta \in (0, \frac{1}{3}]$ case, $a = \frac{\pi}{2}$ and it suffices to prove that $\tau f''(a) > -2$, which is true because $\tau f''(a) = -\frac{\pi}{2} > -2$.

Case of $-0.16 \leq \beta < 0$. From Lemma C.4, we may set $M = \pi$ in Corollary 3.3. Afterwards, we prove that global synchronization occurs in (T1) by applying Corollary 3.3 with $M = \pi$ and $g_i = 1$ for all i. By setting $g_i(x_1, \ldots, x_n) = \sum_{j=1}^n e^{\beta \cos(x_i - x_j)}$, we also show that global synchronization occurs in (T2).

By Corollary 3.3, it suffices to prove that $\tau \langle 1, |f'''|_+ \rangle \leq 4\left(1 + \frac{\tau}{\pi}\right)$. The positive region of f''' over $(-\pi, \pi]$ is $(a, 2\pi - a)$ for some $a \in (\frac{\pi}{2}, \pi)$, see Lemma C.1. Since f''' is even, it suffices to prove that

$$
\langle \tau, \mathbf{1}\{(a, 2\pi - a)\} f''' \rangle \leq 4\left(1 + \frac{\tau}{\pi}\right) \Leftrightarrow \tau f''(a) \geq 2\left(1 + \frac{\tau}{\pi}\right),
$$

which follows from Lemma C.5. ■

Proof of Corollary 2.5. This follows from the same argument as the proof of Corollary 2.4 but with $g_i(x_1, \ldots, x_n) = \sum_{j=1}^{n} e^{\beta \cos(x_i - x_j)}$ for $i \in [n]$ in (S2), which is a smooth function. ■

Proof of Corollary 2.6. From Lemma C.6, if $\beta < -\frac{2}{3}$ and n is divisible by three or n is sufficiently large, then there exists a stable nonsynchronized stationary point. At this point, the Hessian has one zero eigenvalue which corresponds to translating each point by the same displacement and its other eigenvalues are negative. This implies that global synchronization does not occur. ■

6 Generalized system

One of the extensions following Kuramoto's work was introduced by [33] in the following form:

$$\dot{x}_i(t) = -\sum_{j=1}^{n} a_{i,j} f(x_i(t) - x_j(t)), \forall i \in [n], \tag{8}$$

where $A = (a_{i,j})_{1 \le i,j \le n} \in \mathbb{R}^{n \times n}$ is a weight matrix and $f : \mathbb{T} \to \mathbb{R}$ is an interaction function. Taylor [33] showed synchronization result for $f(x) = \sin(x)$ and A being an adjacency matrix of an (undirected) graph with each vertex having degree $\ge 0.94n$. Subsequent works eventually improved the lower bound on degree to $0.75n$ [22], while also showing graphs with each vertex of degree $\ge 0.6838n$, which do not synchronize [37]. It is conjectured that there exists graphs with min-degree approaching $0.75n$ which do not synchronize. Furthermore, expander graphs have been utilized to show that generating A using a random process of adding edges leads to global synchronization once A is connected [1, 20].

In this section, we will extend our criterion to the special case of rank-1 matrices A. Initially, we will only consider the following version:

$$\dot{x}_i(t) = -\sum_{j=1}^{n} c_j f(x_i(t) - x_j(t)), \forall i \in [n], \tag{S3}$$

where $c_j > 0$ for $1 \le j \le n$. This system generalizes (S1) by allowing different weights for each particle. We now state the analogous stationarity and cut-stability conditions.

Suppose $\mathbf{x} = (x_i)_{1 \le i \le n} \in \mathbb{T}^n$. For the system (S3) a point is *stationary* iff

$$\sum_{j=1}^{n} c_j f(x_i - x_j) = 0, \ \forall i \in [n]. \tag{C3}$$

We say that point $\mathbf{x}$ is *cut-stable* if

$$\sum_{i \in S,\, j \in S^C} c_j f'(x_i - x_j) \ge 0 \tag{C4}$$

for all $S \subset [n]$ such that the value of x_i is the same for all $i \in S$.

We state the following result, which generalizes Corollary 3.3. Note that Theorem 2.1 is an implication of this corollary.

Theorem 6.1. *Assume that M is a positive real number such that for all stable, stationary, and non-synchronized points $\mathbf{x}$ of (S3), $\tau_{max}(\mathbf{x}) < M$. If*

$$\tau \int_{-\pi}^{\pi} |f'''(x)|_+ dx \leq 4\left(1 + \frac{\tau}{M}\right) f'(0),$$

then every stationary and stable point $(x_1, \ldots, x_n)$ of system (S3) on $\mathbb{T}^n$ is synchronized, i.e. $x_1 = \cdots = x_n$, where τ is as in Theorem 2.1.

Proof. The same proof of Theorem 2.1 in Section 3 can be used, except with $\Psi(x) \triangleq \sum_{j=1}^{n} c_j f'(x - x_j)$ and $W_i \triangleq \sum_{j \in [n]: x_j = \theta_i} c_j$ replacing N_i. Similarly, we only require (C4) for the proof. ■

It is not immediately clear that global synchronization occurs in this setting, since we no longer have an obvious gradient ascent structure. First, we normalize (S3).

Assume that $g : \mathbb{T}^n \to \mathbb{R}^n_{>0}$ is smooth. Then, we can normalize the system as

$$\dot{x}_i(t) = -\frac{1}{g_i(x_1(t), \ldots, x_n(t))} \sum_{j=1}^{n} c_j f(x_i(t) - x_j(t)),\ 1 \leq i \leq n, \tag{S4}$$

which allows us to state the following result, which generalizes Theorem 2.3 and is stated in the format of Corollary 3.3.

Theorem 6.2. *Assume that $f(x) = \sin(x)h(\cos(x))$, where h is a real-analytic function on an open set containing $[-1, 1]$. Assume that M is a positive real number such that for all stable, stationary, and non-synchronized points $\mathbf{x}$ of (S4), $\tau_{max}(\mathbf{x}) < M$. Furthermore, assume that $\tau \langle 1, |f'''|_+ \rangle \leq 4\left(1 + \frac{\tau}{M}\right) f'(0)$, where τ is as in Theorem 2.1. Then, global synchronization occurs in (S2).*

Similarly to the approach of Section 4, we express the dynamical system in terms of points on $(\mathbb{S}^1)^n$. For $i \in [n]$ and $t \geq 0$, we let $p_i(t) = (\cos(x_i(t)), \sin(x_i(t)))$ to obtain the equivalent dynamical system

$$\dot{\mathfrak{p}}_i(t) = \frac{1}{g_i(\mathfrak{p}(t))} \sum_{j=1}^{n} c_j h(\langle \mathfrak{p}_i(t), \mathfrak{p}_j(t) \rangle) P_{\mathfrak{p}_i(t)^\perp}(\mathfrak{p}_j(t))\ \forall i \in [n], \tag{S4'}$$

where $\mathfrak{p}(t) = (\mathfrak{p}_1(t), \ldots, \mathfrak{p}_n(t)) \in (\mathbb{S}^1)^n$ for $t \geq 0$.

Suppose $\varphi(x) = \int_0^x h(x)dx$ and let

$$E_w(x_1, \ldots, x_n) = \frac{1}{2} \sum_{i,j=1}^{n} c_i c_j \varphi(\langle x_i, x_j \rangle).$$

Note that this energy function is also considered in [11]. The idea is that the c_i correspond to the weights of the particles. We have the following generalization of Lemma 4.1. The result can be proved using the same approach.

Lemma 6.3. *Suppose the Riemannian manifold M over $(\mathbb{S}^1)^n$ has positive-definite inner product $\langle (a_1,\dots,a_n),(b_1,\dots,b_n)\rangle_P = \sum_{i=1}^n \alpha_i(P)a_ib_i$ at $P=(p_1,\dots,p_n)\in(\mathbb{S}^1)^n$. Then,*

$$(\nabla_M)_i E_w(P) = \frac{1}{\alpha_i(P)}\sum_{j=1}^n c_ic_jh(\langle p_i,p_j\rangle)P_{p_i^\perp}(p_j)\,\forall i\in[n].$$

Proof of Theorem 6.2. The same proof as the proof of Theorem 2.3 can be used. The only differences are as follows. The Riemannian manifold R over $(\mathbb{S}^1)^n$ has positive-definite inner product $\langle (a_1,\dots,a_n),(b_1,\dots,b_n)\rangle_P = \sum_{i=1}^n c_ig_i(P)a_ib_i$ at P. Of course, (S4') replaces (S2'), and we implement the remaining analogous replacements; for example, we replace (C1) and (C2) with (C3) and (C4), respectively, as well as Theorem 2.1 with Theorem 6.1.

When verifying that (C4) is true, the final expression we obtain is that for $S\subset[n]$ such that the x_i are all equal to θ for $i\in S$,

$$\sum_{i\in S,j\notin S} c_ic_jf'(x_i-x_j)\geq 0 \Leftrightarrow \left(\sum_{i\in S}c_i\right)\sum_{j\notin S}c_jf'(\theta-x_j)\geq 0.$$

Note that (C4) is clearly true when S is empty. If S is nonempty, since the c_i are positive we have that $\sum_{j\notin S}c_jf'(\theta-x_j)\geq 0$, so (C4) holds. ■

An important implication of Theorem 6.2 is the following result, which allows the setting of A as $w_1w_2^\top$ in (8) while still having global synchronization.

Corollary 6.4. *Assume that f satisfies the conditions of Theorem 6.2, where (S4) is replaced by the system*

$$\dot{x}_i(t) = -\sum_{j=1}^n w_{1i}w_{2j}f(x_i(t)-x_j(t)),\ 1\leq i\leq n,$$

where $w_{1i},w_{2i}>0$ for $1\leq i\leq n$ are fixed. Then, global synchronization occurs in this system.

Proof. This follows from Theorem 6.2 with $g_i=w_{1i}^{-1}$ and $c_i=w_{2i}$ for $1\leq i\leq n$. ■

References

[1] Pedro Abdalla, Afonso S. Bandeira, Martin Kassabov, Victor Souza, Steven H. Strogatz, and Alex Townsend. Expander graphs are globally synchronizing. *arXiv preprint arXiv:2210.12788*, 2024.

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

[3] 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.

[4] Vincent D Blondel, Julien M Hendrickx, Alex Olshevsky, and John N Tsitsiklis. Convergence in multiagent coordination, consensus, and flocking. In *Proceedings of the 44th IEEE Conference on Decision and Control*, pages 2996–3000. IEEE, 2005.

[5] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

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

[7] 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.

[8] 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.

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

[10] Henry Cohn and Abhinav Kumar. Universally optimal distribution of points on spheres. *Journal of the American Mathematical Society*, 20(1):99–148, 2007.

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

[12] Gbètondji JS Dovonon, Michael M Bronstein, and Matt J Kusner. Setting the record straight on transformer oversmoothing. *arXiv preprint arXiv:2401.04301*, 2024.

[13] Ruili Feng, Kecheng Zheng, Yukun Huang, Deli Zhao, Michael Jordan, and Zheng-Jun Zha. Rank diminishing in deep neural networks. *Advances in Neural Information Processing Systems*, 35:33054–33065, 2022.

[14] Amic Frouvelle and Jian-Guo Liu. Long-time dynamics for a simple aggregation equation on the sphere. In *International workshop on Stochastic Dynamics out of Equilibrium*, pages 457–479. Springer, 2017.

[15] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

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

[17] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bull. Amer. Math. Soc.*, 62:427–479, 2025.

[18] Seung-Yeal Ha, Dongnam Ko, and Seung-Yeon Ryoo. On the Relaxation Dynamics of Lohe Oscillators on Some Riemannian Manifolds. *Journal of Statistical Physics*, 172:1427–1478, 06 2018.

[19] Ali Jadbabaie, Jie Lin, and A Stephen Morse. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, 48(6):988–1001, 2003.

[20] Vishesh Jain, Clayton Mizgerd, and Mehtaab Sawhney. The random graph process is globally synchronizing. *arXiv preprint arXiv:2501.12205*, 2025.

[21] Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[22] Martin Kassabov, Steven H. Strogatz, and Alex Townsend. Sufficiently dense kuramoto networks are globally synchronizing. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(7):073135, Jul 2021.

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

[24] Yoshiki Kuramoto. Self-entrainment of a population of coupled non-linear oscillators. In *International Symposium on Mathematical Problems in Theoretical Physics*, pages 420–422, Berlin, Heidelberg, 1975. Springer Berlin Heidelberg.

[25] Stanislaw Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les Équations aux Dérivées Partielles*, 117:87–89, 1963.

[26] Johan Markdahl, Johan Thunberg, and Jorge Gonçalves. Almost global consensus on the n -sphere. *IEEE Transactions on Automatic Control*, 63(6):1664–1675, 2018.

[27] Renato E Mirollo and Steven H Strogatz. Synchronization of pulse-coupled biological oscillators. *SIAM Journal on Applied Mathematics*, 50(6):1645–1662, 1990.

[28] 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.

[29] 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.

[30] Han Shi, Jiahui Gao, Hang Xu, Xiaodan Liang, Zhenguo Li, Lingpeng Kong, Stephen M. S. Lee, and James Kwok. Revisiting over-smoothing in BERT from the perspective of graph. In *International Conference on Learning Representations*, 2022.

[31] Michael Shub. *Global Stability of Dynamical Systems*. Springer New York, NY, 1 edition, 2013.

[32] Steven H Strogatz. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D: Nonlinear Phenomena*, 143 (1-4):1–20, 2000.

[33] Richard Taylor. There is no non-zero stable fixed point for dense networks in the homogeneous Kuramoto model. *Journal of Physics A: Mathematical and Theoretical*, 45(5):055102, Jan 2012.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[35] Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. Novel type of phase transition in a system of self-driven particles. *Physical Review Letters*, 75(6):1226, 1995.

[36] Arthur T Winfree. Biological rhythms and the behavior of populations of coupled oscillators. *Journal of Theoretical Biology*, 16(1):15–42, 1967.

[37] Ryosuke Yoneda, Tsuyoshi Tatsukawa, and Jun-nosuke Teramae. The lower bound of the network connectivity guaranteeing in-phase synchronization. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31(6):063124, Jun 2021.

A Results for $\beta > \frac{1}{3}$

In this subsection, $f(x) = \sin(x)e^{\beta(\cos(x)-1)}$.

Lemma A.1. *Suppose $\beta > \frac{1}{3}$. There exists $0 < b < a < \pi$ such that the positive region of $f'''(x)$ is $(-a, -b) \sqcup (b, a)$.*

Proof. Let

$$p(z) = -(z + 3\beta z^2 - 4\beta(1 - z^2) - 6\beta^2 z(1 - z^2) + \beta^3(1 - z^2)^2).$$

Observe that

$$f'''(x) = p(\cos(x))e^{\beta(\cos(x)-1)},$$

so it suffices to analyze the positive regions of p over $[-1, 1]$. First, observe that

$$p(1) = -1 - 3\beta < 0,\ p(1 - \frac{0.3}{\beta}) = 1.688 - 0.1761\beta^{-1} + 0.24\beta > 0,$$
$$p(-1) = 1 - 3\beta < 0,\ p(-1 - \frac{2}{\beta}) = 5\beta + 6\beta^{-1} + 13 > 0.$$

Since $p(-1) < 0$ and $p(-1 - \frac{2}{\beta}) > 0$, p has two roots less than -1. Furthermore, since $p(1) < 0$, $p(1 - \frac{0.3}{\beta}) > 0$, and $p(-1) < 0$, p has two roots in $[-1, 1]$, and p is positive between these two roots. This finishes the proof. ■

Lemma A.2. $\sqrt{x}\arccos(\frac{\sqrt{1+4x^2}-1}{2x})$ *is strictly increasing over* $(0, \infty)$.

Proof. We compute that

$$\frac{d}{dx}\sqrt{x}\arccos(\frac{\sqrt{1+4x^2}-1}{2x}) = \frac{\arccos(\frac{\sqrt{1+4x^2}-1}{2x}) - \frac{\sqrt{\sqrt{1+4x^2}-1}}{\sqrt{2x^2+\frac{1}{2}}}}{2\sqrt{x}}.$$

Therefore, it suffices to prove that

$$\arccos(\frac{\sqrt{1+4x^2}-1}{2x}) > \frac{\sqrt{\sqrt{1+4x^2}-1}}{\sqrt{2x^2+\frac{1}{2}}} \Leftrightarrow \frac{\sqrt{1+4x^2}-1}{2x} < \cos(\frac{\sqrt{\sqrt{1+4x^2}-1}}{\sqrt{2x^2+\frac{1}{2}}}).$$

Observe that using $\cos(z) \geq 1 - \frac{z^2}{2}$ over $(0, \pi)$ yields

$$\cos(\frac{\sqrt{\sqrt{1+4x^2}-1}}{\sqrt{2x^2+\frac{1}{2}}}) \geq 1 - \frac{\sqrt{1+4x^2}-1}{1+4x^2}.$$

Then, it suffices to prove that

$$\begin{aligned}
&\frac{\sqrt{1+4x^2}-1}{2x} < 1 - \frac{\sqrt{1+4x^2}-1}{1+4x^2} \\
\Leftrightarrow &\sqrt{1+4x^2}(\frac{1}{2x} + \frac{1}{1+4x^2}) < 1 + \frac{1}{1+4x^2} + \frac{1}{2x} \\
\Leftrightarrow &\sqrt{1+4x^2}(1+2x+4x^2) < (1+4x^2)2x + 1 + 2x + 4x^2 \\
\Leftrightarrow &(\sqrt{1+4x^2}-1)(1+2x+4x^2) < (1+4x^2)2x \\
\Leftrightarrow &2x(1+2x+4x^2) < (1+\sqrt{1+4x^2})(1+4x^2),
\end{aligned}$$

which is straightforward to verify. ■

The following corollary also appears in [21, Lemma 4].

Corollary A.3. $\arccos(\frac{\sqrt{1+4x^2}-1}{2x}) < \frac{1}{\sqrt{x}}$ *over* $(0, \infty)$.

Proof. This follows from Lemma A.2 and $\lim_{x\to\infty}\sqrt{x}\arccos(\frac{\sqrt{1+4x^2}-1}{2x}) = 1$. ■

Lemma A.4. *If* $\beta \geq 1$ *then* $\tau(f''(a) - f''(b)) < 2$.

Proof. Let $w = 1 - \cos(x)$ and assume that $x \in [0, \pi]$ so that $\sin(x) \geq 0$. Then,

$$\begin{aligned} f''(x) &= \sqrt{2w - w^2}(-1 - 3\beta(1-w) + \beta^2(2w - w^2))e^{-\beta w} \\ &= \sqrt{2w - w^2}(\beta(-3 + 2\beta w) + (-1 + 3\beta w - \beta^2 w^2))e^{-\beta w}. \end{aligned}$$

Let $z = \beta w$. Hence,

$$f''(x) = \sqrt{\beta}\sqrt{z}\sqrt{2 - \frac{z}{\beta}}(-3 + 2z + \frac{1}{\beta}(-1 + 3z - z^2))e^{-z}.$$

Let

$$g_\beta(z) = \sqrt{z}\sqrt{2 - \frac{z}{\beta}}(-3 + 2z + \frac{1}{\beta}(-1 + 3z - z^2))e^{-z}$$

where $z \in [0, 2\beta]$. Note that $f''(x) = \sqrt{\beta} g_\beta(\beta - \beta\cos(x))$ so

$$f'''(x) = \beta\sqrt{\beta}\sin(x) g'_\beta(\beta - \beta\cos(x))$$

Since $x \in [0, \pi]$ and the positive region of f''' in $[0, \pi]$ is (b, a), the positive region of g'_β is $(\beta(1 - \cos(b)), \beta(1 - \cos(a)))$. Particularly, $\frac{f''(a) - f''(b)}{\sqrt{\beta}} = g_\beta(\beta(1 - \cos(a))) - g_\beta(\beta(1 - \cos(b)))$.

First, observe that $g'_\beta(0.18) < 0$ and $g'_\beta(1.4) > 0$ for all $\beta \geq 1$. Therefore, because the positive region is continuous, we always have that

$$0.18 < \beta(1 - \cos(b)) < 1.4 < \beta(1 - \cos(a)).$$

Moreover, $g_\beta(\beta(1-\cos(a))) > 0 > g_\beta(\beta(1-\cos(b)))$; this is because $g_\beta(0) = 0$, $g_\beta(0.1) < 0$, and $g_\beta(1.9) > 0$, so g_β first decreases to a negative value and then increases to a positive value.

Let

$$g^1(z) = \sqrt{2z}(-3 + 2z)e^{-z},\ g^2(z) = \sqrt{2z}(-1 + 3z - z^2)e^{-z}.$$

Then, g_β is similar to a linear combination of g^1 and g^2, with

$$g_\beta(z) = \frac{\sqrt{2 - \frac{z}{\beta}}}{\sqrt{2}}(g^1(z) + \frac{1}{\beta}g^2(z)).$$

For the following computations, we utilize properties of g^1 and g^2 which are straightforward to verify.

Because $0.18 < \beta(1 - \cos(b)) < 1.4$, the value of $\sqrt{2 - \frac{z}{\beta}}g^2(z)$ at $z = \beta(1 - \cos(b))$ is at least

$$\sqrt{2 - \frac{0.18}{\beta}}g^2(0.18) > -0.247\sqrt{2 - \frac{0.18}{\beta}}.$$

Furthermore, the value of $\sqrt{2 - \frac{z}{\beta}}g^1(z)$ at $z = \beta(1 - \cos(b))$ is at least

$$-1.381\sqrt{2 - \frac{0.18}{\beta}},$$

since $g^1 > -1.381$. Therefore, the value of g_β at $z = \beta(1-\cos(b))$ is greater than

$$-0.247\frac{\sqrt{2-\frac{0.18}{\beta}}}{\sqrt{2}\beta} - 1.381\frac{\sqrt{2-\frac{0.18}{\beta}}}{\sqrt{2}}. \tag{9}$$

Since $\beta(1-\cos(a)) > 1.4$, the value of $\sqrt{2-\frac{z}{\beta}}g^2(z)$ at $z = \beta(1-\cos(a))$ is at most

$$\sqrt{2-\frac{1.4}{\beta}}g^2(1.4) < 0.512\sqrt{2-\frac{1.4}{\beta}}.$$

Furthermore, the value of $\sqrt{2-\frac{z}{\beta}}g^1(z)$ at $z = \beta(1-\cos(a))$ is at most

$$0.375\sqrt{2-\frac{1.5}{\beta}},$$

since $g^1 \leq 0$ for $z \leq 1.5$ and $g^1 < 0.375$. Therefore, the value of g_β at $z = \beta(1-\cos(a))$ is less than

$$0.512\frac{\sqrt{2-\frac{1.4}{\beta}}}{\sqrt{2}\beta} + 0.375\frac{\sqrt{2-\frac{1.5}{\beta}}}{\sqrt{2}}.$$

Using this inequality and (9) gives that

$$g_\beta(\beta(1-\cos(a))) - g_\beta(\beta(1-\cos(b))) <$$
$$0.512\frac{\sqrt{2-\frac{1.4}{\beta}}}{\sqrt{2}\beta} + 0.375\frac{\sqrt{2-\frac{1.5}{\beta}}}{\sqrt{2}} + 0.247\frac{\sqrt{2-\frac{0.18}{\beta}}}{\sqrt{2}\beta} + 1.381\frac{\sqrt{2-\frac{0.18}{\beta}}}{\sqrt{2}}.$$

Let

$$\varphi(\beta) = 0.512\frac{\sqrt{2-\frac{1.4}{\beta}}}{\sqrt{2}\beta} + 0.375\frac{\sqrt{2-\frac{1.5}{\beta}}}{\sqrt{2}} + 0.247\frac{\sqrt{2-\frac{0.18}{\beta}}}{\sqrt{2}\beta} + 1.381\frac{\sqrt{2-\frac{0.18}{\beta}}}{\sqrt{2}}$$

for $\beta \geq 1$, so that $g_\beta(\beta(1-\cos(a))) - g_\beta(\beta(1-\cos(b))) < \varphi(\beta)$.

If $\beta \geq 2$ then $\varphi(\beta) < 2$, so

$$\tau(f''(a) - f''(b)) = \sqrt{\beta}\tau(g_\beta(\beta(1-\cos(a))) - g_\beta(\beta(1-\cos(b)))) < \sqrt{\beta}\tau\varphi(\beta) < 2\sqrt{\beta}\tau.$$

Since $\sqrt{\beta}\tau < 1$ by Corollary A.3, $\tau(f''(a) - f''(b)) < 2$.

Assume that $\beta \in [1, 2)$. Then, from Lemma A.2, $\sqrt{\beta}\tau < \sqrt{2}\tau(2) < 0.96$, so $\tau(f''(a) - f''(b)) < 0.96\varphi(\beta) < 2$. ■

Lemma A.5. *Suppose $\beta \in [0.75, 1)$. Then, $\tau(f''(a) - f''(b)) < 2$.*

Proof. Observe that $g'_\beta(0.165) < 0$ and $g'_\beta(1.24) > 0$ for $\beta \in [0.75, 1)$. Thus,

$$0.165 < \beta(1 - \cos(b)) < 1.24 < \beta(1 - \cos(a)).$$

Similarly, $g_\beta(\beta(1 - \cos(a))) > 0 > g_\beta(\beta(1 - \cos(b)))$, because $g_\beta(0) = 0$, $g_\beta(0.1) < 0$, $g_\beta(1.4) > 0$. Therefore,

$$g_\beta(\beta(1 - \cos(a))) - g_\beta(\beta(1 - \cos(b))) <$$
$$0.54\frac{\sqrt{2 - \frac{1.24}{\beta}}}{\sqrt{2}\beta} + 0.375\frac{\sqrt{2 - \frac{1.5}{\beta}}}{\sqrt{2}} + 0.26\frac{\sqrt{2 - \frac{0.165}{\beta}}}{\sqrt{2}\beta} + 1.381\frac{\sqrt{2 - \frac{0.165}{\beta}}}{\sqrt{2}} =: \varphi(\beta)$$

over $[0.75, 1)$. Using Lemma A.2, $\sqrt{\beta}\tau < \tau(1) < 0.91$, so $\tau(f''(a) - f''(b)) < 0.91\varphi(\beta) < 2$. ■

Lemma A.6. *Suppose $\beta \in [0.5, 0.75)$. Then, $\tau(f''(a) - f''(b)) < 2$.*

Proof. Observe that $g'_\beta(0.14) < 0$ and $g'_\beta(0.9) > 0$ for $\beta \in [0.5, 0.75)$. Thus,

$$0.14 < \beta(1 - \cos(b)) < 0.9 < \beta(1 - \cos(a)).$$

Similarly, $g_\beta(\beta(1 - \cos(a))) > 0 > g_\beta(\beta(1 - \cos(b)))$, because $g_\beta(0) = 0$, $g_\beta(0.1) < 0$, $g_\beta(0.99) > 0$. Therefore,

$$g_\beta(\beta(1 - \cos(a))) - g_\beta(\beta(1 - \cos(b))) <$$
$$0.542\frac{\sqrt{2 - \frac{0.9}{\beta}}}{\sqrt{2}\beta} + 0.28\frac{\sqrt{2 - \frac{0.14}{\beta}}}{\sqrt{2}\beta} + 1.381\frac{\sqrt{2 - \frac{0.14}{\beta}}}{\sqrt{2}} =: \varphi(\beta)$$

over $[0.5, 0.75)$; observe that we have removed the term $0.375\frac{\sqrt{2-\frac{1.5}{\beta}}}{\sqrt{2}}$ for g^1, since g^1 is always non-positive when $\beta \le 0.75$ and $z \le 2\beta$. Using Lemma A.2, $\sqrt{\beta}\tau < \sqrt{0.75}\tau(\sqrt{0.75}) < 0.88$, so $\tau(f''(a) - f''(b)) < 0.88\varphi(\beta) < 2$. ■

Lemma A.7. *Suppose $\beta \in (\frac{1}{3}, 0.5)$. Then, $\tau(f''(a) - f''(b)) < 2$.*

Proof. Observe that $g'_\beta(0.121) < 0$ and $g'_\beta(\frac{2}{3}) > 0$ for $\beta \in (\frac{1}{3}, 0.5)$. Thus,

$$0.12 < \beta(1 - \cos(b)) < \frac{2}{3} < \beta(1 - \cos(a)).$$

Similarly, $g_\beta(\beta(1 - \cos(a))) > 0 > g_\beta(\beta(1 - \cos(b)))$, because $g_\beta(0) = 0$, $g_\beta(0.1) < 0$, $g_\beta(\frac{2}{3}) > 0$.

Since $z \le 2\beta < 1$, the maximal positive value of g_2 is less than $g_2(1) < 0.521$. Therefore,

$$g_\beta(\beta(1 - \cos(a))) - g_\beta(\beta(1 - \cos(b))) <$$
$$0.521\frac{\sqrt{2 - \frac{2/3}{\beta}}}{\sqrt{2}\beta} + 0.285\frac{\sqrt{2 - \frac{0.121}{\beta}}}{\sqrt{2}\beta} + 1.381\frac{\sqrt{2 - \frac{0.121}{\beta}}}{\sqrt{2}} =: \varphi(\beta)$$

over $(\frac{1}{3}, 0.5)$; similarly, we have removed the positive term for g^1. Using Lemma A.2, $\sqrt{\beta}\tau < \sqrt{0.5}\tau(0.5) < 0.809$, so $\tau(f''(a) - f''(b)) < 0.809\varphi(\beta) < 2$. ■

B Results for $0 < \beta \leq \frac{1}{3}$

In this subsection, $f(x) = \sin(x)e^{\beta(\cos(x)-1)}$.

Lemma B.1. *Suppose $\beta \in (0, \frac{1}{3}]$. Then, there exists $a \in (0, \frac{\pi}{2})$ such that the nonnegative region of $f'''(x)$ over $[0, 2\pi)$ is $[a, 2\pi - a]$.*

Proof. We use the same method as the proof of Lemma A.1. Observe that

$$p(1) = -1 - 3\beta < 0,\ p(0) = -\beta(\beta^2 - 4) > 0,\ p(-1) = 1 - 3\beta \geq 0,$$
$$p(-1 - \frac{1}{\beta}) = 5\beta - \beta^{-1} + 1 < 0,\ p(-1 - \frac{2}{\beta}) = 5\beta + 6\beta^{-1} + 13 > 0,$$

and $\lim_{z\to-\infty} p(z) = -\infty$. Thus, p has a root in each of the following intervals: $(0, 1)$, $(-1 - \frac{1}{\beta}, -1]$, $(-1 - \frac{2}{\beta}, -1 - \frac{1}{\beta})$, and $(-\infty, -1 - \frac{2}{\beta})$.

Let r be the root of p in $(0, 1)$. Let $a = \arccos(r)$. Because $r \in (0, 1)$, $a \in (0, \frac{\pi}{2})$. Furthermore, because $p(1) < 0$, $p(0) > 0$, and p has no other roots in $(-1, 1)$, $p(z)$ is nonnegative for $z \in [-1, 1]$ if and only if $z \in [-1, r]$. Therefore, the nonnegative region of f''' is $[a, 2\pi - a]$. ■

Remark B.2. In contrast with Lemma A.1, we consider the nonnegative region of f''' rather than the positive region. The reason for this is that when $\beta = \frac{1}{3}$, $p(-1) = 0$, so the positive region would be $(a, \pi) \cup (\pi, 2\pi - a)$ for this case. For simplicity, we consider the nonnegative region.

Corollary B.3. *Suppose $\beta \in (0, \frac{1}{3}]$. Then, $f'' \leq 0$ over $[0, \pi]$ and $f'' \geq 0$ over $[\pi, 2\pi]$.*

Proof. Using Lemma B.1, assume that the nonnegative region of f''' is $[a, 2\pi - a]$ for $a \in (0, \frac{\pi}{2})$. Then, since $f''(0) = f''(\pi) = 0$, we have that over $[0, \pi]$, f'' first decreases from 0 to its minimal value at a and then increases to 0, so f'' is non-positive. Because f'' is odd, f'' is nonnegative over $[\pi, 2\pi]$. ■

Lemma B.4. *Suppose $\beta \in (0, \frac{1}{3}]$. Then, $\tau f''(a) > -2$.*

Proof. We have that

$$f''(a) \geq -(1 + 3\beta)\sin(a)e^{\beta\cos(a)}e^{-\beta} \geq -(1 + 3\beta)\sin(\tau)e^{\beta\cos(\tau)}e^{-\beta}.$$

Thus, it suffices to prove that

$$(1 + 3\beta)\sin(\tau)e^{\beta\cos(\tau)}e^{-\beta}\tau < 2.$$

Observe that

$$0 \leq \cos(\tau) = \beta\sin(\tau)^2 \leq \beta,$$

so it suffices to prove that

$$(1 + 3\beta)e^{\beta^2-\beta}\sin(\tau)\tau < 2.$$

Assume that $0 < \beta \leq 0.148$. Then, $\tau \leq \frac{\pi}{2}$ and $\sin(\tau) \leq 1$. Because

$$(1+3\beta)e^{\beta^2-\beta} \cdot \frac{\pi}{2} < 2,$$

we have that $\tau f''(a) > -2$. (0.148 is approximately the maximal value of β for which this method is correct.)

Assume that $0.148 < \beta \leq 0.228$. Then, $\tau < \tau(0.148) < 1.43$ and $\sin(\tau) < \sin(1.43)$. Because

$$(1+3\beta)e^{\beta^2-\beta} \cdot 1.43\sin(1.43) < 2,$$

we have that $\tau f''(a) > -2$.

Assume that $0.228 < \beta \leq 0.278$. Then, $\tau < \tau(0.228) < 1.36$ and $\sin(\tau) < \sin(1.36)$. Because

$$(1+3\beta)e^{\beta^2-\beta} \cdot 1.36\sin(1.36) < 2,$$

we have that $\tau f''(a) > -2$.

Assume that $0.278 < \beta \leq 0.321$. Then, $\tau < \tau(0.278) < 1.31$ and $\sin(\tau) < \sin(1.31)$. Because

$$(1+3\beta)e^{\beta^2-\beta} \cdot 1.31\sin(1.31) < 2,$$

we have that $\tau f''(a) > -2$.

Assume that $0.321 < \beta \leq \frac{1}{3}$. Then, $\tau < \tau(0.321) < 1.28$ and $\sin(\tau) < \sin(1.28)$. Because

$$(1+3\beta)e^{\beta^2-\beta} \cdot 1.28\sin(1.28) < 2,$$

we have that $\tau f''(a) > -2$. ■

C Results for $-\frac{1}{3} < \beta < 0$

In this subsection, $f(x) = \sin(x)e^{\beta(\cos(x)-1)}$.

Lemma C.1. *Suppose $\beta \in (-\frac{1}{3}, 0)$. Then, there exists $a \in (\frac{\pi}{2}, \pi)$ such that the nonnegative region of $f'''(x)$ over $[0, 2\pi)$ is $[a, 2\pi - a]$.*

Proof. We use the same method as the proof of Lemma A.1. Observe that

$$p(1) = -1 - 3\beta < 0,\ p(0) = -\beta(\beta^2 - 4) < 0,\ p(-1) = 1 - 3\beta > 0,$$
$$p(1 - \frac{1}{\beta}) = 5\beta - \beta^{-1} - 1 > 0,\ p(1 - \frac{2}{\beta}) = 5\beta + 6\beta^{-1} - 13 < 0,$$

and $\lim_{z\to\infty} p(z) = \infty$. Thus, p has a root in each of the following intervals: $(-1, 0)$, $(1, 1 - \frac{1}{\beta})$, $(1 - \frac{1}{\beta}, 1 - \frac{2}{\beta})$, and $(1 - \frac{2}{\beta}, \infty)$.

Let r be the root of p in $(-1, 0)$. Let $a = \arccos(r)$. Because $r \in (-1, 0)$, $a \in (\frac{\pi}{2}, \pi)$. Furthermore, because $p(-1) > 0$, $p(0) < 0$, and p has no other roots in $(-1, 1)$, $p(z)$ is nonnegative for $z \in [-1, 1]$ if and only if $z \in [-1, r]$. Therefore, the nonnegative region of f''' is $[a, 2\pi - a]$. ■

Remark C.2. If $\beta = -\frac{1}{3}$, the nonnegative region of $f'''(x)$ over $[0, 2\pi)$ is $[a, 2\pi - a] \cup \{0\}$, since $p(1) = 0$.

Corollary C.3. *Suppose $\beta \in (-\frac{1}{3}, 0)$. Then, $f'' \leq 0$ over $[0, \pi]$ and $f'' \geq 0$ over $[\pi, 2\pi]$.*

Proof. Using Lemma B.1, assume that the nonnegative region of f''' is $[a, 2\pi - a]$ for $a \in (\frac{\pi}{2}, \pi)$. Then, since $f''(0) = f''(\pi) = 0$, we have that over $[0, \pi]$, f'' first decreases from to 0 to its minimal value at a and then increases to 0, so f'' is non-positive. Because f'' is odd, f'' is nonnegative over $[\pi, 2\pi]$. ■

Lemma C.4. *Suppose $\beta < 0$. If $\mathbf{x}$ is a stable, stationary, and non-synchronized point of (S1), then $\tau_{max}(\mathbf{x}) < \pi$.*

Proof. This is implied by [11, Lemma 10] with $\varphi(t) = -e^{\beta t}$. ■

Lemma C.5. *Suppose $\beta \in [-0.16, 0)$. Then, $\tau f''(a) \geq -2\left(1 + \frac{\tau}{\pi}\right)$.*

Proof. We have that

$$f''(a) \geq -(1 - 3\beta)\sin(a)e^{\beta\cos(a)}e^{-\beta} \geq -(1 - 3\beta)\sin(\tau)e^{\beta\cos(\tau)}e^{-\beta}.$$

Thus, it suffices to prove that

$$(1 - 3\beta)\sin(\tau)e^{\beta\cos(\tau)}e^{-\beta}\tau \leq 2\left(1 + \frac{\tau}{\pi}\right).$$

Observe that

$$0 \geq \cos(\tau) = \beta\sin(\tau)^2 \geq \beta,$$

so it suffices to prove that

$$(1 - 3\beta)e^{\beta^2-\beta}\sin(\tau)\tau \leq 2\left(1 + \frac{\tau}{\pi}\right).$$

Equivalently, it suffices to prove that

$$\tau\left((1 - 3\beta)e^{\beta^2-\beta}\sin(\tau) - \frac{2}{\pi}\right) \leq 2.$$

Assume that $\beta \in [-0.16, 0)$. Observe that both τ and $(1 - 3\beta)e^{\beta^2-\beta} - \frac{2}{\pi}$ increase as β decreases from 0. At $\beta = -0.16$, we have that

$$\tau\left((1 - 3\beta)e^{\beta^2-\beta} - \frac{2}{\pi}\right) \leq 2,$$

so this must be the case for all $\beta \in [-0.16, 0)$. Thus,

$$\tau\left((1 - 3\beta)e^{\beta^2-\beta}\sin(\tau) - \frac{2}{\pi}\right) \leq \tau\left((1 - 3\beta)e^{\beta^2-\beta} - \frac{2}{\pi}\right) \leq 2$$

for all $\beta \in [-0.16, 0)$. ■

The lower bound -0.16 for β in Lemma C.5 can be improved, but we omit this refinement for simplicity. We establish that global synchronization does not occur for all negative values of β.

Lemma C.6. *Suppose $\beta < -\frac{2}{3}$. Then, if n is a multiple of 3 or sufficiently large, there exists a value of $\mathbf{x} \in \mathbb{T}^n$ such that:*

1. *There exists three elements p_1,p_2, and p_3 of $\mathbb{T}$ such that $\lfloor \frac{n}{3} \rfloor$, $\lfloor \frac{n}{3} \rfloor$, and $n - 2\lfloor \frac{n}{3} \rfloor$ points of $\mathbf{x}$ are placed at p_1, p_2, and p_3, respectively.*
2. *The vector $\mathbf{x}$ is a critical point of E and the Hessian of E over $\mathbb{T}^n$ at $\mathbf{x}$ is negative semidefinite, with only one eigenvalue whose eigenvectors are the scalar multiples of the vector $[1, \dots, 1]^\top$.*

Proof. Suppose $n \geq 3$. Let $p_1 = 0$, $p_2 = \alpha$, and $p_3 = 2\pi - \alpha$, where α is an element of $(0, \pi)$ such that

$$\left\lfloor \frac{n}{3} \right\rfloor \sin(2\alpha) e^{\beta \cos(2\alpha)} + \left(n - 2 \left\lfloor \frac{n}{3} \right\rfloor \right) \sin(\alpha) e^{\beta \cos(\alpha)} = 0;$$

as $n \to \infty$, we can set $\alpha = \frac{2\pi}{3} + o_n(1)$ and in particular we can set $\alpha = \frac{2\pi}{3}$ when n is a multiple of 3.

The Hessian of the energy function E is the Laplacian L of $[-f'(\mathbf{x}_i - \mathbf{x}_j)]_{i,j=1}^n$. Recall that $f'(x) = (\cos(x) - \beta \sin(x)^2) e^{\beta(\cos(x)-1)}$ and the Laplacian of a symmetric $n \times n$ matrix A is $D - A$, where D is the diagonal matrix with diagonal $[\sum_{i=1}^n A_{ij}]_{1 \leq j \leq n}$. Furthermore, for a vector v,

$$v^\top L v = -\sum_{i,j=1}^{n} \frac{1}{2} f'(\mathbf{x}_i - \mathbf{x}_j)(v_i - v_j)^2.$$

However, we always have that $|\mathbf{x}_i - \mathbf{x}_j| \in \{0, \alpha, 2\pi - \alpha\}$. Since $\beta < -\frac{2}{3}$ and $\alpha = \frac{2\pi}{3} + o_n(1)$, $f'(x) > 0$ for all $x \in \{0, \alpha, 2\pi - \alpha\}$ when n is sufficiently large; if n is a multiple of 3, we can set $\alpha = \frac{2\pi}{3}$ to obtain that $f'(x) > 0$ for all $x \in \{0, \alpha, 2\pi - \alpha\}$. Thus, condition 2 is satisfied when n is a multiple of 3 or sufficiently large. ■