Dynamic metastability in the self-attention model

Borjan Geshkovski
Inria & Sorbonne Université

Hugo Koubbi
ENS Paris-Saclay & Yale University

Yury Polyanskiy
MIT

Philippe Rigollet
MIT

January 4, 2025

Abstract

We consider the self-attention model—an interacting particle system on the unit sphere, which serves as a toy model for *Transformers*, the deep neural network architecture behind the recent successes of large language models. We prove the appearance of *dynamic metastability* conjectured in [GLPR23]—although particles collapse to a single cluster in infinite time, they remain trapped near a configuration of several clusters for an exponentially long period of time. By leveraging a gradient flow interpretation of the system, we also connect our result to an overarching framework of *slow motion* of gradient flows proposed by Otto and Reznikoff [OR07] in the context of coarsening and the Allen-Cahn equation. We finally probe the dynamics beyond the exponentially long period of metastability, and illustrate that, under an appropriate time-rescaling, the energy reaches its global maximum in finite time and has a staircase profile, with trajectories manifesting *saddle-to-saddle*-like behavior, reminiscent of recent works in the analysis of training dynamics via gradient descent for two-layer neural networks.

Keywords. Transformers, slow motion, metastability, gradient flows, interacting particle systems

AMS classification. 49Q22, 68T07, 82C22, 37D10, 82C26, 82B26.

Contents

1 Introduction

Introduced in 2017 with the seminal paper [VSP$^+$17], *Transformers* are the neural network architectures behind the recent successes of large language models. Their impressive results are in part due to the way they process data: inputs are length-n sequences of d-dimensional vectors called *tokens* (representing words, or patches of an image, for example), which are processed over several layers of parametrized nonlinearities. Unlike conventional neural networks, all tokens are coupled and mixed at every layer via the so-called *self-attention mechanism.*

    We make this discussion more transparent by following the mathematical framework set out in [GLPR23]—itself based on and inspired by [SABP22, LLH$^+$19]—viewing layers as a continuous time variable t, and tokens as particles, we consider

the toy model

$$\dot{x}_i(t) = \mathbf{P}^\perp_{x_i(t)} \sum_{j=1}^n \frac{e^{\beta\langle x_i(t), x_j(t)\rangle}}{\displaystyle\sum_{k=1}^n e^{\beta\langle x_i(t), x_k(t)\rangle}} x_j(t) \qquad \text{for } t \geqslant 0, \qquad \text{(SA)}$$

for $i \in \{1, \ldots, n\}$; here $\mathbf{P}^\perp_x := I_d - xx^\top$ ensures that the particles $x_i(t)$ evolve on the unit sphere $\mathbb{S}^{d-1}$. We dub (SA) the *self-attention model*: it has a single parameter $\beta \geqslant 0$, designating an inverse temperature, is derived from Transformers, and exhibits a remarkably similar qualitative behavior, as touched upon in [GLPR23].

To analyze the dynamics and the long-time behavior of (SA)—referred as *signal propagation* in the machine learning literature [NAB$^+$22, HMZ$^+$23, CNQG24]—one naturally looks for a Lyapunov function. This endeavor is made simpler upon observing that the partition function

$$\mathfrak{Z}_{\beta,i} := \sum_{k=1}^n e^{\beta\langle x_i, x_k\rangle}$$

satisfies $e^\beta \leqslant \mathfrak{Z}_{\beta,i} \leqslant ne^\beta$. Whereupon, one can, to begin with, consider the *unnormalized self-attention model*

$$\dot{x}_i(t) = n^{-1}\mathbf{P}^\perp_{x_i(t)} \sum_{j=1}^n e^{\beta(\langle x_i(t), x_j(t)\rangle - 1)} x_j(t) \qquad \text{for } t \geqslant 0, \qquad \text{(USA)}$$

which is the (time-reversed) gradient flow for the interaction energy

$$\mathsf{E}_\beta(x_1, \ldots, x_n) := \frac{1}{2\beta e^\beta n^2} \sum_{i=1}^n \sum_{j=1}^n e^{\beta\langle x_i, x_j\rangle}. \qquad (1.1)$$

Namely, $X(t) = (x_1(t), \ldots, x_n(t))$ satisfies

$$\dot{X}(t) = \nabla\mathsf{E}_\beta(X(t)) \qquad \text{for } t \geqslant 0.$$

This observation impels one to also view (SA) as a (reverse-time) gradient flow for E_β, but one in which the gradient is computed with respect to a different metric, obtained by weighting the canonical metric on $\mathsf{T}_X(\mathbb{S}^{d-1})^n$ by $\mathfrak{Z}_{\beta,i}$, as done in [GLPR23]. Consequently E_β increases along trajectories of (SA) and (USA).

Global maxima of E_β are configurations $(x_1, \ldots, x_n) \in (\mathbb{S}^{d-1})^n$ satisfying $x_1 = \ldots = x_n$, which we call *clusters*. With this at hand, using established tools from dynamical systems combined with an analysis of the landscape of E_β, the authors in [GLPR23, MTG17] and the subsequent improvement in [CRMB24] conclude that for almost every initial configuration, and for $\beta \geqslant 0$ when $d \geqslant 3$, or $\beta \leqslant 1 \vee \beta \gtrsim n^2$ when $d = 2$, the unique solution to (SA) or (USA) converges to some cluster as $t \to +\infty$. This behavior has in fact been observed in trained Transformer models, and is referred to as *token uniformity, over-smoothing*

[CZC$^+$22, RZZD23, GWDW23, WAWJ24, WAW$^+$24, DBK24, SWJS24], or *rank-collapse* [DCL21, FZH$^+$22, NAB$^+$22, JDB23, ZMZ$^+$23, ZLL$^+$23, NLL$^+$24, BHK24, CNQG24].

One can then ask whether for almost every initial configuration, the above convergence holds with some rate. The answer is affirmative—and the rate is in fact exponential—when the initial configuration lies in an open hemisphere [GLPR23, Lemma 6.4]. The latter is, generically, is a high-dimensional property ($d \gg n$), and the decay constant is itself exponentially small in $\beta \gg 1$. Prompted by empirical evidence and synthetic simulations, the authors in [GLPR23, Problem 1] posit that the dynamics instead manifest *metastability*: particles quickly approach a few clusters, stay in the vicinity of these clusters for a very long period of time, before eventually coalescing to a single cluster in infinite time. Since the appearance of a single cluster in long time is interpreted as a negative property by practitioners in the empirical literature cited above, alluding to a lack of expressivity, we can view metastability as a desideratum. The goal of this paper is to describe and prove the appearance of the metastability phenomenon for both (SA) and (USA).

1.1 Main result

Recall that $f(x) = \Omega(g(x))$ whenever $\liminf_{x\to\infty} f(x)/g(x) > 0$. We work in the following setting of initial configurations.

Definition 1.1 ((β, ε)-separated configurations)**.** *Suppose $d, n \geqslant 2$, $\beta > 1$ and $\varepsilon \in (0, \frac{1}{16})$. We call $(x_i)_{i=1}^n \in (\mathbb{S}^{d-1})^n$ a (β, ε)-separated configuration if there exist $k \leqslant n$ points $w_1, \ldots, w_k \in \mathbb{S}^{d-1}$ such that*

1. For all $i \in \{1, \ldots, n\}$,

$$x_i(0) \in \bigcup_{q=1}^k \mathcal{S}_q(\varepsilon)$$

where $\mathcal{S}_q(\varepsilon)$ is the spherical cap centered at w_q of radius (or "height") $1 - \varepsilon$:

$$\mathcal{S}_q(\varepsilon) := \left\{ x \in \mathbb{S}^{d-1} \colon \langle x, w_q \rangle \geqslant 1 - \varepsilon \right\}. \tag{1.2}$$

2. Furthermore,

$$\gamma(\beta) := 1 - \alpha - 8\varepsilon - \frac{1}{\beta} \log\left(\frac{2n^2}{\varepsilon} \right) > 0 \quad \text{and} \quad \gamma(\beta) = \Omega(1) \tag{1.3}$$

with

$$\alpha := \max_{\substack{(x,y)\in\mathcal{S}_i(2\varepsilon)\times\mathcal{S}_j(2\varepsilon) \\ i\neq j\in\{1,\ldots,k\}}} \langle x, y \rangle. \tag{1.4}$$

4

Condition (1.3) is particularly indicative in the regime where $d, n \geqslant 2$ are fixed and in the low temperature limit $\beta \to +\infty$. In this regime, which is of interest due to the motivating discussion above, we essentially require ε to be small in comparison to $1 - \alpha$. We provide an illustration of such a configuration in Figure 1.

We now state our main result.

Theorem 1.2. *Suppose $d, n \geqslant 2$ and $\beta > 1$. Consider $(x_i(0))_{i=1}^n \in (\mathbb{S}^{d-1})^n$ which is (β, ε)-separated for some $\varepsilon = \varepsilon(\beta) \in (0, \frac{1}{16})$. Let $(x_i(\cdot))_{i=1}^n \in \mathscr{C}^0(\mathbb{R}_{\geqslant 0}; (\mathbb{S}^{d-1})^n)$ be the unique solution to the corresponding Cauchy problem for* (SA) *or* (USA)*. Take any $\lambda = \lambda(\beta)$ such that*

$$0 < \lambda < 1 - \alpha - O_{\beta, n}\left(\frac{1}{\beta}\right) \tag{1.5}$$

(see Remark 1.3 for the precise upper bound) and

$$\lambda(\beta) = \Omega(1), \tag{1.6}$$

where $\gamma = \gamma(\beta) > 0$ and $\alpha = \alpha(\beta) \in (-1, 1)$ are defined in (1.3) and (1.4) respectively. Then there exist $T_2 > T_1 > 0$ with

$$T_1 \leqslant 2ne^{8\varepsilon\beta} + en\lambda\frac{\beta^2}{\beta - 1} \qquad and \qquad T_2 \geqslant \frac{\varepsilon}{n}e^{(1-\alpha)\beta},$$

such that

1. *If $x_i(0) \in \mathcal{S}_q(\varepsilon)$, then $x_i(t) \in \mathcal{S}_q(2\varepsilon)$ for all $t \in [0, T_2]$;*

2. *For all $q \in \{1, \ldots, k\}$,*

$$\max_{x_i(t), x_j(t) \in \mathcal{S}_q(2\varepsilon)} \|x_i(t) - x_j(t)\|^2 \leqslant 2e^{-\lambda\beta} \tag{1.7}$$
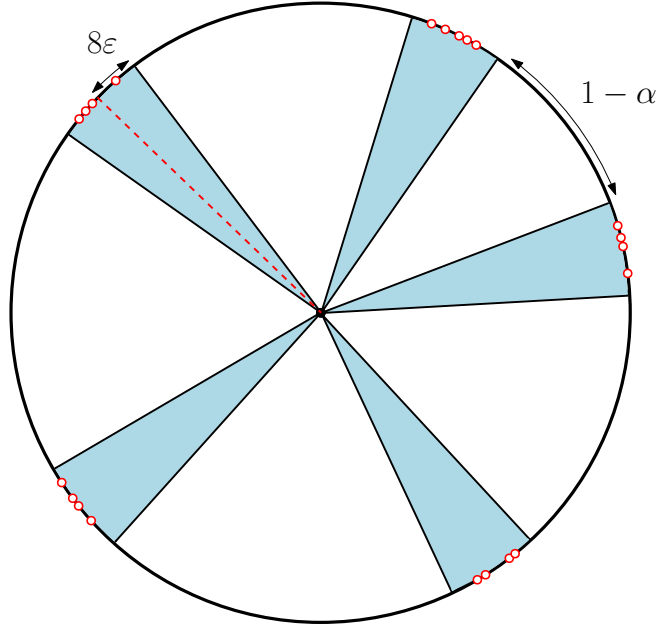
   for all $t \in [T_1, T_2]$.

Since $1 - \alpha > 0$ and $1 - \alpha = \Omega(1)$ by virtue of (1.3), the time T_2 which the particles take to escape from the caps $\mathcal{S}_q(2\varepsilon)$ is exponentially long. Furthermore since $\lambda = \Omega(1)$, after time T_1 all particles within a cap $\mathcal{S}_q(2\varepsilon)$ are exponentially close to each other. This is precisely the dynamic metastability phenomenon alluded to in the introductory remarks: all particles stay in the vicinity of k points for an exponentially long period of time. See Figure 2 for a simulation.

Before delving into a more extensive discussion, we offer some preliminary remarks.

Remark 1.3 (On (1.5)). *The upper bound we require in (1.5) is precisely:*

$$\lambda < \min\left\{e^{\left(1-\alpha-\frac{1}{\beta}\log\frac{(\beta-1)\varepsilon}{\beta^2 n^2 e}\right)\beta}(1 - e^{-\gamma\beta}), 1 - \alpha - \frac{\log\left(\frac{2n^2}{1-e^{-\lambda_*\beta}}\right)}{\beta} - e^{-\lambda_*\beta}\right\} \tag{1.8}$$

Figure 1: An illustration of a (β, ε)-separated configuration on the circle $\mathbb{S}^1$. To clearly visualize distances, we not only show the spherical caps $\mathcal{S}_j(\varepsilon)$, but also their convex hull within the unit disk. The case of interest in our framework is that in which caps have an opening ε that is much smaller than the distance $1 - \alpha$ between them.

where $\lambda_* = \beta^{-1} \log(1/8\varepsilon) > 0$. We use the first upper bound to ensure $T_2 > T_1$, and the second in a "propagation of smallness" argument—see (2.11) in the proof. By straightforward numerical computations, using (1.3) and $\varepsilon \in (0, \frac{1}{16})$ one finds that the second term in the upper bound in (1.8) is greater than λ_* (a useful lower bound that we use in (2.8)). Moreover, since the upper bound in (1.8) is—again by virtue of (1.3)—of order $1 - \alpha$ asymptotically as $\beta \to +\infty$, λ can always be chosen so that (1.6) holds as well.

Remark 1.4 ($\Omega(1)$). *Recall the Bachmann-Landau notation: $f(x) = \omega(g(x))$ whenever $\lim_{x \to +\infty} f(x)/g(x) = +\infty$. We chose to impose $\gamma = \Omega(1)$ in (1.3) and $\lambda = \Omega(1)$ in (1.6) solely to ensure that, in Theorem 1.2, the escape time T_2 is exponentially large, and the smallness rate in (1.7) is exponentially small as functions of β. One can replace $\Omega(1)$ with $\omega(\beta^{-1})$ for both, and provided one doesn't choose an initial configuration that is asymptotically reduced to a point—e.g., $\gamma \sim \log(\beta)/\beta$ or something similar—both the escape time and the smallness rate remain of exponential order.*

Remark 1.5 (Low temperature). *One could equivalently rephrase Definition 1.1 and Theorem 1.2 so that, instead of having "well-separated" initial configurations and an arbitrary β, one rather takes configurations that solely satisfy $1 - \alpha - 8\varepsilon > 0$, and then takes β sufficiently large so that γ defined in (1.3) is positive, and adjusts*

λ appropriately.

Remark 1.6 (Different heights)**.** *In Definition 1.1, all spherical caps are defined using the same height ε. The proof can however be adapted to employ different heights $ε_1, \ldots, ε_k$ per spherical cap without much difficulty. This modified proof yields the same result as discussed in the first step (see 6.1) of the proof of Theorem 6.4 later on. For the sake of simplicity we choose to present the result in less generality.*

Remark 1.7 (Safety caps)**.** *In Theorem 1.2 the time T_2 that up to which all particles remain in the safety caps $S_q(2ε)$. It is possible, at the cost of additional technicalities, to reduce the height parameter $2ε$ to $ε + δ$ with $0 < δ \ll ε$, up to changing to a time T_2^* which is smaller than T_2 and modifying the constant λ. Once again, we choose to present our result in a less general form for the sake of simplicity.*

Remark 1.8 (Time of collapse)**.** *The time T_1 beyond which the particles within caps remain exponentially close to each-other can scale exponentially with β when ε is not of order at least $β^{-1}$. We believe that this estimate is sub-optimal due to coarse bounds in Step 2 of the proof in §2 (see Remark 2.3 as well), and could be improved under further assumptions on the initial distribution of particles inside each spherical cap (for instance, equidistributed within each cap). We leave this open.*
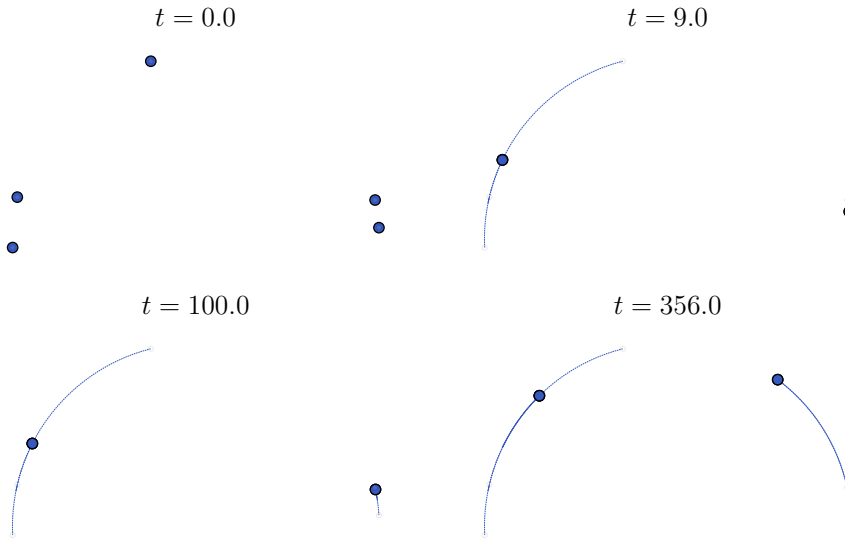
Figure 2: A stylized illustration of Theorem 1.2: here $d = 2$, $n = 5$ and $β = 4$, initial points are distributed uniformly at random, and (SA) is solved using a forward Euler scheme with time step equal to 0.1. Two caps appear and beyond time $T_1 \sim 9$ particles within these caps are essentially merged. The dynamics remains in this metastable state at least up to time $T_2 \sim 356$, a point beyond which the two merged rightmost particles exit the cap, $S_1(2ε)$ say, and Theorem 1.2 is no longer indicative. Continued in Figure 3.

1.2 Discussion and outline

We further discuss the particular framework of Theorem 1.2 as well as extensions thereof, whilst outlining the remainder of the paper.

1.2.1 An energetic reinterpretation (§3)

The proof of Theorem 1.2, which can be found in **§2**, does not make use of the gradient flow interpretation of (SA) nor (USA), as we rather resort to ODE arguments and a fine analysis of the attention nonlinearity. One can however reinterpret Theorem 1.2 almost entirely in terms of the energy E_β by following a general framework introduced by Otto and Reznikoff in [OR07]. To put it briefly: for a gradient flow (descent, say; ascent follows thereupon) of some smooth function E on an abstract manifold $\mathcal{M}$, if there is a subset $\mathcal{N} \subset \mathcal{M}$ on which $\nabla\mathsf{E}$ is of magnitude $\delta \ll 1$, and in whose vicinity E satisfies a *Polyak-Łojasiewicz*-like inequality (see (3.2)), then trajectories are quickly drawn to $\mathcal{N}$ and remain there for time at least δ^{-1}. We present the framework of Otto and Reznikoff in **§3.1**, and prove that our energy E_β fits within this framework in **§3.2**. Finally, in **§3.3**, we point out that outside of the metastable states, the gradient of the energy is *accelerating*.

1.2.2 On (β, ε)-separated configurations (§4)

It is also natural to inquire about the ubiquitousness of (β, ε)-separated configurations per Definition 1.1. In Proposition 4.2 in **§4.1**, we prove that random points drawn from an appropriate Gaussian mixture, projected onto $\mathbb{S}^{d-1}$, satisfy this assumption with high probability. We cover points drawn from the uniform distribution on $\mathbb{S}^{d-1}$ in **§4.2**. When on the circle ($d = 2$) with $n \gg 1$, this condition is rarely true, yet numerical experiments presented in [GLPR23] still indicate the appearance of metastability. We stipulate that the sharp assumption on the initial condition should be related to sufficiently large levels of the energy E_β—see **§4.3**.

1.2.3 The mean field regime (§5)

For the sake of generality we extend Theorem 1.2 to the mean-field limit $n \to +\infty$ in **§5**. Specifically, for an initial measure supported in the union of k spherical caps (akin to Definition 1.1), the corresponding solution to the continuity equation for which (SA) or (USA) are the projected characteristics, also displays metastability.

1.2.4 Beyond the escape time (§6)

Finally, one may wonder if something can be said beyond the exit time T_2 in Theorem 1.2. In Figure 3 we illustrate the continuation of Figure 2, which indicates that all particles eventually coalesce to a single cluster. The issue we encountered in extending our proof of Theorem 1.2 to accommodate further escape times lies in propagating the (β, ε)-separateness assumption.
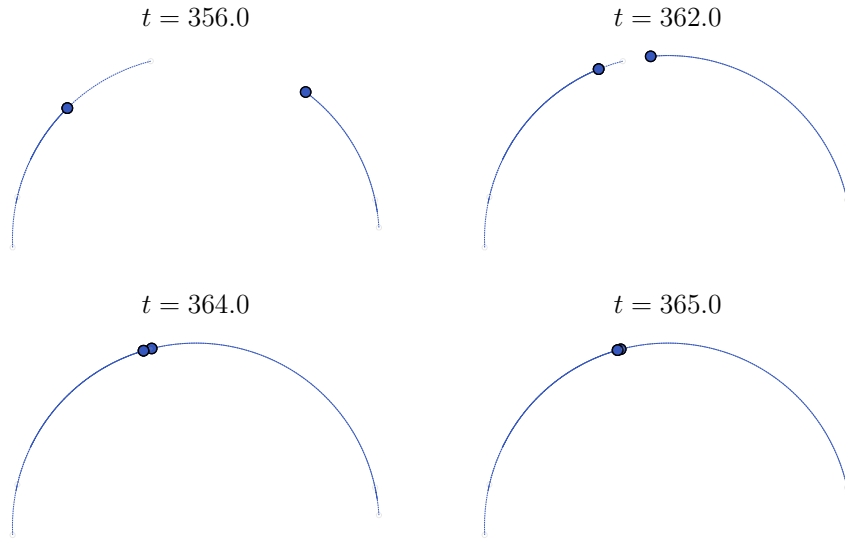
Figure 3: Continuing upon Figure 2, we see that particles keep converging until they meet at a cluster, which is the global maximum of E_β. We recall that in this setup ($d = 2$ and $\beta \gtrsim\!\!\!\!\!\!\diagup\ n^2$, nor are initial particles in some hemisphere), there is no proof of convergence to a cluster as of yet. A movie of the full evolution can be found at https://github.com/HugoKoubbi/2024-transformers-dotm/blob/main/video[tape]/1.gif.

Contrasting Figure 3 to 2 one sees different time-scales: once the two clusters of particles are "sufficiently close", they take little time (compared to the time spent in the spherical caps) to collapse to a single cluster. Leveraging the gradient flow interpretation: the energy stays at a constant level over a long time-scale, before accelerating very quickly over a shorter time-scale, resulting in a jump to another constant level. In the presence of multiple initial clusters, one expects multiple jumps.

To formalize this heuristic, in **§6** we study the low-temperature limit $\beta \to +\infty$ with $d, n \geqslant 2$ fixed ($d = 2$ in our considerations), and seek to find a time-rescaling under which the energy E_β reaches its global maximum in finite time. To this end we need to slightly modify the dynamics: we clump particles that are within a critical window of size roughly β^{-1}, and consider the dynamics associated with a single weighted particle instead of closely spaced particles. Such ideas are commonplace in *renormalization group theory* in statistical physics [Car96, Chapter 3], where systems are simplified by integrating out short-distance degrees of freedom, effectively rescaling the problem to focus on the behavior at larger scales[1]. In this regard, in Theorem 6.4, for particularly well-prepared initial configurations, we construct a time-rescaling along which the energy has a staircase profile as $\beta \to +\infty$, and reaches its global maximum in finite time. See Figure 4 for an illustration.

---

[1] We thank Bruno Loureiro for pointing out this link.
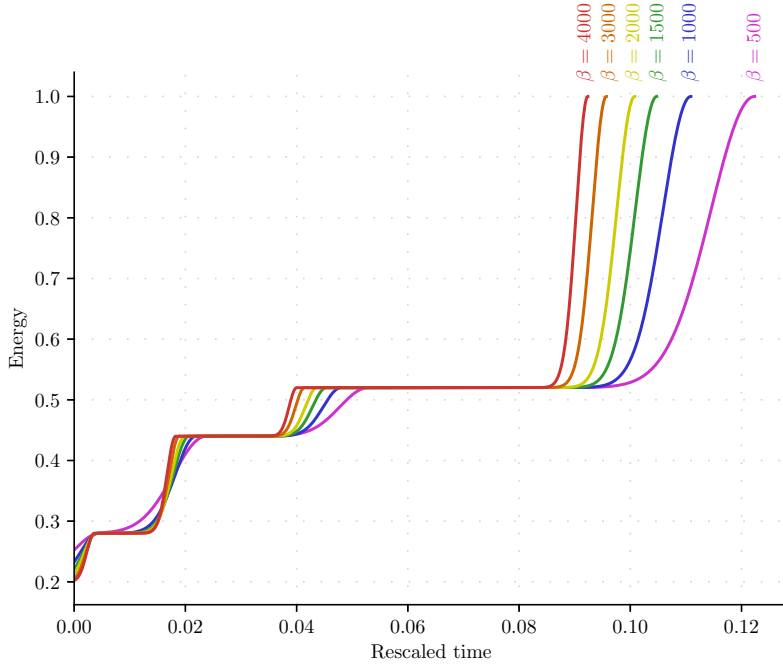
Figure 4: Theorem 6.4 entails that the energy of a trajectory along the time-scale defined in (6.1) converges, uniformly in time, as $\beta \to +\infty$, to a piecewise constant-in time function which equals 1 (designating the maximal value of E_β) beyond some finite time $T_k > 0$. Plateaux indicate metastable zones, and jumps in the energy level indicate rapprochement of nearby clusters.

Upon seeing Figure 4 one can also draw a connection to several works regarding the training dynamics of neural networks, in which a similar staircase profile for the loss function is observed. Therein, this behavior goes under the names *incremental learning* or *saddle-to-saddle dynamics*. The regime of incremental learning has been analyzed in the training dynamics of linear neural networks [GBLJ19, JGŞ+21], diagonal neural networks [Ber23, GSSD19, PF23], more general neural networks [BPVF22, ABBA+24], and also in tensor decomposition [RMC21, JCD23]. An excellent reference for further results and discussions is the thesis [Pes24].

1.3   Related work

Self-attention dynamics   The particle system formulation of Transformers as in (SA) is set out in [SABP22, LLH+19], without using layer normalization so that the particles evolve on $\mathbb{R}^d$. The resulting system is related to many other variants studied in collective behavior—see [Tad23] and the references therein. As a matter of fact, (SA) is itself a generalization of the celebrated Kuramoto model [Kur75]. In [SABP22] a variant of (SA) in which the interaction terms yield a bi-stochastic

10

matrix at every time t instead of solely a stochastic one is additionally introduced; this model is further delved-into in [AHMP24]. The particle system formulation is extended to include *masked* self-attention in [CAP24]—this is relevant for *decoder* Transformer models, in contrast to *encoder* models which largely underpin our motivation. Considering the model on $\mathbb{R}^d$, the authors in [GLPR24] prove clustering in long time, in the presence of various other parameters (other than just β), under an appropriate time-rescaling which renders the equation rather comparable to that on $\mathbb{S}^{d-1}$. These results are first extended in [KBH24], where stability of clustering with respect to perturbations of the initial conditions and of the parameters is shown, and then in [AFZ24], where the zero temperature model is analyzed in discrete time.

The interaction energy The study of minima (or maxima) of interaction energies such as (1.1) is a classical question not only in physics but also in combinatorics, particularly in relation to sphere packing problems. Indeed, for a wide array of monotonic potentials $f : [-1, 1] \to \mathbb{R}_{\geqslant 0}$, encompassing $s \mapsto e^{\beta s}$ which yields (1.1), the minima of $\mathsf{E}(X) = \sum \sum f(\langle x_i, x_j \rangle)$ are optimal configurations on $\mathbb{S}^{d-1}$ [CK07]. Versions thereof on $\mathbb{R}^d$ with radial potentials are also canonical and include the Gaussian core model [Sti76, CdCI18], Coulomb-Riesz potentials [SK97, PS20], and so on.

With regard to the particular example of (1.1), [MTG17, CRMB24] prove that E_β has no local maxima for $\beta \geqslant 0$ and $d \geqslant 3$, improving upon [GLPR23], and also improve $\beta \lesssim \frac{1}{n}$ to $\beta \leqslant 1$ when $d = 2$. See [MB24] for related work in this regard. In the context of the related Kuramoto model, we also refer the reader to [ABK$^+$22, LXB19] for further results on benign landscapes, where the energy contains additional multiplicative coefficients stemming from the adjacency matrix of various random graphs and/or expanders. These are obtained as semidefinite relaxations of diverse combinatorial optimization problems.

Slow motion of gradient flows The starting point of our study is [OR07], which, as alluded to in what precedes, presents an abstract framework for studying slow motion of gradient flows. The application in question is the Allen-Cahn equation in one space dimension

$$\partial_t u - \varepsilon^2 \partial_{xx} u = u(1 - u^2),$$

which is the L^2-gradient flow of the scalar Ginzburg-Landau energy

$$u \mapsto \frac{\varepsilon^2}{2} \int (\partial_x u)^2 \, \mathrm{d}x + \frac{1}{4} \int (u^2 - 1)^2 \, \mathrm{d}x.$$

The dynamics thereof has been a major area of research over the past forty years [CP89, FH89, Peg07], and we only discuss it briefly for completeness. The limit $u_\infty = \lim_{t \to +\infty} u(t)$ exists and satisfies

$$-\varepsilon^2 \partial_{xx} u + u(u^2 - 1) = 0. \tag{1.9}$$

11

There are exactly two stable equilibrium states: $u \equiv 1$ and $u \equiv -1$. When $\varepsilon > 0$, u approaches 1 where $u > 0$ initially and -1 where $u < 0$ initially. Walls form between these domains, at positions corresponding roughly to zeros in the initial data. A domain wall has characteristic width of order ε and can be described explicitly as the solution of an ODE remeniscent to (1.9). The domain structure one then expects to develop consists of arbitrarily placed domain walls of characteristic width ε, separating domains in which u is exponentially close to the stable states ± 1. This is known as *coarsening*. As the domain walls move extremely slowly, this behavior is referred to as dynamic metastability.

Stochastic dynamics Metastability is also extensively studied in the literature on the physics of disordered systems. Contrary to our setting, in disordered systems the energy landscape E may have plenty of local minima. The question of interest is to qualitatively describe the Langevin dynamics

$$\mathrm{d}X_t = -\nabla\mathsf{E}(X_t)\,\mathrm{d}t + \sqrt{2}\,\mathrm{d}B_t$$

where $(B_t)_{t\geqslant 0}$ denotes the standard Brownian motion. The mathematically rigorous analysis of metastability for such dynamics dates back to the work of Freidlin and Wentzell in the early 1970s [FW98] (see [BDH16] for more recent developments), and is based on large deviation theory on path-space. On short time scales, trajectories of the system follow those of the system without stochasticity and thus converge toward one of the attractors. On much longer time scales, the stochastic perturbation allows and facilitates the system to perform transitions between stable attractors. Although metastability is inherently dynamic, there are methods based on a study of the energy landscape and the critical points relying on replica theory—see [RF22] for a recent treatise.

2 A direct proof

We first provide an ODE-based proof of Theorem 1.2 which uses solely the specific structure of the equation and does not rely on any abstract arguments.

Proof of Theorem 1.2. We focus on (SA), but the arguments are identical for (USA). Set

$$a_{ij}(t) = \frac{e^{\beta\langle x_i(t), x_j(t)\rangle}}{\displaystyle\sum_{k=1}^{n} e^{\beta\langle x_i(t), x_k(t)\rangle}}.$$

In the following, we drop the dependence on β for α and ε for the sake of readability. For ease of reading, we also split the proof in several steps.

Step 1. Lower-bounding the escape time

Naturally,

$$T_2 = T_{\mathrm{esc}} := \inf\left\{ t \geqslant 0 \colon \exists i \in \{1, \ldots, n\} \text{ such that } x_i(t) \notin \bigcup_{q=1}^{k} \mathcal{S}_q(2\varepsilon) \right\}.$$

For $q \in \{1, \ldots, k\}$ we also define

$$T_{\mathrm{esc}}(q) := \inf\{ t \geqslant 0 \colon x_i(0) \in \mathcal{S}_q(\varepsilon) \text{ but } x_i(t) \notin \mathcal{S}_q(2\varepsilon) \}.$$

Observe that

$$T_{\mathrm{esc}} = \min_{q \in \{1, \ldots, k\}} T_{\mathrm{esc}}(q),$$

whence we can localize our analysis to a single cap to begin with. Take an arbitrary $q \in \{1, \ldots, k\}$ and consider

$$\eta_q(t) := \min_{x_i(t) \in \mathcal{S}_q(2\varepsilon)} \langle x_i(t), w_q \rangle.$$

Setting

$$i(t) \in \operatorname*{arg\,min}_{i \colon x_i(t) \in \mathcal{S}_q(2\varepsilon)} \langle x_i(t), w_q \rangle,$$

for $t \in [0, T_{\mathrm{esc}}]$ we compute[2]

$$\dot{\eta}_q(t) = \sum_{j \colon x_j(t) \in \mathcal{S}_q(2\varepsilon)} a_{i(t)j}(t) \left\langle \mathbf{P}^{\perp}_{x_{i(t)}(t)}(x_j(t)), w_q \right\rangle$$
$$+ \sum_{j \colon x_j(t) \notin \mathcal{S}_q(2\varepsilon)} a_{i(t)j}(t) \left\langle \mathbf{P}^{\perp}_{x_{i(t)}(t)}(x_j(t)), w_q \right\rangle.$$

On one hand, we have

$$\left| \sum_{j \colon x_j(t) \notin \mathcal{S}_{2q}(\varepsilon)} a_{i(t)j} \left\langle \mathbf{P}^{\perp}_{x_{i(t)}(t)}(x_j(t)), w_q \right\rangle \right| \leqslant n e^{-(1-\alpha)\beta}, \tag{2.1}$$

---

[2]To compute this derivative, we first fix t_0, compute the derivative of $\langle x_{i_0(t)}(t), w_q \rangle$, and evaluate at $t = t_0$.

and by plugging (2.1) in the previous identity, we find

$$\dot{\eta}_q(t) \geqslant \sum_{j:\, x_j(t) \in \mathcal{S}_q(2\varepsilon)} a_{i(t)j}(t) \left\langle \mathbf{P}^{\perp}_{x_{i(t)}(t)}(x_j(t)), w_q \right\rangle - n e^{-(1-\alpha)\beta}. \qquad (2.2)$$

By definition of $i(t)$, for all indices j such that $x_j(t) \in \mathcal{S}_q(2\varepsilon)$ we have

$$\langle x_j(t), w_q \rangle \geqslant \langle x_{i(t)}(t), w_q \rangle, \qquad (2.3)$$

so by expanding $\langle \mathbf{P}^{\perp}_{x_{i(t)}(t)}(x_j(t)), w_q \rangle$ we see that the sum in (2.2) is nonnegative. Going back to (2.2) we end up with

$$\dot{\eta}_q(t) \geqslant -n e^{-(1-\alpha)\beta}.$$

Thence

$$\eta_q(t) - \eta_q(0) \geqslant -nt e^{-(1-\alpha)\beta},$$

and so

$$\eta_q(t) \geqslant 1 - \varepsilon - nt e^{-(1-\alpha)\beta}.$$

Consequently, as long as $t \leqslant \frac{\varepsilon}{n} e^{(1-\alpha)\beta}$ we have $\eta_q(t) \geqslant 1 - 2\varepsilon$, and so

$$T_{\mathrm{esc}}(q) \geqslant \frac{\varepsilon}{n} e^{(1-\alpha)\beta}.$$

Step 2. Monotonicity within caps

We now show that beyond time $T_1 \in (0, T_{\mathrm{esc}})$, all particles within a cap will remain exponentially close. To this end, for $q \in \{1, \dots, k\}$ and $t \in [0, T_{\mathrm{esc}}]$ we consider

$$\rho_q(t) := \min_{x_i(t), x_j(t) \in \mathcal{S}_q(2\varepsilon)} \langle x_i(t), x_j(t) \rangle.$$

(Note that

$$\frac{1}{2} \max_{x_i(t), x_j(t) \in \mathcal{S}_q(2\varepsilon)} \|x_i(t) - x_j(t)\|^2 = 1 - \rho_q(t)$$

for reference.) We also consider $i(t), j(t)$ (both depending on q) such that

$$(i(t), j(t)) \in \underset{\substack{(i,j) \in \{1,\dots,n\}^2 \\ x_i(t) \neq x_j(t) \in \mathcal{S}_q(\varepsilon)}}{\arg\min} \langle x_i(t), x_j(t) \rangle.$$

We compute as before

$$\dot{\rho}_q(t) = \sum_{k=1}^{n} a_{i(t)k} \left\langle \mathbf{P}^{\perp}_{x_{i(t)}(t)}(x_k(t)), x_{j(t)}(t) \right\rangle + \sum_{k=1}^{n} a_{j(t),k} \left\langle \mathbf{P}^{\perp}_{x_{j(t)}(t)}(x_k(t)), x_{i(t)}(t) \right\rangle.$$

Bounding any of the two sums in the above identity is clearly the same, so we focus on a single one, the first one say. As in the first step, we split the sum into particles lying in the cap $\mathcal{S}_q(2\varepsilon)$ and those in the complement; we first see that

$$\sum_{k:\, x_k(t)\in\mathcal{S}_q(2\varepsilon)} a_{i(t)k}\left(\left\langle x_k(t), x_{j(t)}(t)\right\rangle - \left\langle x_k(t), x_{i(t)}(t)\right\rangle\left\langle x_{i(t)}(t), x_{j(t)}(t)\right\rangle\right)$$

$$\geqslant \sum_{k:\, x_k(t)\in\mathcal{S}_q(2\varepsilon)} a_{i(t)k}\left\langle x_{i(t)}(t), x_{j(t)}(t)\right\rangle\left(1-\left\langle x_k, x_{i(t)}(t)\right\rangle\right) \qquad (2.4)$$

where we used $\langle x_k(t), x_{j(t)}(t)\rangle \geqslant \langle x_{i(t)}(t), x_{j(t)}(t)\rangle$. Since $a_{ij}(t) \geqslant \frac{1}{n}e^{\beta(\rho_q(t)-1)}$, we also find

$$\sum_{k:\, x_k(t)\in\mathcal{S}_q(2\varepsilon)} a_{i(t)k}\left\langle x_{i(t)}(t), x_{j(t)}(t)\right\rangle\left(1-\left\langle x_k, x_{i(t)}(t)\right\rangle\right)$$

$$\geqslant \frac{1}{n}\rho_q(t)(1-\rho_q(t))e^{\beta(\rho_q(t)-1)}, \qquad (2.5)$$

where we only keep the $j(t)$-th term in the sum above. On the other hand, we also have

$$\left|\sum_{k:\, x_k(t)\notin\mathcal{S}_q(2\varepsilon)} a_{i(t)k}\left\langle \mathbf{P}^{\perp}_{x_{i(t)}(t)}(x_k(t)), x_{i(t)}(t)\right\rangle\right| \leqslant ne^{-(1-\alpha)\beta}. \qquad (2.6)$$

All in all, combining (2.4), (2.5), and (2.6), we deduce

$$\dot{\rho}_q(t) \geqslant \frac{2}{n}\rho_q(t)(1-\rho_q(t))e^{\beta(\rho_q(t)-1)} - 2ne^{-(1-\alpha)\beta}. \qquad (2.7)$$

Step 3. The collapse time

Fix $\lambda > 0$ as in the statement, and assume furthermore that

$$\lambda > \frac{1}{\beta}\log\left(\frac{1}{8\varepsilon}\right). \qquad (2.8)$$

Assuming (2.8) is without loss of generality, since if (1.7) holds for such λ, it also holds for all smaller, positive λ. (See also Remark 1.3.) We wish to use (2.7) to find a time beyond which $1 - \rho_q(t)$ is exponentially small. To this end, consider

$$T_*(q) := \inf\left\{t \in [0, T_{\mathrm{esc}}]\colon \rho_q(t)(1-\rho_q(t))e^{\beta(1-\rho_q(t))} \leqslant 2n^2 e^{-(1-\alpha)\beta}\right\},$$

with $\inf\varnothing = +\infty$. We show that $\max\limits_{q\in\{1,\ldots,k\}} T_*(q) \leqslant T_{\mathrm{esc}}$. Suppose $T_*(q) > T_{\mathrm{esc}}$. Then

$$\dot{\rho}_q(t) \geqslant \frac{1}{n}\rho_q(t)(1-\rho_q(t))e^{\beta(\rho_q(t)-1)} \qquad (2.9)$$

for all $t \in [0, T_{\mathrm{esc}}]$. In particular, $t \mapsto \rho_q(t)$ is increasing, and recall that it is bounded from above by 1. The following lemma is of crucial use.

15

Lemma 2.1 (Until collapse). *Suppose $\beta > 1$ and $c > 0$. For $u_0 \in (0, 1]$ consider $u \in \mathscr{C}^0(\mathbb{R}_{\geqslant 0}; [0, 1])$ the unique solution to the Cauchy problem*

$$\begin{cases} \dot{u}(t) = u(t)(1 - u(t))e^{\beta(u(t)-1)} & \text{for } t \geqslant 0 \\ u(0) = u_0. \end{cases}$$

Then,

$$\inf\left\{t \geqslant 0 \colon 1 - u(t) \leqslant e^{-c\beta}\right\} \leqslant \frac{e^{\beta(1-u_0)}}{u_0} + \frac{\beta^2 \cdot c \cdot e}{\beta - 1}.$$

We postpone the elementary proof to Appendix A.1.1. We combine (2.9) and the comparison principle for scalar ODEs along with Lemma 2.1 with $u_0 = \rho_q(0)$: we have $\rho_q(tn) \geqslant u(t)$, thence

$$\inf\left\{t \geqslant 0 \colon 1 - \rho_q(tn) \leqslant e^{-\lambda\beta}\right\} \wedge T_{\text{esc}} \leqslant \inf\left\{t \geqslant 0 \colon 1 - u(t) \leqslant e^{-\lambda\beta}\right\} \wedge T_{\text{esc}}.$$

So

$$T_1(q) := \inf\left\{t \geqslant 0 \colon 1 - \rho_q(t) \leqslant e^{-\lambda\beta}\right\} \wedge T_{\text{esc}} \leqslant \frac{ne^{\beta(1-\rho_q(0))}}{\rho_q(0)} + \frac{n \cdot \beta^2 \cdot \lambda \cdot e}{\beta - 1}$$

$$\leqslant 2ne^{8\varepsilon\beta} + \frac{n \cdot \beta^2 \cdot \lambda \cdot e}{\beta - 1}$$

where we used $\rho_q(0) \geqslant 1 - 8\varepsilon > \frac{1}{2}$. The upper bound above is independent of q and is strictly smaller than $\frac{\varepsilon}{n}e^{(1-\alpha)\beta}$ because of the first of the upper bounds in (1.8), which is a contradiction with the lower bound on T_{esc} deduced in Step 1. Therefore

$$T_1 := \max_{q \in \{1,\ldots,k\}} T_1(q) \wedge T_*(q) < T_{\text{esc}}.$$

Step 4. Within caps, particle stick

The previous step entails

$$1 - \rho_q(T_1) \leqslant e^{-\lambda\beta}. \tag{2.10}$$

We seek to propagate this smallness for all times up to T_2. This follows from the following lemma.

Lemma 2.2 (Propagation). *Fix $\beta > 1$, and consider $\delta \in (0, 1)$ and $\alpha \in (-1, 1)$ such that*

$$\frac{1}{n}\delta(1 - \delta)e^{-\delta\beta} > ne^{-(1-\alpha)\beta}. \tag{2.11}$$

Suppose $(x_i(0))_{i=1}^n \in (\mathbb{S}^{d-1})^n$ is such that

$$\langle x_i(0), x_j(0) \rangle \geqslant 1 - \delta$$

for some $I \subset \{1, \ldots, n\}$ and for all $(i, j) \in I^2$. Let $(x_i(\cdot))_{i=1}^n \in \mathscr{C}^0(\mathbb{R}_{\geqslant 0}; (\mathbb{S}^{d-1})^n)$ be the unique solution to the corresponding Cauchy problem for (SA) or (USA), and suppose that for all $i \in I$ and $k \in I^c$,

$$\langle x_i(t), x_k(t) \rangle \leqslant \alpha \qquad \text{for all } t \in [0, T],$$

Then for all $(i, j) \in I^2$,

$$\langle x_i(t), x_j(t) \rangle \geqslant 1 - \delta \qquad \text{for all } t \in [0, T].$$

We postpone the proof to Appendix A.1.2. We apply Lemma 2.2 to $\rho_q(t)$ with $\delta = e^{-\lambda\beta}$, starting from time T_1 instead of 0—all conditions in the statement being satisfied by virtue of the second of the upper bounds in (1.8) and (2.8) (for (2.11)), (2.10) and the definition of T_2 respectively—to deduce that

$$1 - \rho_q(t) \leqslant e^{-\lambda\beta} \qquad \text{for all } t \in [T_1, T_2].$$

This concludes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square$

Remark 2.3. *We can provide an even more refined picture of the dynamics: within each cap—the quantity $t \mapsto \eta_q(t)$ is actually increasing up to a certain time. Indeed in the first step of the proof, for all $t \geqslant 0$ we saw that*

$$\dot{\eta}_q(t) \geqslant \eta_q(t) \sum_{j \,:\, x_j(t) \in \mathcal{S}_q(2\varepsilon)} a_{i(t)j}(t) \frac{\|x_j(t) - x_{i(t)}(t)\|^2}{2} - n e^{-(1-\alpha)\beta}.$$

This shows that the variance within a cap $\mathcal{S}_q(2\varepsilon)$, defined as

$$\text{Var}_q(t) := \sum_{j \,:\, x_j(t) \in \mathcal{S}_q(2\varepsilon)} a_{i(t)j}(t) \frac{\|x_j(t) - x_{i(t)}(t)\|^2}{2},$$

controls the rate of convergence within the cap. It is however not straightforward to show the monotonicity of this variance. Indeed, consider a spherical cap which contains two sub-caps, which are separated. Then the variance will first increase, and then decrease, exponentially fast.

3   An energetic reinterpretation

We now provide a rewriting of our metastability result by leveraging the gradient flow structure, following the framework proposed by Otto and Reznikoff in [OR07].

3.1 The Otto-Reznikoff framework

We begin by reviewing the framework and result proposed in [OR07]. Consider an abstract gradient flow[3] evolving on a manifold $\mathcal{M} \subset \mathbb{R}^d$

$$\begin{cases} \dot{u}(t) = -\nabla \mathsf{E}(u(t)) & \text{for } t \geqslant 0 \\ u(0) = u_0 \end{cases} \tag{3.1}$$

for a given $u_0 \in \mathcal{M}$. Infinite-dimensional versions can also be considered—we keep the presentation formal, as done in [OR07]. Here $\mathsf{E} : \mathcal{M} \to \mathbb{R}_{\geqslant 0}$ is assumed smooth, but more importantly, we assume that there exists $\mathcal{N} \subset \mathcal{M}$ such that

(H1) For every $u \in \mathcal{M}$ there exists $v \in \mathcal{N}$ such that

$$\frac{1}{2}\|u - v\|^2 \leqslant \mathsf{E}(u) - \mathsf{E}(v) \leqslant \frac{1}{2}\|\nabla \mathsf{E}(u)\|^2; \tag{3.2}$$

(H2) There exists some constant $\delta > 0$ such that for all $v_1, v_2 \in \mathcal{N}$,

$$|\mathsf{E}(v_1) - \mathsf{E}(v_2)| \leqslant \delta \|v_1 - v_2\|.$$

The upper bound of the energy discrepancy in **(H1)** is reminiscent of a *Polyak-Łojasiewicz (PL)* inequality ([BDL07, KNS16]) in the vicinity of the slow manifold $\mathcal{N}$. This makes $\mathcal{N}$ attractive for points $u \in \mathcal{M} \setminus \mathcal{N}$. On the other hand, should $\delta \ll 1$, **(H2)** entails that, along $\mathcal{N}$, the landscape is essentially flat since the energy gradient is of order δ. Hence, the flow ought to remain trapped in $\mathcal{N}$. The manifold $\mathcal{N}$ is determined by the above hypotheses and, because of **(H2)**, is referred to as the *slow manifold*.

The following result is then shown in [OR07]—we repeat the statement verbatim.

Theorem 3.1 ([OR07, Theorem 1.1]). *Suppose that **(H1)**–**(H2)** hold, and let v be such that $v(t)$ and $u(t)$ satisfy (3.2). Then the solution of (3.1) is drawn into a δ-neighborhood of $\mathcal{N}$ with an exponential rate close to 1; that is, for any $\varepsilon \in (0, 1)$, there exists a finite constant $C_\varepsilon > 0$ such that*

$$\|u(t) - v(t)\| + \sqrt{\mathsf{E}(u(t)) - \mathsf{E}(v(t))} \leqslant e^{-(1-\varepsilon)t}\sqrt{\mathsf{E}(u(0)) - \mathsf{E}(v(0))} + C_\varepsilon \delta. \tag{3.3}$$

Moreover, we have for any $0 < s < t$ that

$$\|u(t) - u(s)\| \leqslant \sqrt{\mathsf{E}(u(s)) - \mathsf{E}(v(s))} + \delta(t - s + 1). \tag{3.4}$$

---

[3] To stay faithful to [OR07] we review the framework in the case of gradient descent, but all results apply for gradient ascent under appropriate sign changes.

As the statement of Theorem 3.1 appears[4] different from Theorem 1.2, we reformulate the concrete conclusion as in [OR07, Remark 1] before proceeding. Suppose $\delta \ll 1$ and set $\varepsilon = \frac{1}{2}$ in Theorem 3.1. For short times, if the initial energy gap is of order 1, then the first term in the upper bound is dominant in (3.3). After time

$$t_1 \sim \log \left(\frac{\mathsf{E}(u(0)) - \mathsf{E}(v(0))}{\delta^2} \right),$$

one sees that the energy gap in (3.3) is reduced to order δ:

$$\sqrt{\mathsf{E}(u(t_1)) - \mathsf{E}(v(t_1))} \lesssim \delta. \tag{3.5}$$

Furthermore, after this initial layer t_1 comes the "slow motion phase", which lasts for a time of order δ^{-1}; setting $s = t_1$ in (3.4),

$$\|u(t) - u(t_1)\| \lesssim \sqrt{\mathsf{E}(u(t_1)) - \mathsf{E}(v(t_1))} + \delta(t - t_1 + 1) \lesssim \delta + \delta(t - t_1)$$

from (3.5). This is precisely like Theorem 1.2 with $\delta \sim e^{-\lambda\beta}$, which we confirm in Corollary 3.6.

3.2 Application to the self-attention model
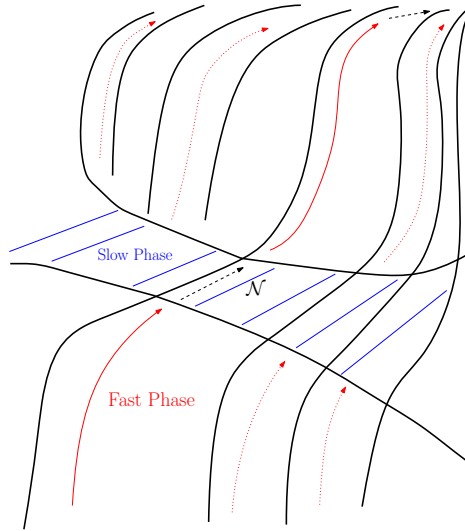
Figure 5: An illustration of the landscape of E_β. The slow manifold $\mathcal{N}$ is an almost-flat zone, thus one where the gradient flow moves very little, and is surrounded by zones where E_β satisfies a PL inequality.

We wish to apply Theorem 3.1 to (USA) and (SA). This requires checking the concave analogue of (**H1**), as well as (**H2**), for the interaction energy E_β defined

---

[4]One can, however, refer to [OR07, Theorem 1.2], concerning the application to the Allen-Cahn equation, where the statement is almost identical to that of our main result.

in (1.1). This in turn requires determining the slow manifold $\mathcal{N}$. A first guess could be be to consider $\mathcal{N}$ as the set consisting of configurations with isolated points, or clustered in isolated points, which we can define as

$$\mathcal{N}_\beta := \left\{ (x_1, \ldots, x_n) \in (\mathbb{S}^{d-1})^n : \max_{1 \leqslant i,j \leqslant n} \|x_i - x_j\| e^{-\frac{\beta}{2}\|x_i - x_j\|^2} \leqslant 2\delta_\beta \right\};$$

here $0 < \delta_\beta \ll 1$ is fixed and to be determined later on. There is two cases in which the left-hand side term in the above definition is be small: either $\|x_i - x_j\|$ is small, or $\|x_i - x_j\|$ is big, since then the exponential part renders the entire term small. The important observation is that for $(x_1, \ldots, x_n) \in \mathcal{N}_\beta$, **(H2)** holds:

$$\|\nabla \mathsf{E}_\beta(x_1, \ldots, x_n)\| \leqslant \delta_\beta.$$

We are therefore left with checking the concave analogue of **(H1)**. PL inequalities are often proven globally by using concavity of the function in question, namely by controlling the spectrum of the Hessian. But it is known that on any compact connected Riemannian manifold, all geodesically concave functions are constant. Alternative to this stationary proof is a *dynamic version* which relies on evaluating the Hessian along trajectories, in the spirit of Bakry-Émery calculus ([BE85], and also [Vil21, Chapter 9], [OV00]). This is the main clue in our proof—we can adapt this strategy, allowing us to localize near $\mathcal{N}$.

Lemma 3.2. *Let* $\mathsf{E} : \mathcal{M} \to \mathbb{R}_{\geqslant 0}$ *be smooth, and let* $\mathcal{N} \subset \mathcal{M}$. *Fix* $u \in \mathcal{M}$ *and consider*

$$\begin{cases} \dot{X}(t) = \nabla \mathsf{E}(X(t)) & \text{for } t \geqslant 0 \\ X(0) = u. \end{cases}$$

Suppose that there exist $v \in \mathcal{N}$ *and* $T > 0$ *such that*

$$X(T) = v,$$

and a numerical constant $c > 0$ *such that*

$$\langle \nabla \mathsf{E}(X(t)), \operatorname{Hess} \mathsf{E}(X(t)) \nabla \mathsf{E}(X(t)) \rangle \leqslant -c\|\nabla \mathsf{E}(X(t))\|^2 \qquad (3.6)$$

for all $t \in [0, T]$. *Then,*

$$\mathsf{E}(v) - \mathsf{E}(u) \leqslant \frac{1}{2c}\|\nabla \mathsf{E}(u)\|^2.$$

The proof is elementary.

Proof of Lemma 3.2. We compute

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\nabla \mathsf{E}(X(t))\|^2 = 2\langle \nabla \mathsf{E}(X(t)), \operatorname{Hess} \mathsf{E}(X(t)) \nabla \mathsf{E}(X(t)) \rangle$$
$$\leqslant -2c\|\nabla \mathsf{E}(X(t))\|^2.$$

by using (3.6). By virtue of the Grönwall lemma we find

$$\|\nabla \mathsf{E}(X(t))\|^2 \leqslant e^{-2ct}\|\nabla \mathsf{E}(X(0))\|^2.$$

We integrate to find

$$\begin{aligned}
\mathsf{E}(X(t)) - \mathsf{E}(X(0)) &= \int_0^t \|\nabla \mathsf{E}(X(s))\|^2 \, \mathrm{d}s \\
&\leqslant \int_0^t e^{-2cs} \, \mathrm{d}s \|\nabla \mathsf{E}(X(0))\|^2 \\
&\leqslant \frac{1}{2c}\|\nabla \mathsf{E}(X(0))\|^2.
\end{aligned}$$

Since $X(T) = v$,

$$\mathsf{E}(v) - \mathsf{E}(u) \leqslant \frac{1}{2c}\|\nabla \mathsf{E}(u)\|^2. \qquad \square$$

Thus, provided a curated definition of the slow manifold $\mathcal{N}_\beta$, we are reduced to showing (3.6). To this end, it is necessary to have a tractable form of the Hessian of E_β. We focus on the case of the circle $\mathbb{S}^1$ to carry out the computations, but we believe the general idea should extend to the higher-dimensional case. Additionally, we primarily focus on (USA)—the extension to (SA) is discussed in Remark 3.7. We can reparametrize the problem to work with angles on the torus $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$: for $\theta_i(t) = \arccos\langle x_i(t), e_1\rangle$, (USA) equivalently rewrites as

$$\dot{\Theta}(t) = \nabla \mathsf{E}_\beta(\Theta(t)) \tag{3.7}$$

with

$$\mathsf{E}_\beta(\theta_1, \ldots, \theta_n) = \frac{1}{2\beta e^\beta n^2} \sum_{i=1}^n \sum_{j=1}^n e^{\beta \cos(\theta_i - \theta_j)}.$$

In other words,

$$\dot{\theta}_i(t) = \sum_{j=1}^n e^{\beta(\cos(\theta_j(t) - \theta_i(t)) - 1)} \sin(\theta_j(t) - \theta_i(t)) \qquad \text{for } t \geqslant 0.$$

We now reformulate Definition 1.1 in this setting.

Definition 3.3. *Suppose $\beta > 1$ and $\tau \in (0, \frac{1}{16})$. We call $(\theta_1, \ldots, \theta_n) \in \mathbb{T}^n$ a (β, τ)-separated configuration if there exist $k \leqslant n$ points $\omega_1, \ldots, \omega_k \in \mathbb{T}$ such that*

1. For all $i \in \{1, \ldots, n\}$,

$$\theta_i \in \bigcup_{q \in \{1, \ldots, k\}} \mathcal{S}_q(\tau)$$

where

$$\mathcal{S}_q(\tau) := \{\theta \in \mathbb{T} : \cos(\theta - \omega_q) \geqslant 1 - \tau\}. \tag{3.8}$$

2. Furthermore,

$$\gamma := 1 - \alpha - 8\tau - \frac{1}{\beta}\log\left(\frac{2n^2}{\tau}\right) > 0 \quad \text{and} \quad \gamma(\beta) = \Omega(1), \quad (3.9)$$

where

$$\alpha(\tau) := \max_{\substack{(\theta,\phi)\in\mathcal{S}_q(2\tau)\times\mathcal{S}_p(2\tau) \\ q\neq p\in\{1,\dots,k\}}} \cos(\theta - \phi).$$

Lemma 3.4 (PL inequality). *Suppose $\beta > 1$ and $n \geqslant 2$. Consider a configuration $\Theta := (\theta_1(0), \dots, \theta_n(0)) \in \mathbb{T}^n$ which is (β, τ)-separated for some $\tau = \tau(\beta) > 0$ which is such that for all $q \in \{1, \dots, k\}$, and for all $(u, v) \in \mathcal{S}_q(2\tau)$, we have*

$$|u - v| \leqslant \frac{1}{8}\sqrt{\frac{1-\delta}{\beta + \frac{1}{2}}}, \quad (3.10)$$

for some $8(1 + \beta)e^{-(1-\alpha)\beta}e^{-\frac{1}{2}} < \delta < 1$. Take any $\lambda > 0$ as in (1.5)–(1.6), and consider

$$\mathcal{N}_\beta := \left\{(\theta_1, \dots, \theta_n) \in \mathbb{T}^n : \max_{q\in\{1,\dots,k\}} \max_{\theta_i,\theta_j\in\mathcal{S}_q(2\tau)} |\theta_i - \theta_j| \leqslant e^{-\frac{\lambda}{2}\beta}\right\},$$

Then there exist $U \in \mathcal{N}_\beta$ and $\kappa(\beta, n) > 0$ such that

$$\mathsf{E}_\beta(U) - \mathsf{E}_\beta(\Theta) \leqslant \frac{1}{2\kappa(\beta, n)}\|\nabla\mathsf{E}_\beta(\Theta)\|^2.$$

Proof of Lemma 3.4. Consider

$$\begin{cases} \dot{\Theta}(t) = \nabla\mathsf{E}_\beta(\Theta(t)) & \text{for } t \geqslant 0, \\ \Theta(0) = \Theta. \end{cases}$$

From (1.7) in Theorem 1.2, we gather that there exists a time $T > 0$ such that

$$\Theta(T) \in \mathcal{N}_\beta, \quad \text{and} \quad \Theta(t) \in \mathbb{T}^n \setminus \mathcal{N}_\beta \quad \text{for all } t \in [0, T].$$

We now seek to check (3.6). For any $i, j \in \{1, \dots, n\}$ one has

$$\partial_{\theta_i}\mathsf{E}_\beta(\theta_1, \dots, \theta_n) = -\frac{1}{n^2}\sum_{m=1}^n \sin(\theta_i - \theta_m)e^{\beta(\cos(\theta_i-\theta_m)-1)},$$

and

$$\partial_{\theta_i}\partial_{\theta_j}\mathsf{E}_\beta(\theta_1, \dots, \theta_n) = \frac{1}{n^2} \cdot \begin{cases} g(\theta_i - \theta_j) & i \neq j \\ -\displaystyle\sum_{m\in\{1,\dots,n\}\setminus\{i\}} g(\theta_i - \theta_m) & i = j, \end{cases}$$

where we set $g(x) := (\cos(x) - \beta \sin^2(x))e^{\beta(\cos(x)-1)}$. One can observe that the Hessian has the structure of a Laplacian matrix. Let $v \in \mathbb{R}^n$; by standard computations for such matrices, we find

$$\langle \text{Hess}\, \mathsf{E}_\beta(\Theta)v, v \rangle = \sum_{i=1}^n \sum_{j \in \{1,\dots,n\} \setminus \{i\}} \partial_{\theta_i} \partial_{\theta_j} \mathsf{E}_\beta(\Theta) v_i v_j$$

$$= -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \partial_{\theta_i} \partial_{\theta_j} \mathsf{E}_\beta(\Theta)(v_i - v_j)^2.$$

For the sake of concise notation, we henceforth denote

$$\mathsf{H}(t) := \langle \text{Hess}\, \mathsf{E}_\beta(\Theta(t)) \nabla \mathsf{E}_\beta(\Theta(t)), \nabla \mathsf{E}_\beta(\Theta(t)) \rangle .$$

We apply the above computations with $v = \nabla \mathsf{E}_\beta(\Theta(t))$ to find

$$\mathsf{H}(t) = -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \partial_{\theta_i} \partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \left(\partial_{\theta_i} \mathsf{E}_\beta(\Theta(t)) - \partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \right)^2 .$$

Recall that we seek to upper bound $\mathsf{H}(t)$ by $-\|\nabla \mathsf{E}_\beta(\Theta)\|^2$. Let $q \in \{1, \dots, k\}$ and $i \in \{1, \dots, n\}$. By arguing as in (2.1), if $\theta_i(t) \in \mathcal{S}_q(2\tau)$ and $\theta_j(t) \notin \mathcal{S}_q(2\tau)$, we have

$$\left| \partial_{\theta_i} \partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \right| \leqslant \frac{(1+\beta)}{n^2} e^{-(1-\alpha)\beta}.$$

Thus

$$\sum_{j=1}^n \partial_{\theta_i} \partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \left(\partial_{\theta_i} \mathsf{E}_\beta(\Theta(t)) - \partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \right)^2$$

$$\geqslant \sum_{j\, : \, \theta_j(t) \in \mathcal{S}_q(2\tau)} \partial_{\theta_i} \partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \left(\partial_{\theta_i} \mathsf{E}_\beta(\Theta(t)) - \partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \right)^2$$

$$- \frac{4(1+\beta)}{n} e^{-(1-\alpha)\beta} \max_{j \in \{1,\dots,n\}} \left(\partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \right)^2 . \tag{3.11}$$

Since q is fixed, for simplicity we relabel the $1 \leqslant r < n$ particles in $\mathcal{S}_q(2\tau)$ in such a way that $\theta_1(t) < \dots < \theta_r(t)$. First observe that over $\{j : \theta_j(t) \in \mathcal{S}_q(2\tau)\} \setminus \{i\}$, by virtue of the definition of g,

$$\partial_{\theta_i} \partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \geqslant \frac{1}{n^2} \left(1 - \left(\beta + \frac{1}{2} \right) |\theta_i - \theta_j|^2 \right) e^{-\frac{\beta|\theta_j - \theta_i|^2}{2}}$$

$$\geqslant \frac{\delta}{n^2} e^{-\frac{\beta}{2\beta+1}(1-\delta)}$$

$$\geqslant \frac{\delta e^{-\frac{1}{2}}}{n^2}, \tag{3.12}$$

where we used $\sin(x) \leqslant x$, as well as $\cos(x) \geqslant 1 - \frac{x^2}{2}$ when $|x| \leqslant 1$ for the first inequality, and (3.10) for the second. In view of (3.11), we are left with lower

bounding the discrepancy between components of the gradient. Since $\Theta(t) \notin \mathcal{N}_\beta$ for $t \in [0, T)$, there necessarily exists some $q \in \{1, \ldots, k\}$ such that

$$\max_{\theta_a(t), \theta_b(t) \in \mathcal{S}_q(2\tau)} |\theta_a(t) - \theta_b(t)| \geqslant e^{-\frac{\lambda}{2}\beta}. \tag{3.13}$$

With this at hand, let $q \in \{1, \ldots, k\}$ be any index for which the corresponding cap satisfies (3.13). We see that

$$\partial_{\theta_1} \mathsf{E}_\beta(\Theta(t)) \geqslant \frac{1}{n^2} \sum_{j=1}^r \sin(\theta_j(t) - \theta_1(t)) e^{\beta(\cos(\theta_j(t) - \theta_1(t)) - 1)} - \frac{1}{n} e^{-(1-\alpha)\beta}$$

$$\geqslant \frac{1}{n^2} \sin(\theta_r(t) - \theta_1(t)) e^{\beta(\cos(\theta_r(t) - \theta_1(t)) - 1)} - \frac{1}{n} e^{-(1-\alpha)\beta} > 0,$$

and similarly

$$\partial_{\theta_r} \mathsf{E}_\beta(\Theta(t)) \leqslant \sum_{j=1}^r \sin(\theta_j(t) - \theta_r(t)) e^{\beta(\cos(\theta_j(t) - \theta_r(t)) - 1)} + \frac{1}{n} e^{-(1-\alpha)\beta}$$

$$\leqslant -\frac{1}{n^2} \sin(\theta_r(t) - \theta_1(t)) e^{\beta(\cos(\theta_r(t) - \theta_1(t)) - 1)} + \frac{1}{n} e^{-(1-\alpha)\beta} < 0,$$

both by virtue of (3.13) and the choice of λ. We use the following inequality, the proof of which we postpone to after the present one due to its technical nature.

Claim 1. *We have*

$$\max_{\ell \in \{1, \ldots, r\}} |\partial_{\theta_\ell} \mathsf{E}_\beta(\Theta(t))| \leqslant \frac{e}{2} \max \{|\partial_{\theta_1} \mathsf{E}_\beta(\Theta(t))|, |\partial_{\theta_r} \mathsf{E}_\beta(\Theta(t))|\}. \tag{3.14}$$

Now observe first of all that the coordinate j for which $(\partial_{\theta_j} \mathsf{E}(\Theta(t)))^2$ is largest must correspond to a particle $\theta_j(t)$ lying in a spherical cap $\mathcal{S}_q(2\tau)$ satisfying (3.13). Using this information, by virtue of (3.12), and since $\partial_{\theta_r} \mathsf{E}_\beta(\Theta(t))$ and $\partial_{\theta_1} \mathsf{E}_\beta(\Theta(t))$ are of opposite signs and thus

$$(\partial_{\theta_r} \mathsf{E}_\beta(\Theta(t)) - \partial_{\theta_1} \mathsf{E}_\beta(\Theta(t)))^2 \geqslant \max \left\{ (\partial_{\theta_1} \mathsf{E}_\beta(\Theta(t)))^2, (\partial_{\theta_r} \mathsf{E}_\beta(\Theta(t)))^2 \right\},$$

and taking (3.14) into account, using (3.11) we deduce that

$$\sum_{i=1}^n \sum_{j=1}^n \partial_{\theta_i} \partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \left(\partial_{\theta_i} \mathsf{E}_\beta(\Theta(t)) - \partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \right)^2$$

$$\geqslant \kappa(\beta, n) \max_{j \in \{1, \ldots, n\}} \left(\partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \right)^2,$$

where

$$\kappa(\beta, n) := \frac{1}{n} \cdot \left(\frac{\delta e^{\frac{1}{2}}}{2} - 4(1 + \beta) e^{-(1-\alpha)\beta} \right) > 0,$$

because of the assumption on δ given by (3.10). All in all, we gather that

$$\mathsf{H}(t) \leqslant -\frac{\kappa(\beta, n)}{2} \max_{j \in \{1, \ldots, n\}} \left(\partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \right)^2 \leqslant -\frac{\kappa(\beta, n)}{2n} \|\nabla \mathsf{E}_\beta(\Theta(t))\|^2.$$

We can apply Lemma 3.2 to conclude. $\qquad\square$

Proof of Claim 1. We omit time dependence for the sake of readability. Without loss of generality, suppose $|\partial_{\theta_1} \mathsf{E}_\beta(\Theta)| \geqslant |\partial_{\theta_r} \mathsf{E}_\beta(\Theta)|$. Let $\ell \in \{1, \ldots, r\}$, and suppose that $\partial_{\theta_\ell} \mathsf{E}_\beta(\Theta(t)) \partial_{\theta_1} \mathsf{E}_\beta(\Theta) > 0$, We now compute:

$$n^2 \left| \partial_{\theta_j} \mathsf{E}_\beta(\Theta) \right| = n^2 \partial_{\theta_j} \mathsf{E}_\beta(\Theta) \leqslant \sum_{k=1}^r \sin(\theta_j - \theta_k) e^{\beta(\cos(\theta_j - \theta_k) - 1)}$$
$$+ \sum_{k \,:\, \theta_k \notin \mathcal{S}_q(2\tau)} \sin(\theta_j - \theta_k) e^{\beta(\cos(\theta_j - \theta_k) - 1)}.$$

We focus on the first term; recalling that $\theta_1 < \ldots < \theta_r$, we end up with

$$\sum_{k=1}^r \sin(\theta_j - \theta_k) e^{\beta(\cos(\theta_j - \theta_k) - 1)}$$

$$\leqslant \sum_{k=1}^j \sin(\theta_k - \theta_j) e^{\beta(\cos(\theta_j - \theta_k) - 1)} + \sum_{k=j+1}^r \sin(\theta_k - \theta_1) e^{\beta(\cos(\theta_j - \theta_k) - 1)}$$

$$\leqslant \sum_{k=1}^j \sin(\theta_k - \theta_1) e^{\beta(\cos(\theta_1 - \theta_k) - 1)}$$

$$+ e^{\beta|\theta_j - \theta_1| \max_{\ell \in \{1, \ldots, r\}} \sin(\theta_\ell - \theta_1)} \sum_{k=j+1}^r \sin(\theta_k - \theta_1) e^{\beta \cos(\theta_k - \theta_1)}$$

$$\leqslant e^{\beta|\theta_j - \theta_1| \max_{\ell \in \{1, \ldots, r\}} \sin(\theta_\ell - \theta_1)} \sum_{k=1}^r \sin(\theta_k - \theta_1) e^{\beta \cos(\theta_k - \theta_1)},$$

where we used $\frac{\pi}{2} > \theta_k - \theta_1 > \theta_j - \theta_1 > 0$ and the monotonicity of $\sin(\cdot)$ for the first inequality, whereas we used

$$\sum_{k=1}^j \sin(\theta_k - \theta_j) e^{\beta(\cos(\theta_j - \theta_k) - 1)} \leqslant 0 \leqslant \sum_{k=1}^j \sin(\theta_k - \theta_1) e^{\beta(\cos(\theta_j - \theta_k) - 1)}$$

for the second, and

$$|\cos(\theta_j - \theta_k) - \cos(\theta_k - \theta_1)| \leqslant \sin(\theta_k - \theta_1)|\theta_j - \theta_1|$$

for the third (which follows by the mean-value theorem). Besides, by definition of spherical caps we have $|\sin(\theta_j - \theta_i)| \leqslant \beta^{-1/2}$, so we can conclude that

$$\max_{\ell \in \{1, \ldots, r\}} |\partial_{\theta_\ell} \mathsf{E}_\beta(\Theta(t))| \leqslant e \max \{|\partial_{\theta_1} \mathsf{E}_\beta(\Theta(t))|, |\partial_{\theta_r} \mathsf{E}_\beta(\Theta(t))|\} - (1 + \beta) e^{-(1-\alpha)\beta}.$$

Using the fact that $\Theta(t) \notin \mathcal{N}_\beta$, we get

$$\max_{\ell \in \{1, \ldots, r\}} |\partial_{\theta_\ell} \mathsf{E}_\beta(\Theta(t))| \geqslant 2e(1 + \beta) e^{-(1-\alpha)\beta},$$

whence,

$$\max_{\ell \in \{1, \ldots, r\}} |\partial_{\theta_\ell} \mathsf{E}_\beta(\Theta(t))| \leqslant \frac{e}{2} \max \{|\partial_{\theta_1} \mathsf{E}_\beta(\Theta(t))|, |\partial_{\theta_r} \mathsf{E}_\beta(\Theta(t))|\}. \qquad \square$$

We now focus on the lower bound of the energy discrepancy in (3.2). To this end, we simply adapt [OV00, Proposition 1] to the framework of slow manifolds.

Lemma 3.5. *Let* $\mathsf{E} : \mathcal{M} \to \mathbb{R}_{\geqslant 0}$ *be smooth, and let* $\mathcal{N} \subset \mathcal{M}$. *Fix* $u \in \mathcal{M}$ *and consider*

$$\begin{cases} \dot{X}(t) = \nabla\mathsf{E}(X(t)) & \text{for } t \geqslant 0 \\ X(0) = u. \end{cases}$$

Suppose that there exist $v \in \mathcal{N}, T > 0$, *and* $c > 0$ *such that*

$$\mathsf{E}(v) - \mathsf{E}(u) \leqslant \frac{1}{2c}\|\nabla\mathsf{E}(u)\|^2. \tag{3.15}$$

Then,

$$2c\|u - v\|^2 \leqslant \mathsf{E}(v) - \mathsf{E}(u).$$

Proof of Lemma 3.5. Consider

$$\varphi(t) := \|u - X(t)\| + \frac{\sqrt{\mathsf{E}(v) - \mathsf{E}(X(t))))}}{\sqrt{2c}}.$$

We compute

$$\dot{\varphi}(t) = -\left\langle \nabla\mathsf{E}(X(t)), \frac{u - X(t)}{\|u - X(t)\|} \right\rangle - \frac{\|\nabla\mathsf{E}(X(t))\|^2}{\sqrt{2c(\mathsf{E}(v) - \mathsf{E}(X(t)))}}$$

Using (3.15) we get

$$-\frac{\|\nabla\mathsf{E}(X(t)\|^2}{\sqrt{2c(\mathsf{E}(v) - \mathsf{E}(X(t)))}} \leqslant -\|\nabla\mathsf{E}(X(t)\|,$$

and Cauchy-Schwarz,

$$\left| \left\langle \nabla\mathsf{E}(X(t)), \frac{u - X(t)}{\|u - X(t)\|} \right\rangle \right| \leqslant \|\nabla\mathsf{E}(X(t))\|.$$

It follows that φ is non-increasing, and we conclude the proof by evaluating φ at $t = 0$ and $t = T$. $\qquad\square$

As a result of Theorem 3.1 and Lemmas 3.2 and 3.5, we conclude the following.

Corollary 3.6. *Suppose* $\beta > 1$, *and consider a* (β, τ)-*separated configuration* $\Theta(0) \in \mathbb{T}^n$ *for some* $\tau = \tau(\beta)$ *satisfying the conditions of Definition 3.3 as well as* (3.10). *Let* $\mathcal{N}_\beta \subset \mathbb{T}^n$ *be defined as in Lemma 3.4. Then the conclusions of Theorem 3.1 hold for* (3.7) *with* $\delta = e^{-\lambda\beta/2}$, *for* λ *as in Lemma 3.4.*

Remark 3.7 ((SA)). *We demonstrate how the proof of Lemma 3.4 can be adapted to (SA) when $d = 2$. Written in angles, (SA), for $i \in \{1, \ldots, n\}$, reads*

$$\dot{\theta}_i(t) = \sum_{j=1}^{n} \frac{e^{\beta \cos(\theta_j(t) - \theta_i(t))}}{\displaystyle\sum_{k=1}^{n} e^{\beta \cos(\theta_i(t) - \theta_k(t))}} \sin(\theta_j(t) - \theta_i(t)) \qquad \text{for } t \geqslant 0. \qquad (3.16)$$

As implied in the introduction, (SA) is also the gradient flow for E_β, but for a gradient taken with respect to a different metric $\mathfrak{g}$. We do not go into the details here—see [GLPR23]—all we need to know is that

$$\dot{\Theta}(t) = \operatorname{grad}_{\mathfrak{g}} \mathsf{E}_\beta(\Theta(t)) \qquad \text{for } t \geqslant 0,$$

and the i-th coordinate of $\operatorname{grad}_{\mathfrak{g}} \mathsf{E}_\beta(\Theta(t)) \in \mathbb{T}^n$ is precisely the right-hand side in (3.16). With this in hand, we wish to compute the Hessian—with respect to $\mathfrak{g}$—of E_β. Since $\mathbb{T}^n$ a submanifold of $\mathbb{R}^n$, we actually have[5]

$$\operatorname{Hess}_{\mathfrak{g}} \mathsf{E}_\beta(\Theta)[v] = \operatorname{proj}_\Theta \left(\frac{\mathrm{d}}{\mathrm{d}\varepsilon} \mathsf{G}(\Theta + \varepsilon v) \Big|_{\varepsilon = 0} \right)$$

at a point $\Theta \in \mathbb{T}^n$ and direction $v \in \mathsf{T}_\Theta \mathbb{T}^n$. Here $\operatorname{proj}_\Theta$ is the orthogonal projection onto $\mathsf{T}_\Theta \mathbb{T}^n$, and G is any smooth vector field with $\mathsf{G}(\Theta) = \operatorname{grad}_{\mathfrak{g}} \mathsf{E}_\beta(\Theta)$ for $\Theta \in \mathbb{T}^n$. Since $\mathbb{T}^n$ is locally flat, the tangent space can be identified with $\mathbb{R}^n$ itself, and the orthogonal projection is the idenitity map. Whereupon, we can simply use the trivial extension to $\mathbb{R}^n$ of the right-hand side in (3.16) to compute the Hessian: the i-th coordinate of $\operatorname{Hess}_{\mathfrak{g}} \mathsf{E}_\beta(\Theta)[v] \in \mathbb{T}^n$ reads

$$(\operatorname{Hess}_{\mathfrak{g}} \mathsf{E}_\beta(\Theta)[v])_i = \sum_{j=1}^{n} b_{ij}(v_i - v_j), \qquad (3.17)$$

where

$$b_{ij} := a_{ij} \left[\cos(\theta_i - \theta_j) - \beta \sin^2(\theta_i - \theta_j) + \beta \sin(\theta_i - \theta_j) \sum_{k=1}^{n} a_{ik} \sin(\theta_i - \theta_k) \right],$$

and $a_{ij} := e^{\beta \cos(\theta_i - \theta_j)} / \sum_{\ell=1}^{n} e^{\beta \cos(\theta_i - \theta_\ell)}$. We can identify (3.17) with a $n \times n$ matrix that has a Laplacian structure—denoting it again by $\operatorname{Hess}_{\mathfrak{g}} \mathsf{E}_\beta(\Theta)$ and its entries by $\overline{\partial}_{\theta_i} \overline{\partial}_{\theta_j} \mathsf{E}_\beta(\Theta)$, we have

$$\overline{\partial}_{\theta_i} \overline{\partial}_{\theta_j} \mathsf{E}_\beta(\Theta) = \begin{cases} -b_{ij} & i \neq j \\ \displaystyle\sum_{k \neq i} b_{ik} & i = j. \end{cases}$$

Therefore the proof of Lemma 3.4 can be repeated to this case, and one solely needs to check if $\overline{\partial}_{\theta_i} \overline{\partial}_{\theta_j} \mathsf{E}_\beta(\Theta)$ satisfy similar bounds to those by $\partial_{\theta_i} \partial_{\theta_j} \mathsf{E}_\beta(\Theta)$. Because the first two terms in b_{ij} are the same as before, we only have to manage the third term of this expression, and the previous arguments can be adapted to this case as well.

---

[5]For instance, see [Bou23, Chapter 5] for details.

3.3   Acceleration of the gradient between metastable states

The dynamics of separated particles is in fact *accelerating* over time as distances between particles decrease. Actually when particles are sufficiently separated we can show a *reverse PL inequality*.

We first motivate this acceleration in a general framework as before. Suppose $\mathsf{E} : \mathcal{M} \to \mathbb{R}_{\geqslant 0}$ is smooth, fix $u \in \mathcal{M}$, and consider

$$\begin{cases} \dot{X}(t) = \nabla\mathsf{E}(X(t)) & \text{for } t \geqslant 0 \\ X(0) = u. \end{cases}$$

Let $\mathscr{A} \subset \mathcal{M}$ designate the *accelerating* manifold: setting

$$T_u = \inf\{t \geqslant 0 \colon X(t) \notin \mathscr{A}\},$$

suppose, for some $c > 0$ and all $t \in [0, T_u]$, that

$$\langle \operatorname{Hess} \mathsf{E}(X(t))\nabla\mathsf{E}(X(t)), \nabla\mathsf{E}(X(t))\rangle \geqslant c\|\nabla\mathsf{E}(X(t))\|^2. \tag{3.18}$$

We say that a *reverse PL inequality* holds if for all $u \in \mathscr{A}$, there exist $v \notin \mathscr{A}$ and $c > 0$ such that

$$\mathsf{E}(v) - \mathsf{E}(u) \geqslant c\|\nabla\mathsf{E}(v)\|^2.$$

We briefly explain the argument allowing one to establish this inequality. Suppose that (3.18) holds. Then,

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\nabla\mathsf{E}(X(t))\|^2 = 2\langle\nabla\mathsf{E}(X(t)), \operatorname{Hess}\mathsf{E}(X(t))\nabla\mathsf{E}(X(t))\rangle$$
$$\geqslant 2c\|\nabla\mathsf{E}(X(t))\|^2.$$

Using Grönwall's lemma, we get the differential inequality

$$\|\nabla\mathsf{E}(X(t))\|^2 \geqslant e^{2ct}\|\nabla\mathsf{E}(X(0))\|^2,$$

resulting in an acceleration of the gradient[6]. Moreover if $X(t) \notin \mathscr{A}$ for some $t < +\infty$, we then have

$$\mathsf{E}(v) - \mathsf{E}(u) \geqslant c\|\nabla\mathsf{E}(v)\|^2.$$

We can derive a bound of the mould (3.18) for the Hessian of E_β defined in (1.1). Using the shorthand

$$\mathsf{H}(t) := \langle\nabla\mathsf{E}_\beta(X(t)), \operatorname{Hess}\mathsf{E}_\beta(X(t))\nabla\mathsf{E}_\beta(X(t))\rangle,$$

---

[6]One could potentially use such an inequality to answer Problem 2. Indeed, to escape a metastable state, in which we recall the gradient is exponentially small, one needs the gradient to start growing exponentially. We believe that this acceleration mechanism is behind the escape of such metastable states, and thus jumps in the energy level as seen in the staircase profile.

we recall that

$$\mathsf{H}(t) = -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \partial_{\theta_i} \partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \left(\partial_{\theta_i} \mathsf{E}_\beta(\Theta(t)) - \partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \right)^2.$$

For any $i \in \{1, \ldots, n\}$ suppose that there exists $j_i \in \{1, \ldots, n\}$ such that

$$|\theta_{j_i}(t) - \theta_i(t)| = \min_{k \in \{1,\ldots,n\} \setminus \{i\}} |\theta_k - \theta_i| \quad \text{and} \quad \partial_{\theta_i} \mathsf{E}_\beta(\Theta(t)) \partial_{\theta_{j_i}} \mathsf{E}_\beta(\Theta(t)) < 0.$$

Suppose that all particles are separated by at least $\tau(\beta)$; then

$$\begin{aligned}
\mathsf{H}(t) &= -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \partial_{\theta_i} \partial_{\theta_j} \mathsf{E}_\beta(\Theta(t)) \left(\partial_{\theta_i} \mathsf{E}_\beta(\Theta(t)) - \partial_{\theta_{j_i}} \mathsf{E}_\beta(\Theta(t)) \right)^2 \\
&\geqslant -\frac{1}{2} \sum_{i=1}^{n} \partial_{\theta_i} \partial_{\theta_{j_i}} \mathsf{E}_\beta(\Theta(t)) \left(\partial_{\theta_i} \mathsf{E}_\beta(\Theta(t)) - \partial_{\theta_{j_i}} \mathsf{E}_\beta(\Theta(t)) \right)^2 \\
&\geqslant -\frac{\mathsf{L}(t)}{2} \|\nabla \mathsf{E}_\beta(\Theta(t))\|^2.
\end{aligned}$$

where

$$\mathsf{L}(t) := \max_{i \in \{1,\ldots,n\}} \partial_{\theta_i} \partial_{\theta_{j_i}} \mathsf{E}_\beta(\Theta(t)),$$

and where we used the fact that $\partial_{\theta_i} \mathsf{E}_\beta(\Theta(t)) \partial_{\theta_{j_i}} \mathsf{E}_\beta(\Theta(t)) < 0$ and that $g(s)$ is non-positive and increasing for $|s| \geqslant \tau(\beta)$. Then

$$\frac{\mathrm{d}}{\mathrm{d}t} \|\nabla \mathsf{E}_\beta(\Theta(t))\|^2 \geqslant -\frac{\mathsf{L}(t)}{2} \|\nabla \mathsf{E}_\beta(\Theta(t))\|^2.$$

By Grönwall's lemma we deduce

$$\|\nabla \mathsf{E}_\beta(\Theta(t))\|^2 \geqslant \exp\left(-\frac{1}{2} \int_0^t \mathsf{L}(s) \, \mathrm{d}s \right) \|\nabla \mathsf{E}_\beta(\Theta(0))\|^2.$$

4   On the initial configuration

We now discuss a couple of examples of randomly generated initial configurations which may or may not fall in the setting of Definition 1.1.

4.1   Projected Gaussian mixtures

The first case of interest are Gaussian mixtures, namely random variables X with a density of the form

$$f(x) = \frac{1}{r\sqrt{2\pi\sigma^2}} \sum_{i=1}^{r} e^{-\frac{\|x - \sqrt{r}w_i\|^2}{2\sigma^2}} \qquad x \in \mathbb{R}^d, \tag{4.1}$$

where $\sigma > 0$, and $w_1, \ldots, w_r \in \mathbb{S}^{d-1}$ with $r \geqslant 1$.

Definition 4.1. *Suppose $d, n \geqslant 2$, $r \in \{1, \ldots, n\}$ and $\varepsilon > 0$. We say that the configuration $(w_1, \ldots, w_r) \in (\mathbb{S}^{d-1})^n$ is (β, ε)-centered if the corresponding spherical caps $(\mathcal{S}_1(\varepsilon), \ldots, \mathcal{S}_r(\varepsilon))$ satisfy* (1.3) *in Definition 1.1.*

We show the following result.

Proposition 4.2. *Suppose $\beta > 0$, $d, n \geqslant 2$, $r \in \{1, \ldots, n\}$ and $\varepsilon > 0$. Let $(w_1, \ldots, w_r) \in (\mathbb{S}^{d-1})^n$ be (β, ε)-centered per Definition 4.1. Let $X_1, \ldots, X_n$ be i.i.d. random variables following the Gaussian mixture law with density* (4.1) *and such that*

$$\frac{6\delta\sqrt{d}}{1 + \delta\sqrt{d}} + \delta\sqrt{2d \log n} \leqslant \varepsilon,$$

where $\delta := \frac{\sigma}{\sqrt{r}}$. Then

$$\mathbb{P}\left(\left(\frac{X_1}{\|X_1\|}, \ldots, \frac{X_n}{\|X_n\|}\right) \text{ is } (\beta, \varepsilon)\text{-separated}\right) \geqslant 1 - 2e^{-d}.$$

Proof of Proposition 4.2. We can write

$$X = \sum_{k=1}^{r} \varepsilon_k Z_k,$$

where $Z_1, \ldots, Z_r$ are independent $\mathcal{N}(w_i, \sigma I_d)$ random variables, whereas $\varepsilon_1, \ldots, \varepsilon_k$ are random variables defined as

$$\varepsilon_i = \mathbf{1}\left(\sum_{q=1}^{i-1} p_q \leqslant U \leqslant \sum_{q=1}^{i} p_q\right)$$

for $i \in \{1, \ldots, k\}$, where U is a random variable following the uniform distribution on $[0, 1]$. We also define $(N_1, \ldots, N_r) \sim \mathcal{N}(0, I_d)$ as

$$Z_i = w_i + \sigma N_i$$

for $i \in \{1, \ldots, r\}$. Consider a fixed $i \in \{1, \ldots, r\}$. Conditioned on the event $\left\{U \in \left[\sum_{q=1}^{i-1} p_q, \sum_{q=1}^{i} p_q\right]\right\}$, we can write X as a function of standard Gaussian variables:

$$\min_{1 \leqslant j \leqslant r} \left\|\frac{X}{\|X\|} - w_j\right\|^2 \leqslant \left\|\frac{Z_i}{\|Z_i\|} - w_i\right\|^2 = f(Z_i).$$

We can also show that f is roughly $\frac{1}{\|X\|}$-Lipschitz:

$$|f(X) - f(Y)| = \left\|\frac{X}{\|X\|} - w_i\right\|^2 - \left\|\frac{Y}{\|Y\|} - w_i\right\|^2 = 2\left|\left\langle w_i, \frac{X}{\|X\|} - \frac{Y}{\|Y\|}\right\rangle\right|.$$

Then

$$|f(X) - f(Y)| \leqslant 2\left\|\frac{Y}{\|Y\|} - \frac{X}{\|X\|}\right\| \leqslant \frac{\|X - Y\|}{\min\{\|X\|, \|Y\|\}}.$$

30

Focusing on the event $\{\|X\|\| \geqslant x'\}$, by the Gausian concentration inequality [BLM13, Theorem 5.6] we have

$$\mathbb{P}\left(f(Z_i) - \mathbb{E}[f(Z_i)] \geqslant t, \|Z_i\| \geqslant x' \,\middle|\, U \in \left[\sum_{q=1}^{i-1} \lambda_q, \sum_{q=1}^{i} \lambda_q\right]\right) \leqslant e^{-\frac{(x' \cdot t)^2}{2\sigma^2}}.$$

Whence,

$$\mathbb{P}\left(\min_{1 \leqslant j \leqslant r} \left\|\frac{X}{\|X\|} - w_j\right\|^2 - \mathbb{E}[f(Z_i)] \geqslant t, \|Z_i\| \geqslant x' \,\middle|\, U \in \left[\sum_{q=1}^{i-1} \lambda_q, \sum_{q=1}^{i} \lambda_q\right]\right)$$
$$\leqslant e^{-\frac{(x' \cdot t)^2}{2\sigma^2}}.$$

We then have by union bound

$$\mathbb{P}\left(\min_{1 \leqslant j \leqslant r} \left\|\frac{X}{\|X\|} - w_j\right\|^2 - \mathbb{E}[f(Z_i)] \geqslant t \,\middle|\, U \in \left[\sum_{q=1}^{i-1} \lambda_q, \sum_{q=1}^{i} \lambda_q\right]\right)$$
$$\leqslant e^{-\frac{(x' \cdot t)^2}{2\sigma^2}} + \mathbb{P}\left(\|Z_i\| \leqslant x'\right).$$

We can use the Gaussian concentration inequality (applied to the 1-Lipschitz function $x \mapsto \|x\|$) to bound the rightmost term, for all $i \in \{1, \ldots, n\}$, as

$$\mathbb{P}\Big(\|w_i + \sigma N_i\| \leqslant \mathbb{E}[\|w_i + \sigma N_i\|] - t\Big) \leqslant e^{-\frac{t^2}{2\sigma^2}}.$$

Using the triangle inequality,

$$\mathbb{P}\Big(\sigma\|Z_i\| \leqslant -\sigma\mathbb{E}[\|N_i\|] + \|w_i\| - t\Big) \leqslant \mathbb{P}\left(\|Z_i\| \leqslant \mathbb{E}[\|Z_i\|] - t\right),$$

and we then get

$$\mathbb{P}\left(\|Z_i\| \leqslant \sqrt{r} - \sigma\sqrt{d} - t\right) \leqslant e^{-\frac{t^2}{2\sigma^2}}.$$

So,

$$\mathbb{P}\left(\|Z_i\| \leqslant x'\right) \leqslant e^{-\frac{1}{2\delta^2}\left(1 - \delta\sqrt{d} - \frac{x'}{\sqrt{r}}\right)^2}.$$

We now bound $\mathbb{E}[f(Z_i)]$. Note that

$$1 - \frac{f(Z_i)}{2} = \left\langle \frac{Z_i}{\|Z_i\|}, w_i \right\rangle = \frac{1}{\|w_i + \delta N_i\|} + \frac{2\delta\langle N_i, w_i \rangle}{\|w_i + \delta N_i\|}.$$

We bound the first term from below as

$$\frac{1}{\|w_i + \delta N_i\|} \geqslant \frac{1}{1 + \delta\|N_i\|},$$

31

and the second term from above as

$$\left| \frac{2\delta\langle N_i, w_i\rangle}{\|w_i + \delta N_i\|} \right| \leqslant \frac{2\delta\|N_i\|}{1 + \delta\|N_i\|}.$$

Because of convexity of $x \mapsto \frac{1}{1+\delta x}$ and of concavity of $x \mapsto \frac{x}{1+\delta x}$, an application of Jensen's inequality yields

$$\mathbb{E}\left[1 - \frac{f(Z_i)}{2}\right] \geqslant \frac{1 - 2\delta\sqrt{d}}{1 + \delta\sqrt{d}} = 1 - \frac{3\delta\sqrt{d}}{1 + \delta\sqrt{d}}.$$

And, so

$$\mathbb{E}[f(Z_i)] \leqslant \frac{6\delta\sqrt{d}}{1 + \delta\sqrt{d}}.$$

Combining all the bounds, we end up with

$$\mathbb{P}\left(\min_{1 \leqslant j \leqslant r}\left\|\frac{X}{\|X\|} - w_j\right\|^2 \geqslant \frac{6\delta\sqrt{d}}{1 + \delta\sqrt{d}} + t\right) \leqslant e^{-\frac{(x' \cdot t)^2}{2\sigma^2}} + e^{-\frac{1}{2\delta^2}\left(1 - \delta\sqrt{d} - \frac{x'}{\sqrt{r}}\right)^2}.$$

We can consider $x' = \sqrt{r}\left(1 - \delta\sqrt{d} + t\right)$ to get

$$\mathbb{P}\left(\min_{1 \leqslant j \leqslant r}\left\|\frac{X}{\|X\|} - w_j\right\|^2 \geqslant \frac{6\delta\sqrt{d}}{1 + \delta\sqrt{d}} + t\right) \leqslant e^{-\frac{t^2}{2\delta^2}} + e^{-\frac{1}{2\delta^2}\left(1 - \delta\sqrt{d} + t\right)^2}$$

Now, there exist $\varepsilon_1^i, \ldots, \varepsilon_r^i$ which follow the law as ε_k above such that for all $i \in \{1, \ldots, n\}$

$$X_i := \sum_{k=1}^{r} \varepsilon_k^i Z_k^i, \tag{4.2}$$

where $Z_k^i \sim \mathcal{N}(w_k, \sigma_k)$, $w_k \in \sqrt{r}\mathbb{S}^{d-1}$ and $\sigma_k > 0$. We consider the random variable Z defined as

$$Z := \max_{1 \leqslant i \leqslant n} \min_{1 \leqslant j \leqslant r}\left\|\frac{X_i}{\|X_i\|} - w_j\right\|^2.$$

By the union bound we get

$$\mathbb{P}\left(Z \geqslant \frac{6\delta\sqrt{d}}{1 + \delta\sqrt{d}} + t\right) \leqslant n\left(e^{-\frac{t^2}{2\delta^2}} + e^{-\frac{1}{2\delta^2}\left(1 - \delta\sqrt{d} + t\right)^2}\right).$$

Because of the fact that $1 - \delta\sqrt{d} > 0$, we get

$$\mathbb{P}\left(Z \geqslant \frac{6\delta\sqrt{d}}{1 + \delta\sqrt{d}} + t\right) \leqslant 2ne^{-\frac{t^2}{2\delta^2}}.$$

Taking $t = \delta\sqrt{2d \log n}$, we find

$$\mathbb{P}\left(Z \geqslant \frac{6\delta\sqrt{d}}{1 + \delta\sqrt{d}} + \delta\sqrt{2d \log n}\right) \leqslant 2e^{-d}.$$

Noticing that we have

$$\{Z \leqslant \varepsilon\} = \left\{\left(\frac{X_1}{\|X_1\|}, \ldots, \frac{X_n}{\|X_n\|}\right) \text{ is } (\beta, \varepsilon)\text{-separated}\right\},$$

we obtain the desired result. $\qquad\square$

4.2 Uniformly distributed points

The second example which we discuss is that of uniformly distributed points.

4.2.1 High dimension

Recall the following consequence of the concentration of measure phenomenon.

Proposition 4.3. *Suppose $n \geqslant 2$. Then there exists some $d^*(n) > n$ such that for all $d \geqslant d^*(n)$, the following holds. Consider a sequence $(x_1, \ldots, x_n)$ of n i.i.d. uniformly distributed points on $\mathbb{S}^{d-1}$. Then, with probability at least $1 - 2n^2 d^{-1/64}$, there exist $(w_1, \ldots, w_n) \in (\mathbb{S}^{d-1})^n$ which are pairwise orthogonal ($\langle w_i, w_j \rangle = \delta_{ij}$), such that*

$$\|x_i - w_i\| \leqslant \sqrt{\frac{4 \log d}{d}}.$$

Proof. See Step 2 in the proof of Theorem 6.9 in [GLPR23]. $\qquad\square$

The following then holds.

Corollary 4.4. *Suppose $n \geqslant 2$. Then there exists some $d^*(n) > n$ such that for all $d \geqslant d^*(n) \vee 381$ and $\beta > 0$ satisfying*

$$\frac{16 \log^2 d}{d^2} + \frac{40 \log d}{d} + \frac{1}{\beta} \log\left(\frac{n^2 d}{2 \log d}\right) < 1, \tag{4.3}$$

the following holds. Consider a sequence $(x_1, \ldots, x_n)$ of n i.i.d. uniformly distributed points on $\mathbb{S}^{d-1}$. Then with probability at least $1 - 2n^2 d^{-1/64}$, $(x_1, \ldots, x_n)$ is (β, ε)-separated in the sense of Definition 1.1 with $\varepsilon = 4 \log d/d$.

Proof of Corollary 4.4. Since $d \geqslant 381$ we have $\varepsilon < \frac{1}{16}$. According to Proposition 4.3, there exist unit vectors $w_1, \ldots, w_n$ such that

$$x_i \in \bigcup_{q=1}^{n} \mathcal{S}_q(\varepsilon).$$

For $\alpha(\varepsilon)$ defined as in (1.4) we have $\alpha(\varepsilon) \leqslant \varepsilon^2 + 2\varepsilon$, and (4.3) is then simply a rewriting of (1.3). $\qquad\square$

Remark 4.5 (Freezing). *Corollary 4.4 has as a consequence that particles initialized uniformly at random when $d \gg n$ remain frozen and do not move for exponentially long times. This is remeniscent to the case of zero temperature ($\beta = +\infty$), in which all configurations are stationary.*

4.2.2 Low dimension

We comment on the case $d < n$ by specializing to the circle ($d = 2$). It can be seen that the probability of having separated configurations decays exponentially with n.

Claim 2. *Fix $\beta > 0$ and let $(x_1, \ldots, x_n)$ be n i.i.d uniformly distributed points on $\mathbb{S}^1$. Then, for $\varepsilon \in (0, \frac{1}{16})$ and $k \leqslant n$, there exists some $c \in (0, 1)$ such that*

$$\mathbb{P}\Big((x_1, \ldots, x_n) \text{ is } (\beta, \varepsilon)\text{-separated }\Big) \leqslant c^n.$$

We briefly explain how to heuristically derive this bound. Using independence of the random variables, we can first compute the probability that there are k points satisfying the (β, ε)-separated hypothesis:

$$\mathbb{P}\Big((X_1, \ldots, X_k) \text{ is } (\beta, \varepsilon)\text{-separated}\Big) \leqslant (1 - 2\alpha - 8\varepsilon)^{r-1}.$$

We also have that

$$\mathbb{P}\left(X_{k+1}, \ldots, X_n \in \bigcup_{i=1}^{k} [X_i - \varepsilon, X_i + \varepsilon] \;\middle|\; (X_1, \ldots, X_k) \text{ is } (\beta, \varepsilon)\text{-separated} \right)$$
$$= (2r\varepsilon)^{n-r}.$$

Thence

$$(2r\varepsilon)^{n-1} \leqslant \mathbb{P}\Big((x_1, \ldots, x_n) \text{ is } (\beta, \varepsilon)\text{-separated}\Big) \leqslant (1 - \alpha - 4\varepsilon)^{n-1}.$$

This may be a fundamental limitation of the spherical cap framework, and raises the question on the sharp assumption needed for the initial configuration to have metastability when d if fixed and $n \gg 1$.

4.3 A discussion on energy levels

In view of many of the previous considerations, it is natural to look for an assumption on the initial condition, yielding metastability, written solely in terms of the energy. We posit the following question.

Problem 1. *Fix $d, n \geqslant 2$ and $\beta > 0$. Let $U_1, \ldots, U_n$ be n i.i.d random variables following the uniform distribution on $\mathbb{S}^{d-1}$. Can one find $1 > c_2 > c_1 > 0$ depending on β such that for all $(x_1, \ldots, x_n) \in (\mathbb{S}^{d-1})^n$ satisfying*

$$c_2 \geqslant \mathsf{E}_\beta(x_1, \ldots, x_n) - \mathbb{E}\big[\mathsf{E}_\beta(U_1, \ldots, U_n)\big] \geqslant c_1,$$

metastability, as stated in Theorem 1.2, holds?

One way to interpret this condition is that any configuration which breaks the symmetry of uniformly distributed random points will lead to metastability.

On one hand, it is not obvious to see if one can simply truncate the gradient over different energy levels instead of spherical caps in our proof. On the other hand, the energetic assumption on the initial configuration is weaker than the one given in Definition 1.1, as the latter implies

$$c - \frac{ke^{-(1-\alpha)\beta}}{n^2} \geqslant \mathsf{E}_\beta(X_1, \ldots, X_n) \geqslant \frac{1}{n} + \frac{ke^{-8\beta\varepsilon}}{n^2} = O(e^{-8\beta\varepsilon}).$$

The converse can then be asked, should one wish to retain the proof of Theorem 1.2 as is—namely, does an energetic assumption as the one above imply quantitative clustering of the configuration in the mould of Definition 1.1? One approach could involve the so-called *Stolarsky invariance principle* [BDM18, BD19].

5 The mean-field regime

For the sake of generality, we now demonstrate that dynamic metastability also holds in the mean-field regime. Consider

$$\begin{cases} \partial_t \mu(t) + \operatorname{div}\left(v[\mu(t)]\mu(t)\right) = 0 & \text{on } \mathbb{R}_{\geqslant 0} \times \mathbb{S}^{d-1} \\ \mu(0) = \mu_0 & \text{on } \mathbb{S}^{d-1}, \end{cases} \tag{5.1}$$

where $-\operatorname{div}$ is the adjoint of the spherical gradient ∇, and

$$v[\mu](x) = \int \frac{e^{\beta\langle x, x'\rangle}}{\int e^{\beta\langle x, \zeta\rangle}\mu(\mathrm{d}\zeta)} \mathbf{P}_x^\perp(x')\mu(\mathrm{d}x')$$

for $x \in \mathbb{S}^{d-1}$. (All arguments carry through for the mean-field analogue of (USA).) We recall that (5.1) is well-posed in the sense that for any $\mu_0 \in \mathscr{P}(\mathbb{S}^{d-1})$ there exists a unique weak solution $\mu \in \mathscr{C}^0(\mathbb{R}_{\geqslant 0}; \mathscr{P}(\mathbb{S}^{d-1}))$. Equation (5.1) can also be seen as the mean-field limit for (SA) when $n \to +\infty$, a limit which is fully rigorous due to classical Dobrushin estimates. We refer the reader to [GLPR24, GLPR23] for all the details.

We consider the following generalization of Definition 1.1.

Definition 5.1. *Let $\beta > 1$ and $\varepsilon \in (0, \frac{1}{16})$. We say $\mu_0 \in \mathscr{P}(\mathbb{S}^{d-1})$ is a (β, ε)-separated measure if there exist $k \leqslant n$ points $w_1, \ldots, w_k \in \mathbb{S}^{d-1}$ and measures $\nu_1, \ldots, \nu_k \in \mathscr{P}(\mathbb{S}^{d-1})$ satisfying*

$$\operatorname{supp}(\nu_q) \subset \mathcal{S}_q(\varepsilon),$$

with $\mathcal{S}_q(\varepsilon)$ denoting the spherical caps of Definition 1.1 centered at w_q, such that

$$\mu_0 = \frac{1}{k}\sum_{q=1}^k \nu_q,$$

holds, where

$$\gamma(\beta) := 1 - \alpha - 8\varepsilon - \frac{1}{\beta} \log\left(\frac{2k^2}{\varepsilon}\right) > 8\varepsilon \quad \textit{and} \quad \gamma(\beta) = \Omega(1), \qquad (5.2)$$

with

$$\alpha := \max_{\substack{(x,y)\in\mathcal{S}_i(2\varepsilon)\times\mathcal{S}_j(2\varepsilon) \\ i\neq j\in\{1,\ldots,k\}}} \langle x, y \rangle. \qquad (5.3)$$

The following partial generalization of Theorem 1.2 holds.

Theorem 5.2. *Let $\beta > 1$, and let $\mu_0 \in \mathscr{P}(\mathbb{S}^{d-1})$ be a $(\beta, \varepsilon(\beta))$-separated measure for some $\varepsilon(\beta) \in (0, \frac{1}{16})$. Let $\mu \in \mathscr{C}^0(\mathbb{R}_{\geqslant 0}; \mathscr{P}(\mathbb{S}^{d-1}))$ denote the corresponding unique solution to (5.1). Then there exist $T_2 > T_1 > 0$ with*

$$T_1 < \frac{\varepsilon}{k} e^{\beta(1-\alpha-8\varepsilon)} \quad \textit{and} \quad T_2 > \frac{\varepsilon}{k} e^{\beta(1-\alpha-8\varepsilon)},$$

such that for any $q \in \{1, \ldots, k\}$,

$$\mathrm{supp}\left(\left(\Phi_{v[\mu(t)]}^t\right)_{\#} \nu_q\right) \subset \mathcal{S}_q(2\varepsilon)$$

for all $t \in [0, T_2]$, where $\Phi_{v[\mu(t)]}^t$ is the flow map defined in (5.4), as well as

$$\int_{\mathcal{S}_q(2\varepsilon)} \left\| \Phi_{v[\mu(t)]}^t(x') - \operatorname*{arg\,min}_{x \in \Phi_{v[\mu(t)]}^t(\mathcal{S}_q(\varepsilon))} \langle x, w_q \rangle \right\|^2 \mu_0(\mathrm{d}x') \leqslant e^{-\lambda\beta}$$

for all $t \in [T_1, T_2]$ and for all $0 < \lambda < \gamma$, where $\gamma = \gamma(\beta) > 0$ is defined in (5.2).

Before proceeding with the proof we make a couple of comments.

Remark 5.3. *Theorem 5.2 differs slightly from Theorem 1.2 in that 1). the collapse time T_1 is of the same order of magnitude as the escape time T_2, and 2). only the variance of the particles within a cap is exponentially small. Both are due to the fact that we only study the distance of the particle farthest to the center of the spherical cap, as in Step 1 of the proof of Theorem 1.2. Since Theorem 5.2 serves only to illustrate the generality of the metastability phenomenon, we circumvented a complete generalization thereof, which only requires additional technicalities.*

Remark 5.4 ((Sub-)Gaussian case)**.** *One can naturally inquire about generalizing the above result to measures ν_q which are not exactly supported in $\mathcal{S}_q(\varepsilon)$, but have "most" of their mass in $\mathcal{S}_q(\varepsilon)$. A case of interest is the Gaussian mixture law on $\mathbb{S}^{d-1}$ with density*

$$\rho(x) := \frac{1}{k} \sum_{q=1}^{k} \frac{1}{\mathcal{Z}_q} e^{-\frac{\|x-w_q\|^2}{2\sigma_q^2}},$$

where $\mathcal{Z}_q$ is the normalizing constant. This example eluded our proof due to the difficulty of lower bounding the partition function $\mathcal{Z}_{\beta,\mu(t)}(x)$, partly due to possible interactions with particles outside the cap $\mathcal{S}_q(\varepsilon)$. We leave this question open.

Proof of Theorem 5.2. We recall that since (5.1) is well-posed, given the solution $\mu \in \mathscr{C}^0(\mathbb{R}_{\geqslant 0}; \mathscr{P}(\mathbb{S}^{d-1}))$, we know that any $x(t) \in \text{supp}(\mu(t))$ satisfies

$$\dot{x}(t) = v[\mu(t)](x(t)) \qquad \text{for } t \geqslant 0.$$

We can define the Lipschitz-continuous and invertible map $\Phi^t_{v[\mu(t)]} : x(0) \mapsto x(t)$, and then

$$\mu(t) = \left(\Phi^t_{v[\mu(t)]} \right)_\# \mu_0. \tag{5.4}$$

With this at hand, the proof is an adaptation of that of Theorem 1.2, mostly by replacing sums with integrals. We provide some details nonetheless.

Step 1. Lower-bounding the escape time

For $q \in \{1, \ldots, k\}$ we define

$$\mathscr{B}_q(t) := \Phi^t_{v[\mu(t)]} \left(\mathcal{S}_q(\varepsilon) \right).$$

Just as before,

$$T_{\text{esc}} := \left\{ t \geqslant 0 \colon \exists q \in \{1, \ldots, k\} \text{ such that } \mathscr{B}_q(t) \not\subset \bigcup_{q=1}^k \mathcal{S}_q(2\varepsilon) \right\}.$$

For $q \in \{1, \ldots, k\}$ we also define

$$T_{\text{esc}}(q) := \inf \left\{ t \geqslant 0 \colon \mathscr{B}_q(t) \not\subset \mathcal{S}_q(2\varepsilon) \right\}.$$

Observe that

$$T_{\text{esc}} = \min_{q \in \{1, \ldots, k\}} T_{\text{esc}}(q).$$

So let $q \in \{1, \ldots, k\}$ be arbitrary. We define

$$\eta_q(t) := \min_{x \in \mathscr{B}_q(t)} \langle x, w_q \rangle,$$

and take

$$x(t) \in \arg\min_{x \in \mathscr{B}_q(t)} \langle x, w_q \rangle.$$

Set $\mathfrak{Z}_{\beta,\mu}(x) := \int e^{\beta \langle x, x' \rangle} \mu(\,dx')$. We compute the derivative of η_q as

$$\dot{\eta}_q(t) = \left\langle v[\mu(t)](x(t)), w_q \right\rangle$$
$$= \frac{1}{\mathfrak{Z}_{\beta,\mu(t)}(x(t))} \int e^{\beta \langle x', x(t) \rangle} \left\langle \mathbf{P}^\perp_{x(t)}(x'), w_q \right\rangle \mu(t,\,dx').$$

Using $|\langle \mathbf{P}^{\perp}_{x(t)}(x'), w_q \rangle| \leqslant 1$ and the change of variable formula, we find

$$\dot{\eta}_q(t) \geqslant \frac{1}{\mathcal{F}_{\beta,\mu(t)}(x(t))} \int_{\mathcal{S}_q(2\varepsilon)} e^{\beta\langle \Phi^t_{v[\mu(t)]}(x'),x(t)\rangle} \left\langle \mathbf{P}^{\perp}_{x(t)} \left(\Phi^t_{v[\mu(t)]}(x') \right), w_q \right\rangle \mu_0(\,\mathrm{d}x')$$
$$- \frac{1}{\mathcal{F}_{\beta,\mu(t)}(x(t))} \sum_{r\in\{1,\dots,k\}\setminus\{q\}} \int_{\mathcal{S}_r(2\varepsilon)} e^{\beta\langle \Phi^t_{v[\mu(t)]}(x'),x(t)\rangle} \mu_0(\,\mathrm{d}x').$$

For $t \in [0, T_{\mathrm{esc}}]$ and $x \in \mathcal{S}_q(2\varepsilon)$ we have

$$\mathcal{F}_{\beta,\mu(t)}(x) \geqslant \int_{\mathcal{B}_q(t)} e^{\beta\langle x,x'\rangle} \mu(t,\,\mathrm{d}x') \geqslant e^{(1-8\varepsilon)\beta} \int_{\mathcal{S}_q(2\varepsilon)} \mu_0(\,\mathrm{d}x') = \frac{1}{k} e^{(1-8\varepsilon)\beta},$$

and also

$$\sum_{r\in\{1,\dots,k\}\setminus\{q\}} \int_{\mathcal{S}_r(2\varepsilon)} e^{\beta\langle x',x\rangle} \mu(t,\,\mathrm{d}x') \leqslant e^{\alpha\beta}.$$

Using these two inequalities, we get

$$\dot{\eta}_q(t) \geqslant \frac{1}{\mathcal{F}_{\beta,\mu(t)}(x(t))} \int_{\mathcal{S}_q(2\varepsilon)} e^{\beta\langle \Phi^t_{v[\mu(t)]}(x'),x(t)\rangle} \left\langle \mathbf{P}^{\perp}_{x(t)} \left(\Phi^t_{v[\mu(t)]}(x') \right), w_q \right\rangle \mu_0(\,\mathrm{d}x')$$
$$- ke^{-(1-\alpha-8\varepsilon)\beta}.$$

Now as in Step 1 of the proof of Theorem 1.2, since $x(t) \in \arg\min_{x\in\mathcal{B}_q(t)}\langle x, w_q \rangle$, we have

$$\left\langle \mathbf{P}^{\perp}_{x(t)} \left(\Phi^t_{v[\mu(t)]}(x') \right), w_q \right\rangle \geqslant 0$$

for all $x' \in \mathcal{S}_q(2\varepsilon)$. Thus

$$\dot{\eta}_q(t) \geqslant -ke^{-(1-\alpha-8\varepsilon)\beta}.$$

The same argument as in Step 1 of the proof of Theorem 1.2 then yields

$$T_{\mathrm{esc}} \geqslant \frac{\varepsilon}{k} e^{(1-\alpha-8\varepsilon)\beta}.$$

Step 2. The variance is decreasing

Let $t \in [0, T_{\mathrm{esc}}]$. From the previous step,

$$\dot{\eta}_q(t) \geqslant \frac{1}{\mathcal{F}_{\beta,\mu(t)}(x(t))} \int_{\mathcal{S}_q(2\varepsilon)} e^{\beta\langle \Phi^t_{v[\mu(t)]}(x'),x(t)\rangle} \left\langle \mathbf{P}^{\perp}_{x(t)}(\Phi^t_{v[\mu(t)]}(x')), w_q \right\rangle \mu_0(\,\mathrm{d}x')$$
$$- ke^{-(1-\alpha-8\varepsilon)\beta}$$
$$=: (a) - ke^{-(1-\alpha-8\varepsilon)\beta}.$$

38

Elementary algebraic manipulations yield

$$(a) \geqslant \frac{1}{\mathcal{Z}_{\beta,\mu(t)}(x(t))} \int_{\mathcal{S}_q(2\varepsilon)} e^{\beta\langle \Phi^t_{v[\mu(t)]}(x'), x(t)\rangle} \langle x(t), w_q\rangle \frac{\left\| \Phi^t_{v[\mu(t)]}(x') - x(t)\right\|^2}{2} \mu_0(\,\mathrm{d}x')$$

$$\geqslant \frac{\eta_q(t)}{\mathcal{Z}_{\beta,\mu(t)}(x(t))} \int_{\mathcal{S}_q(2\varepsilon)} e^{\beta\langle \Phi^t_{v[\mu(t)]}(x'), x(t)\rangle} \frac{\left\| \Phi^t_{v[\mu(t)]}(x') - x(t)\right\|^2}{2} \mu_0(\,\mathrm{d}x').$$

Then

$$\dot{\eta}_q(t) \geqslant \frac{\eta_q(t)}{2\mathcal{Z}_{\beta,\mu(t)}(x(t))} \int_{\mathcal{S}_q(2\varepsilon)} e^{\beta\langle \Phi^t_{v[\mu(t)]}(x'), x(t)\rangle} \left\| \Phi^t_{v[\mu(t)]}(x') - x(t)\right\|^2 \mu_0(\,\mathrm{d}x')$$
$$- k e^{-(1-\alpha-8\varepsilon)\beta}.$$

Therefore,

$$\dot{\eta}_q(t) \geqslant \frac{k}{2}\eta_q(t) \int_{\mathcal{S}_q(2\varepsilon)} e^{\beta(\langle \Phi^t_{v[\mu(t)]}(x'), x(t)\rangle - 1 + 8\varepsilon)} \left\| \Phi^t_{v[\mu(t)]}(x') - x(t)\right\|^2 \mu_0(\,\mathrm{d}x')$$
$$- k e^{-(1-\alpha-8\varepsilon)\beta}.$$

Since for all $x' \in \mathcal{S}_q(2\varepsilon)$ we have

$$e^{\beta(\langle \Phi^t_{v[\mu(t)]}(x'), x(t)\rangle - 1 + 8\varepsilon)} \geqslant e^{\beta(\eta_q(t) - 1 + 8\varepsilon)},$$

we deduce that

$$\dot{\eta}_q(t) \geqslant \frac{k}{2}\eta_q(t) e^{-(1-\eta_q(t)-8\varepsilon)\beta} \int_{\mathcal{S}_q(2\varepsilon)} \left\| \Phi^t_{v[\mu(t)]}(x') - x(t)\right\|^2 \mu_0(\,\mathrm{d}x')$$
$$- k e^{-(1-\alpha-8\varepsilon)\beta}.$$

For $q \in \{1, \ldots, k\}$, we define

$$\mathsf{V}_q(t) := \frac{1}{2} \int_{\mathcal{S}_q(2\varepsilon)} \left\| \Phi^t_{v[\mu(t)]}(x') - x(t)\right\|^2 \mu_0(\,\mathrm{d}x').$$

Then

$$\dot{\eta}_q(t) \geqslant k e^{8\varepsilon\beta} \left(\eta_q(t) e^{-(1-\eta_q(t))\beta} \mathsf{V}_q(t) - e^{-(1-\alpha)\beta}\right). \tag{5.5}$$

For $q \in \{1, \ldots, k\}$ and $c > 0$, we define

$$T_*(q, c) := \inf \left\{ t \in [0, T_{\mathrm{esc}}] \colon \eta_q(t)\mathsf{V}_q(t) e^{-(1-\eta_q(t))\beta} \leqslant 2 e^{-c\beta}\right\}.$$

Claim 3. *We have*

$$\left\{ t \in [0, T_{\mathrm{esc}}] \colon \eta_q(t)\mathsf{V}_q(t) e^{-(1-\eta_q(t))\beta} \leqslant 2 e^{-c\beta}\right\} \neq \varnothing$$

and

$$\inf \left\{ t \in [0, T_{\mathrm{esc}}] \colon \eta_q(t)\mathsf{V}_q(t) e^{-(1-\eta_q(t))\beta} \leqslant 2 e^{-c\beta}\right\} < \frac{4\varepsilon}{k} e^{(c-8\varepsilon)\beta}.$$

39

We provide the proof after the present one. Setting $c := \lambda$ for an arbitrary but fixed $\lambda \in (8\varepsilon, \gamma)$—this is without loss of generality, since if the bound in the statement holds for $\lambda > 8\varepsilon$, it also holds for $\lambda \leqslant 8\varepsilon$—, using the fact that $\eta_q(T_*(q,c)) \geqslant 1 - 8\varepsilon > \frac{1}{2}$ we find

$$\mathsf{V}_q(T_*(q,c)) \leqslant e^{-\lambda\beta}. \tag{5.6}$$

Step 3. Propagation of smallness

We can conclude as in Step 4 in the proof of Theorem 1.2. Consider

$$T := \inf\left\{ t \geqslant T_*(q,c) \colon \mathsf{V}_q(t) \geqslant e^{-\lambda\beta} \right\},$$

and suppose that $T < T_{\mathrm{esc}}$. By continuity, we have $\mathsf{V}_q(T) = e^{-\lambda\beta}$. We can compute the derivative of V_q at a given time t as

$$\dot{\mathsf{V}}_q(t) = -2 \int_{\mathcal{S}_q(2\varepsilon)} \frac{\mathrm{d}}{\mathrm{d}t} \langle x(t), \Phi^t_{v[\mu(t)]}(x') \rangle \mu_0(\,\mathrm{d}x')$$

$$= -2 \int_{\mathcal{S}_q(2\varepsilon)} \left(\langle \dot{x}(t), \Phi^t_{v[\mu(t)]}(x') \rangle + \left\langle x(t), \frac{\mathrm{d}}{\mathrm{d}t} \Phi^t_{v[\mu(t)]}(x') \right\rangle \right) \mu_0(\,\mathrm{d}x').$$

We begin with the left term in the above identity:

$$\int_{\mathcal{S}_q(2\varepsilon)} \langle \dot{x}(t), \Phi^t_{v[\mu(t)]}(x') \rangle \mu_0(\,\mathrm{d}x') = \int_{\mathcal{S}_q(2\varepsilon)} \langle v[\mu(t)](x(t)), \Phi^t_{v[\mu(t)]}(x') \rangle \mu_0(\,\mathrm{d}x').$$

Further computations yield

$$\int_{\mathcal{S}_q(2\varepsilon)} \langle v[\mu(t)](x(t)), \Phi^t_{v[\mu(t)]}(x') \rangle \mu_0(\,\mathrm{d}x')$$

$$= \int_{\mathcal{S}_q(2\varepsilon)} \left\langle \int_{\mathcal{B}_q(t)} \frac{e^{\beta\langle x(t),y \rangle}}{\mathcal{Z}_{\beta,\mu(t)}(x(t))} \mathbf{P}^\perp_{x(t)}(y)\mu(t,\,\mathrm{d}y), \Phi^t_{v[\mu(t)]}(x') \right\rangle \mu_0(\,\mathrm{d}x')$$

$$+ \sum_{i \neq q} \int_{\mathcal{S}_q(2\varepsilon)} \left\langle \int_{\mathcal{B}_i(t)} \frac{e^{\beta\langle x(t),y \rangle}}{\mathcal{Z}_{\beta,\mu(t)}(x(t))} \mathbf{P}^\perp_{x(t)}(y)\mu(t,\,\mathrm{d}y), \Phi^t_{v[\mu(t)]}(x') \right\rangle \mu_0(\,\mathrm{d}x').$$

The second term in the above identity can be bounded as

$$\sum_{i \neq q} \int_{\mathcal{S}_q(2\varepsilon)} \left\langle \int_{\mathcal{B}_i(2\varepsilon)} \frac{e^{\beta\langle x(t),y \rangle}}{\mathcal{Z}_{\beta,\mu(t)}(x(t))} \mathbf{P}^\perp_{x(t)}(y)\mu(t,\,\mathrm{d}y), \Phi^t_{v[\mu(t)]}(x') \right\rangle \mu_0(\,\mathrm{d}x')$$

$$\leqslant k e^{-(1-\alpha-8\varepsilon)\beta}.$$

We use the following bound for the other term:

$$\left\langle \mathbf{P}^\perp_{x(t)}(\Phi^t_{v[\mu(t)]}(y)), \Phi^t_{v[\mu(t)]}(x') \right\rangle$$

$$= \left\langle \Phi^t_{v[\mu(t)]}(y), \Phi^t_{v[\mu(t)]}(x') \right\rangle - \left\langle x(t), \Phi^t_{v[\mu(t)]}(y) \right\rangle \left\langle x(t), \Phi^t_{v[\mu(t)]}(x') \right\rangle$$

$$\geqslant \left\langle x(t), \Phi^t_{v[\mu(t)]}(x') \right\rangle \left(1 - \left\langle x(t), \Phi^t_{v[\mu(t)]}(y) \right\rangle \right)$$

$$\geqslant \frac{1}{2} \left\langle x(t), \Phi^t_{v[\mu(t)]}(x') \right\rangle \left\| x(t) - \Phi^t_{v[\mu(t)]}(y) \right\|^2.$$

40

We now integrate this inequality to get

$$
\int_{\mathcal{S}_q(2\varepsilon)} \left\langle \dot{x}(t), \Phi^t_{v[\mu(t)]}(x') \right\rangle \mu_0(\,\mathrm{d}x')
$$

$$
\geqslant \int_{\mathcal{S}_q(2\varepsilon)} \left\langle \int_{\mathcal{S}_q(2\varepsilon)} \frac{e^{\beta\langle x(t), \Phi^t_{v[\mu(t)]}(y)\rangle}}{\mathcal{Z}_{\beta,\mu(t)}(x(t))} \mathbf{P}^{\perp}_{x(t)}\left(\Phi^t_{v[\mu(t)]}(y)\right) \mu_0(\,\mathrm{d}y), \Phi^t_{v[\mu(t)]}(x') \right\rangle \mu_0(\,\mathrm{d}x')
$$

$$
\geqslant \frac{1}{k} \int_{\mathcal{S}_q(2\varepsilon)} \int_{\mathcal{S}_q(2\varepsilon)} \left\langle \mathbf{P}^{\perp}_{x(t)}(\Phi^t_{v[\mu(t)]}(y), \Phi^t_{v[\mu(t)]}(x') \right\rangle \mu_0(\,\mathrm{d}x')\mu_0(\,\mathrm{d}y)
$$

$$
\geqslant \frac{1}{k} \int_{\mathcal{S}_q(2\varepsilon)} \int_{\mathcal{S}_q(2\varepsilon)} \frac{1}{2} \left\langle x(t), \Phi^t_{v[\mu(t)]}(x') \right\rangle \left\| x(t) - \Phi^t_{v[\mu(t)]}(y) \right\|^2 \mu_0(\,\mathrm{d}x')\mu_0(\,\mathrm{d}y)
$$

$$
\geqslant \frac{(1-2\varepsilon)}{2k^2}\mathsf{V}_q(t).
$$

We can argue similarly for the second term resulting in:

$$
\int_{\mathcal{S}_q(2\varepsilon)} \left\langle x(t), \frac{\mathrm{d}}{\mathrm{d}t}\Phi^t_{v[\mu(t)]}(x') \right\rangle \mu_0(\,\mathrm{d}x') \geqslant -e^{-(1-\alpha-8\varepsilon)\beta}.
$$

All in all, for $t \geqslant T$,

$$
\dot{\mathsf{V}}_q(t) \leqslant -\frac{(1-2\varepsilon)}{k^2}\mathsf{V}_q(t) + 2e^{-(1-\alpha+8\varepsilon)\beta} \leqslant -\frac{(1-2\varepsilon)}{k^2}e^{-\lambda\beta} + 2e^{-(1-\alpha+8\varepsilon)\beta}.
$$

Because of the condition on λ, and the definition of T, we can conclude that

$$
\dot{\mathsf{V}}_q(T) < 0.
$$

By continuity, this implies that there exists $t < T$ such that $\mathsf{V}_q(T) \geqslant e^{-\lambda\beta}$. Therefore, necessarily $T \geqslant T_{\mathrm{esc}}$. □

Proof of Claim 3. For all $t \in [0, T_*(q,c)]$ we have

$$
\frac{2e^{\beta(1-\eta_q(t)-c)}}{\eta_q(t)} \leqslant \mathsf{V}_q(t) \leqslant \frac{4(1-\eta_q(t))}{k}. \tag{5.7}
$$

(The second inequality is actually always true.) Also,

$$
\dot{\eta}_q(t) \geqslant ke^{8\varepsilon\beta}\eta_q(t)\mathsf{V}_q(t)e^{-\beta(1-\eta_q(t))}. \tag{5.8}
$$

By plugging (5.7) into (5.8), we find

$$
\dot{\eta}_q(t) \geqslant 2ke^{-(c-8\varepsilon)\beta}. \tag{5.9}
$$

So, using both (5.7) and (5.9), we deduce

$$
\eta_q(t)\mathsf{V}_q(t)e^{-\beta(1-\eta_q(t))} \leqslant 4\eta_q(t)(1-\eta_q(t))
$$

$$
\leqslant 4\eta_q(0)\left(1 - \eta_q(0) - 2tke^{-(c-8\varepsilon)\beta}\right).
$$

We deduce

$$
T_*(q,c) < \frac{4\varepsilon}{k}e^{(c-8\varepsilon)\beta}. \tag{□}
$$

6 Beyond metastability

Theorem 1.2 entails that the dynamics take an exponential time to escape the metastable state. This raises the question of describing the dynamics beyond this escape time. It is for instance tempting to iterate the arguments using spherical caps presented in the proof of Theorem 1.2. We did not succeed in this endeavor as it appears challenging to propagate the (β, ε)-separateness condition beyond the first cone collapse. We leave this question open as a subject for future investigation.

In the same vein, here we are interested in understanding the dynamics in the low-temperature limit: d and n are fixed, and $\beta \to +\infty$. As alluded to in §1.2.4, existing results of this kind in related literature mostly rely on explicit time-rescalings strongly linked to the particular problem at hand that accelerate the dynamics. Finding an explicit rescaling is not straightforward in our setting due to particle interactions.

As a starting point for our study we posit the following question.

Problem 2 (Staircase profile). *Fix $d, n \geqslant 2$. Let $(x_1(0), \ldots, x_n(0)) \in (\mathbb{S}^{d-1})^n$ and consider the unique solution $(x_1(\cdot), \ldots, x_n(\cdot)) \in \mathscr{C}^0(\mathbb{R}_{\geqslant 0}, (\mathbb{S}^{d-1})^n)$ to the corresponding Cauchy problem for (SA) or (USA). Do there exist a number of jumps $k \in \{1, \ldots, n\}$, jumping times $0 = T_0 < T_1 < \ldots < T_k < T_{k+1} = +\infty$, and a sequence $(\tau_\beta)_{\beta \geqslant 0} \subset \mathscr{C}^0(\mathbb{R}_{\geqslant 0}; \mathbb{R}_{\geqslant 0})$, such that the function $\varphi_\beta : \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ defined by*

$$\varphi_\beta(t) := \mathsf{E}_\beta\left(x_1(\tau_\beta(t)), \ldots, x_n(\tau_\beta(t))\right),$$

converges uniformly on (T_i, T_{i+1}) for $i \in \{0, \ldots, k-1\}$ towards some piecewise constant $\varphi_\infty \in L^\infty(\mathbb{R}_{\geqslant 0}; [0, 1])$ as $\beta \to +\infty$? Otherwise said, φ_∞ is defined as

$$\varphi_\infty(t) = \varphi_\infty(T_i) \qquad \text{for } t \in [T_i, T_{i+1}),$$

for $i \in \{0, \ldots, k+1\}$.

We believe this to be a challenging problem due to the singular nature of the limit $\beta \to +\infty$. At a fixed time instance, the Laplace method ensures that, as $\beta \to +\infty$, the softmax converges to the argmax. But issues arise along the flow due to the fact that particles cannot collide in finite time—when two particles are too near, most of the interaction is not between the two, but rather with the others.

6.1 Staircase on the circle

We present a stylized example in which the staircase profile of the energy can be proven to occur. We focus on dynamics on the circle with the following class of initial configurations.
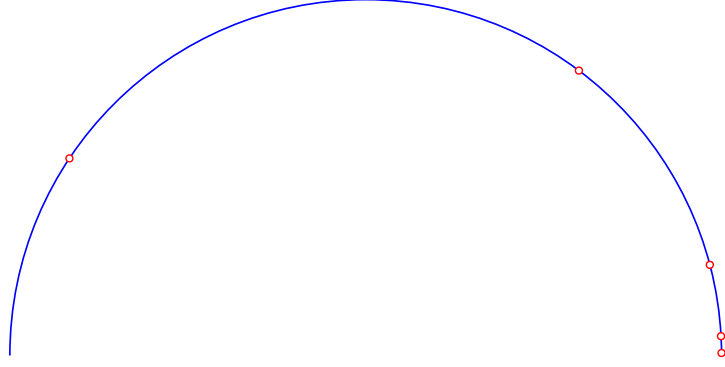
Definition 6.1 (Well-prepared configuration). *Let $\beta > 1$. We call a configuration $(\theta_1, \ldots, \theta_n) \in \mathbb{T}^n$ well-prepared if*

$$0 \leqslant \theta_1 < \theta_2 < \ldots < \theta_n \leqslant \pi$$

and there exists a numerical constant $c > 1$ such that for all $i \in \{2, \ldots, n-1\}$ and $k > i$,

$$\cos(\theta_i - \theta_1) > \cos(\theta_k - \theta_i) + \frac{c \log \beta}{\beta}.$$

Remark 6.2. *The configuration $(\theta_1, \ldots, \theta_n)$ where $\theta_j = c \cdot 2^j$ is well-prepared for sufficiently small $c > 0$, and β large enough.*

Figure 6: A well-prepared configuration.

We can give an affirmative answer to Problem 2 but for a slightly modified version of (USA) in which we enforce collisions.

Definition 6.3 (Modified (USA)). *Suppose $\beta > 0$, $n \geqslant 2$, and $(\theta_i(0))_{i=1}^n \in \mathbb{T}^n$. Given the unique solution $(\theta_i(\cdot))_{i=1}^n \in \mathscr{C}^0(\mathbb{R}_{\geqslant 0}; \mathbb{T}^n)$ to the corresponding Cauchy problem for (3.7), define*

$$T_* = \inf \left\{ t \geqslant 0 \colon \exists i \neq j \in \{1, \ldots, n\}^2 \text{ such that } |\theta_i(t) - \theta_j(t)| \leqslant \frac{1}{\sqrt{\beta \log \beta}} \right\},$$

and suppose that

$$\mathrm{card} \left\{ i \in \{1, \ldots, n\} \colon \exists j \text{ such that } |\theta_i(T_*) - \theta_j(T_*)| \leqslant \frac{1}{\sqrt{\beta \log \beta}} \right\} = 2.$$

Without loss of generality let $(1, 2)$ be these two indices.

1. If $T_* < +\infty$, define

$$
\left(\overline{\theta}_i(t)\right)_{i=1}^n := \begin{cases} (\theta_i(t))_{i=1}^n & \text{for } t < T_* \\ (\theta_i^*(t))_{i=1}^n & \text{for } t \geqslant T_*, \end{cases}
$$

where $\theta_1^*(t) := \theta_2^*(t)$ for $t \geqslant T_*$, and $(\theta_i^*(\cdot))_{i=2}^n \in \mathscr{C}^0([T_*, +\infty); \mathbb{T}^n)$ denotes the unique solution to (3.7) with initial data $(\theta_i(T_*))_{i=2}^n$;

2. If $T = +\infty$, set $(\overline{\theta}_i(t))_{i=1}^n := (\theta_i(t))_{i=1}^n$ for all $t \geqslant 0$.

We call $(\overline{\theta}_i(\cdot))_{i=1}^n$ the modified USA dynamics.

These dynamics have some enhanced aspects compared to (3.7); for instance, it is not obvious to prove that two particles remain in a neighborhood of size $\beta^{-1/2}$ of each other over time. The only statement we can prove in this direction is Appendix A.1.2, which requires that these two particles stay isolated enough from all the others. Furthermore, our numerical simulation in Figure 4 is actually done by merging these nearby particles since we otherwise encounter numerical overflow. The following holds.

Theorem 6.4. *Let $n \geqslant 2$. For $\beta > 0$, let $(\theta_i(0))_{i=1}^n \in \mathbb{T}^n$ be a well-prepared configuration in the sense of Definition 6.1, let $\Theta(\cdot) = (\theta_i(\cdot))_{i=1}^n$ be the dynamics defined in Definition 6.3, and consider $\tau_\beta \in \mathscr{C}^0(\mathbb{R}_{\geqslant 0}; \mathbb{R}_{\geqslant 0})$ defined as a solution to*

$$
\begin{cases} \dot{\tau}_\beta(t) = \log \beta \displaystyle\max_{\substack{(i,j) \in \{1,\dots,n\}^2 \\ |\theta_i(t) - \theta_j(t)| > \frac{1}{\sqrt{\beta \log \beta}}}} e^{\beta(1 - \cos(\theta_i(t) - \theta_j(t)))} & \text{for } t \geqslant 0, \\ \tau_\beta(0) = 0. \end{cases} \tag{6.1}
$$

Then there exist a sequence of times $0 = T_0 < T_1 < T_2 < \ldots < T_k < T_{k+1} = +\infty$ with $k \leqslant n$, and a piecewise constant $\varphi_\infty \in L^\infty(\mathbb{R}_{\geqslant 0}; [0,1])$ such that

$$
\lim_{\beta \to +\infty} \max_{i \in \{1,\dots,k\}} \sup_{t \in (T_i, T_{i+1})} |\mathsf{E}_\beta(\Theta(\tau_\beta(t))) - \varphi_\infty(t)| = 0.
$$

Proof of Theorem 6.4. For $i \in \{1, \dots, n\}$ and $t \geqslant 0$, set $\widetilde{\theta}_i(t) := \theta_i(\tau_\beta(t))$.

Step 1. Preliminary spherical caps

Consider the spherical caps

$$
\begin{aligned}
\mathcal{S}_1 &:= [\theta_1(0), \theta_2(0)], \\
\mathcal{S}_q &:= [\theta_q(0) - e^{-\beta K}, \theta_q(0) + e^{-\beta K}] \qquad \text{for } q \in \{3, \dots, n\},
\end{aligned}
$$

where $\frac{\log \beta}{\beta} < K < \cos\left(\frac{\theta_2(0) - \theta_1(0)}{2}\right) - \cos(\theta_3(0) - \theta_2(0))$. We also define

$$
t_1(\beta) := \inf\left\{ t \geqslant 0 : \min_{(i,j) \in \{1,\dots,n\}^2} |\theta_i(t) - \theta_j(t)| \leqslant \sqrt{\frac{\log \beta}{\beta}} \right\},
$$

44

and

$$T_1(\beta) := \inf \left\{ t \geqslant 0 \colon \min_{(i,j) \in \{1,\dots,n\}^2} |\theta_i(t) - \theta_j(t)| \leqslant \frac{1}{\sqrt{\beta \log \beta}} \right\}.$$

We can slightly modify the proof of Theorem 1.2 to ensure that the particles $\theta_i(t)$ do not leave their respective spherical caps $\mathcal{S}_q$ up to a time $T > 0$ which is exponentially large with respect to β, and that $|\theta_1(t) - \theta_2(t)|$ becomes exponentially small with respect to β. We briefly explain how to adapt the proof. The first step of the proof of Theorem 1.2 can be reproduced in this setting with a lower bound on the time of escape which is of the form

$$T_{\mathrm{esc}} \geqslant \max_{q \in \{1,\dots,n\}} \frac{e^{\beta(1-\alpha_q - 4e^{-\beta K} - K)}}{n},$$

where $\alpha_q = \cos(\theta_q(0) - \theta_{q-1}(0))$. We can reproduce the cone collapse argument and ensure that the time $T_c > 0$ of clustering within the first spherical cap satisfies

$$T_c \leqslant 4n e^{\beta \varepsilon},$$

with $\varepsilon = 1 - \cos\left(\frac{\theta_2(0) - \theta_1(0)}{2}\right)$. So, asymptotically, the time scales are of different orders if for all $q \in \{2,\dots,n\}$,

$$\frac{\log \beta}{\beta} < K < \cos\left(\frac{\theta_2(0) - \theta_1(0)}{2}\right) - \alpha_q.$$

We end up with two particles for the (USA) dynamics that come exponentially near each other while the others do not escape their original spherical caps.

As a result of the above discussion, all particles remain in their original caps up to time T, and for $t \in [0, T_1(\beta)]$,

$$\underset{\substack{(i,j) \in \{1,\dots,n\}^2 \\ |\theta_i(t) - \theta_j(t)| > \frac{1}{\sqrt{\beta \log \beta}}}}{\arg\max} \cos\left(\theta_i(t) - \theta_j(t)\right) = (1,2)$$

if β is large enough. Then for all $i \in \{1,\dots,n\}$ and $t \in [0, T_1(\beta)]$,

$$\dot{\widetilde{\theta}}_i(t) = \dot{\tau}_\beta(t) \sum_{j=1}^n \frac{e^{\beta\left(\cos\left(\widetilde{\theta}_j(t) - \widetilde{\theta}_i(t)\right) - 1\right)}}{n^2} \sin\left(\widetilde{\theta}_j(t) - \widetilde{\theta}_i(t)\right). \tag{6.2}$$

Plugging (6.1) into (6.2), we gather that for all $i \in \{1,\dots,n\}$ and $t \in [0, T_1(\beta)]$,

$$\dot{\widetilde{\theta}}_i(t) = \log \beta \sum_{j=1}^n \frac{e^{\beta(\cos(\widetilde{\theta}_j(t) - \widetilde{\theta}_i(t)) - \cos(\widetilde{\theta}_1(t) - \widetilde{\theta}_2(t)))}}{n^2} \sin(\widetilde{\theta}_j(t) - \widetilde{\theta}_i(t)).$$

Set $u_1(t) := \widetilde{\theta}_2(t) - \widetilde{\theta}_1(t)$. We then have

$$\dot{u}_1(t) = -\frac{2\log\beta}{n^2}\sin(u_1(t)) + \sum_{j\notin\{1,2\}}\frac{e^{\beta\left(\cos\left(\tilde{\theta}_j(t)-\tilde{\theta}_1(t)\right)-d_1(t)\right)}}{n^2}\sin(\tilde{\theta}_j(t) - \tilde{\theta}_1(t))$$

$$-\sum_{j\notin\{1,2\}}\frac{e^{\beta\left(\cos\left(\tilde{\theta}_j(t)-\tilde{\theta}_2(t)\right)-d_1(t)\right)}}{n^2}\sin(\tilde{\theta}_j(t) - \tilde{\theta}_2(t)).$$

We then can bound the right-hand side using the inequality on d_i as

$$\left|\dot{u}_1(t) + \frac{2\log\beta}{n^2}\sin(u_1(t))\right| \leqslant \frac{2\log\beta}{n}e^{-c\beta}.$$

Then we have the following lemma.

Lemma 6.5. *Suppose $u_0 \in [0,1]$, $\beta \geqslant e$, $c > 0$, $K > 0$ and $\kappa > \frac{\log\beta}{\beta}$. Consider $u \in \mathscr{C}^0(\mathbb{R}_{\geqslant 0})$ a solution to the Cauchy problem*

$$\begin{cases} \dot{u}(t) = -c\log\beta\sin(u(t)) + c(\beta) & \text{for } t \geqslant 0 \\ u(0) = u_0, \end{cases}$$

where

$$|c(\beta)| \leqslant Ke^{-\kappa\beta}\log\beta.$$

Let

$$t(\beta) := \inf\left\{t \geqslant 0 : u(t) \leqslant \sqrt{\frac{\log\beta}{\beta}}\right\},$$

and

$$T(\beta) := \inf\left\{t \geqslant 0 : u(t) \leqslant \frac{1}{\sqrt{\beta\log\beta}}\right\}.$$

Then, as $\beta \to +\infty$,

$$\left|t(\beta) - \frac{2}{c}\right| \leqslant \frac{2\log\left(\tan\left(\frac{u(0)}{2}\right)\right)}{c\log\beta} + \frac{\log\log\beta}{c\log\beta},$$

and

$$T(\beta) - t(\beta) \leqslant \frac{2\log\log\beta}{c\log\beta} + O\left(\frac{1}{\beta^2\log\beta}\right)$$

We postpone the proof to Appendix A.1.3.

Step 2. Repeating the arguments

We now argue by induction. By definition of the modified USA dynamics (6.3), at time $t = T_k(\beta)$ (defined analogously to $T_1(\beta)$), the particles $\theta_1(t)$ and $\theta_{k+1}(t)$ are fusioned. Thus at time $T_k(\beta)$ we consider the spherical caps

$$\begin{aligned} S_1 &:= [\theta_1(T_k), \theta_{k+2}(T_k)], \\ S_q &:= [\theta_q(T_k) - e^{-c_k\beta}, \theta_q(T_k) + e^{-c_k\beta}] && \text{for } q \geqslant k+3, \end{aligned}$$

46

where
$$c_k = \cos\left(\frac{\theta_{k+2}(T_k) - \theta_1(T_k)}{2}\right) - \cos\left(\theta_{k+3}(T_k) - \theta_{k+2}(T_k)\right) > 0$$

because of the hypothesis on the initial configuration, and $\theta_1(T_k(\beta)) \geqslant \theta_1(0)$. One can redo the argument of Step 1, to establish that the particles do not escape their spherical caps because of the hypothesis on the distance of spherical caps at initialization[7]. Moreover, similar bounds on $T_k(\beta)$ and $t_k(\beta)$ can be provided by virtue of a result similar to Lemma 6.5, but considering weighted particles as to handle the particles which are already merged[8]. So by induction, there exist $t_1(\beta) < T_1(\beta) < t_2(\beta) < T_2(\beta) < \ldots < t_k(\beta) < T_k(\beta)$ such that for all $i \in \{1, \ldots, k\}$

$$T_i(\beta) - t_i(\beta) \leqslant \frac{2}{c_i}\frac{\log\log\beta}{\log\beta} + O\left(\frac{1}{\beta^2\log\beta}\right).$$

Besides, we notice that for all $i \in \{1, \ldots, k\}$

$$\mathsf{E}_\beta\left(\tilde{\theta}_1(T_i(\beta)), \ldots, \tilde{\theta}_n(T_i(\beta))\right) - \mathsf{E}_\beta\left(\tilde{\theta}_1(t_i(\beta)), \ldots, \tilde{\theta}_n(t_i(\beta))\right)$$
$$= \sum_{k=1}^{n}\sum_{j=1}^{n}\frac{e^{\beta(\cos(\theta_j(T_i(\beta)) - \theta_k(T_i(\beta))) - 1)}}{n^2} - \sum_{k=1}^{n}\sum_{j=1}^{n}\frac{e^{\beta(\cos(\theta_j(t_i(\beta)) - \theta_k(t_i(\beta))) - 1)}}{n^2}$$
$$\geqslant \frac{2i}{n^2}e^{\beta(\cos(\theta_{i+1}(T_i(\beta)) - \theta_1(T_i(\beta))) - 1)} - e^{\beta(\cos(\theta_{i+1}(T_i(\beta)) - \theta_{i+2}(T_i(\beta))) - 1)}$$
$$- e^{\beta(\cos(\theta_1(t_i(\beta)) - \theta_{i+1}(t_i(\beta))) - 1)}$$
$$\geqslant \frac{2i}{n^2} + O\left(\frac{1}{\log\beta}\right),$$

and we also have

$$\mathsf{E}_\beta\left(\tilde{\theta}_1(T_i(\beta)), \ldots, \tilde{\theta}_n(T_i(\beta))\right) - \mathsf{E}_\beta\left(\tilde{\theta}_1(t_i(\beta)), \ldots, \tilde{\theta}_n(t_i(\beta))\right)$$
$$= \sum_{k=1}^{n}\sum_{j=1}^{n}\frac{e^{\beta(\cos(\theta_j(T_i(\beta)) - \theta_k(T_i(\beta))) - 1)}}{n^2} - \sum_{k=1}^{n}\sum_{j=1}^{n}\frac{e^{\beta(\cos(\theta_j(t_i(\beta)) - \theta_k(t_i(\beta))) - 1)}}{n^2}$$
$$\leqslant \frac{2i}{n^2} + \left(\frac{1}{\log\beta}\right).$$

Using the monotonicity of the energy, and the definitions of $T_i(\beta)$ and $t_{i+1}(\beta)$, we gather that for all $i \in \{1, \ldots, k\}$ and $t \in (T_i(\beta), t_{i+1}(\beta))$,

$$0 \leqslant \mathsf{E}_\beta\left(\tilde{\theta}_1(t), \ldots, \tilde{\theta}_n(t)\right) - \mathsf{E}_\beta\left(\tilde{\theta}_1(T_i(\beta)), \ldots, \tilde{\theta}_n(T_i(\beta))\right) = O\left(\frac{1}{\log\beta}\right).$$

---

[7]The hypothesis on the initial configuration implies that all the spherical caps are sufficiently separated to apply the adapted proof of Theorem 1.2, by using the fact that $\theta_1(T_k(\beta)) \geqslant \theta_1(0)$ for all $k \in \{1, \ldots, n-1\}$.

[8]The proof is a straightforward adaptation, the only difference being that the scalar differential equation is not for $u(t) = \tilde{\theta}_k(t) - \tilde{\theta}_1(t)$, but rather for a weighted difference of the two particles, namely $v(t) = \lambda_k\tilde{\theta}_k(t) - (1 - \lambda_k)\tilde{\theta}_1(t)$.

Thence there exist $\ell_0, \ell_1, \ldots, \ell_n \in \mathbb{R}_{\geqslant 0}$, defined as limits of $T_k(\beta)$ as $\beta \to +\infty$, with $\ell_0 := 0$, such that, defining $\varphi_\infty : \mathbb{R}_{\geqslant 0} \to \mathbb{R}_{\geqslant 0}$ as

$$\varphi_\infty(t) = \frac{1}{n} + \frac{2}{n^2} \sum_{k=1}^{i} k \qquad \text{for } t \in (\ell_i, \ell_{i+1}),$$

we have, for all $i \in \{1, \ldots, n\}$ and $t \in (\ell_i, \ell_{i+1})$,

$$\left| \mathsf{E}_\beta \left(\tilde{\theta}_1(t), \ldots, \tilde{\theta}_n(t) \right) - \varphi_\infty(t) \right| \xrightarrow[\beta \to +\infty]{} 0,$$

as desired. $\qquad\qquad\square$

6.2 A reparametrization candidate

A naive way of accelerating the dynamics is to introduce the time reparametrization τ_β defined by

$$\begin{cases} \dot{\tau}_\beta(t) = \frac{\log \beta}{\|\nabla \mathsf{E}_\beta(\tau_\beta(t))\|} & \text{for } t \geqslant 0, \\ \tau_\beta(0) = 0. \end{cases}$$

One sees that when the gradient is small, the dynamics is accelerated. Therefore, the hope is that the dynamics would take a constant time (not depending on β asymptotically) to induce a jump in the energy. Denoting $\varphi_\beta(t) := \mathsf{E}_\beta(u(\tau_\beta(t)))$, we have

$$\dot{\varphi}_\beta(t) = \log \beta \cdot \|\nabla \mathsf{E}_\beta(u(\tau_\beta(t)))\|,$$

since the gradient has different scales of magnitude depending on β. We posit the following question.

Problem 3. *Does the statement of Problem 2 hold for τ_β as above?*

A Toolkit

A.1 Technical lemmas

A.1.1 Proof of Lemma 2.1

Proof of Lemma 2.1. First of all, observe that $F(u) = u(1-u)e^{\beta(u-1)}$ is zero at $u = 0, 1$ and strictly positive in $(0, 1)$. Whence $t \mapsto u(t)$ is increasing for all $\beta \geqslant 0$. Now suppose $\beta > 1$, and consider

$$t_1 := \inf \left\{ t \geqslant 0 \colon 1 - u(t) \leqslant \frac{1}{\beta} \right\}.$$

Since $u(t) \geqslant u_0 > 0$ and also $1 - u(t) \geqslant \beta^{-1}$ for all $t \in [0, t_1]$, we have

$$\dot{u}(t) \geqslant \frac{u_0}{\beta} e^{\beta(u(t)-1)}$$

for $t \in [0, t_1]$. Setting $f(t) := e^{\beta(1-u(t))}$, we find

$$\dot{f}(t) = -\beta e^{\beta(1-u(t))}\dot{u}(t) \leqslant -u_0,$$

whence

$$f(t) \leqslant f(0) - u_0 t.$$

It follows that

$$t_1 \leqslant \frac{f(0) - f(t_1)}{u_0} \leqslant \frac{f(0)}{u_0} = \frac{e^{\beta(1-u_0)}}{u_0}.$$

We then define

$$t_2 := \inf\left\{t \geqslant t_1 : 1 - u(t) \leqslant e^{-c\beta}\right\}.$$

For all $t \geqslant t_2 \geqslant t_1$ we have $u(t) \geqslant 1 - \beta^{-1}$ and $e^{\beta(u(t)-1)} \geqslant e^{-1}$. So

$$\dot{u}(t) \geqslant (1 - u(t))\left(1 - \frac{1}{\beta}\right)\frac{1}{e} = \frac{1 - u(t)}{\frac{\beta}{\beta-1} \cdot e}.$$

By the Grönwall lemma, for all $t \geqslant t_1$

$$1 - u(t) \leqslant (1 - u(t_1))e^{-\frac{(t-t_1)}{\frac{\beta}{\beta-1} \cdot e}}.$$

We then deduce that

$$t_2 - t_1 \leqslant \frac{\beta^2 \cdot c \cdot e}{\beta - 1}.$$

We conclude by using the bound on t_1. $\qquad\square$

A.1.2 Proof of Lemma 2.2

Proof of Lemma 2.2. We define

$$T_* := \inf\left\{t \geqslant 0 : \min_{(i,j)\in I^2}\langle x_i(t), x_j(t)\rangle \leqslant 1 - \delta\right\}.$$

By contradiction suppose that $T_* < T$. Let $t \in [0, T_*]$ and

$$\rho(t) := \min_{(i,j)\in I^2}\langle x_i(t), x_j(t)\rangle$$

We also consider $i(t), j(t)$ such that

$$(i(t), j(t)) \in \arg\min_{(i,j)\in I^2}\langle x_i(t), x_j(t)\rangle.$$

Following exactly the same arguments as in Step 2 of the proof of Theorem 1.2, we get

$$\dot{\rho}(t) \geqslant \frac{2}{n}\rho(t)(1 - \rho(t))e^{-\beta(1-\rho(t))} - 2ne^{-(1-\alpha)\beta}$$

49

for $t \in [0, T_*]$. Now for $t = T_*$, by continuity we have

$$\frac{1}{n}\rho(T_*)(1 - \rho(T_*))e^{-\beta(1-\rho(T_*))} > ne^{-(1-\alpha)\beta}.$$

Plugging the former inequality into the latter we get $\dot\rho(T_*) > 0$. So for all times t in a neighborhood of T_*, $\dot\rho(t) > 0$, whence $\rho(t) < 1 - \delta$ for t in a neighborhood of T_*. This is in contradiction with the definition of T_*. Therefore $T_* \geqslant T$, as desired. $\qquad\square$

A.1.3   Proof of Lemma 6.5

Proof of Lemma 6.5. Define $v(t) := \log\left(\tan\left(\frac{u(t)}{2}\right)\right)$. Note that

$$\dot v(t) = \frac{\dot u(t)}{2\sin(u(t))}.$$

For all $t \geqslant 0$, we then have

$$\dot v(t) = -c\log\beta + \frac{c(\beta)}{2\sin(u(t))}.$$

Furthermore, for all $t \in [0, T(\beta)]$,

$$\tan\left(\frac{u(t)}{2}\right) \leqslant \tan\left(\frac{u(0)}{2}\right) e^{-\left(\frac{c\log\beta}{2} - \frac{c(\beta)\sqrt{\beta\log\beta}}{2}\right)t}, \tag{A.1}$$

as well as

$$\tan\left(\frac{u(t)}{2}\right) \geqslant \tan\left(\frac{u(0)}{2}\right) e^{-\left(\frac{c\log\beta}{2} + \frac{c(\beta)\sqrt{\beta\log\beta}}{2}\right)t}. \tag{A.2}$$

We now turn our attention to deriving bounds on $t(\beta)$ and $T(\beta)$. The inequalities (A.2) and (A.1) yield

$$e^{-\left(\frac{c(\beta)\sqrt{\beta\log\beta}}{2}\right)t(\beta)} \leqslant \frac{\tan\left(\frac{u(t(\beta))}{2}\right)}{e^{\frac{-ct(\beta)\log\beta}{2}}\tan\left(\frac{u(0)}{2}\right)} \leqslant e^{\left(\frac{c(\beta)\sqrt{\beta\log\beta}}{2}\right)t(\beta)}$$

Since $c(\beta) = O(e^{-\kappa\beta}\log\beta)$ with $\kappa > 0$ and $t(\beta)$ is bounded uniformly in β, as $\beta \to +\infty$ we have

$$\frac{\tan\left(\frac{u(t(\beta))}{2}\right)}{e^{\frac{-c(\log\beta)t(\beta)}{2}}\tan\left(\frac{u(0)}{2}\right)} = 1 + O\left(c(\beta)\sqrt{\beta\log\beta}\right).$$

Besides, Taylor-expanding the $\tan$ we get

$$e^{\frac{-c(\log\beta)t(\beta)}{2}}\tan\left(\frac{u(0)}{2}\right) = \sqrt{\frac{\log\beta}{\beta}} + O\left(\left(\frac{\log\beta}{\beta}\right)^{-\frac{3}{2}}\right)$$

$$+ O\left(e^{\frac{-c(\log\beta)t(\beta)}{2}}c(\beta)\sqrt{\beta\log\beta}\right).$$

Whence, as $\beta \to +\infty$,

$$t(\beta) = \frac{2}{c} + \frac{2\log\left(\tan\left(\frac{u(0)}{2}\right)\right)}{c\log\beta} - \frac{\log\log\beta}{c\log\beta} + O\left(\frac{1}{\beta^2\log\beta}\right).$$

Following the above chain of computations, we can also gather that, as $\beta \to +\infty$,

$$T(\beta) = \frac{2}{c} + \frac{2\log\left(\tan\left(\frac{u(0)}{2}\right)\right)}{c\log\beta} + \frac{\log\log\beta}{c\log\beta} + O\left(\frac{1}{\beta^2\log\beta}\right).$$

Therefore, asymptotically as $\beta \to +\infty$,

$$T(\beta) - \tau(\beta) = \frac{2\log\log\beta}{c\log\beta} + O\left(\frac{1}{\beta^2\log\beta}\right). \qquad \square$$

A.2 Numerical considerations

Code can be found at https://github.com/HugoKoubbi/2024-transformers-dotm. In Figure 4 we used the initial configuration displayed in Figure 7, and discretized the equation using a forward Euler scheme with time-step equal to 10^{-6}.
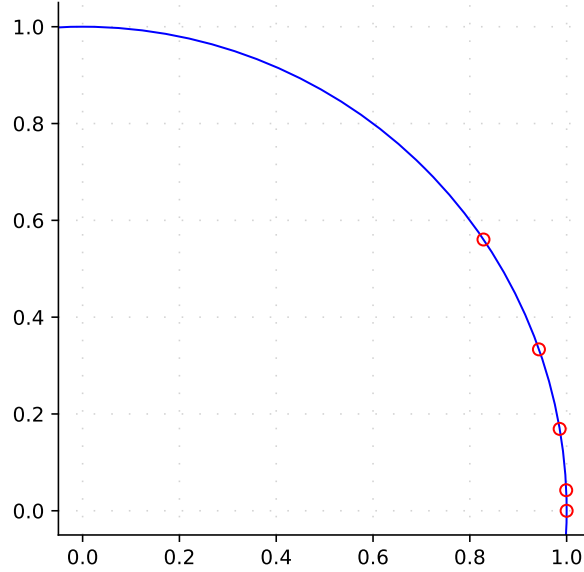

Figure 7: The initial configuration with $n = 5$ points used for Figure 4.

References

[ABBA$^+$24] Emmanuel Abbe, Samy Bengio, Enric Boix-Adsera, Etai Littwin, and Joshua Susskind. Transformers learn through gradual rank increase. *Advances in Neural Information Processing Systems*, 36, 2024.

[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.

[AFZ24] Albert Alcalde, Giovanni Fantuzzi, and Enrique Zuazua. Clustering in pure-attention hardmax transformers and its role in sentiment analysis. *arXiv preprint arXiv:2407.01602*, 2024.

[AHMP24] Medha Agarwal, Zaid Harchaoui, Garrett Mulcahy, and Soumik Pal. Iterated Schrödinger bridge approximation to Wasserstein Gradient Flows. *arXiv preprint arXiv:2406.10823*, 2024.

[BD19] Dmitriy Bilyk and Feng Dai. Geodesic distance Riesz energy on the sphere. *Transactions of the American Mathematical Society*, 372(5):3141–3166, 2019.

[BDH16] Anton Bovier and Frank Den Hollander. *Metastability: a potential-theoretic approach*, volume 351. Springer, 2016.

[BDL07] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[BDM18] Dmitriy Bilyk, Feng Dai, and Ryan Matzke. The Stolarsky principle and energy optimization on the sphere. *Constructive Approximation*, 48(1):31–60, 2018.

[BE85] Dominique Bakry and Michel Émery. Diffusions hypercontractives. *Séminaire de probabilités de Strasbourg*, 19:177–206, 1985.

[Ber23] Raphaël Berthier. Incremental learning in diagonal linear networks. *Journal of Machine Learning Research*, 24(171):1–26, 2023.

[BHK24] Han Bao, Ryuichiro Hataya, and Ryo Karakida. Self-attention networks localize when qk-eigenspectrum concentrates. *arXiv preprint arXiv:2402.02098*, 2024.

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford Press University, 2013.

[Bou23] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

[BPVF22] Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *Advances in Neural Information Processing Systems*, 35:20105–20118, 2022.

[CAP24]   Valérie Castin, Pierre Ablin, and Gabriel Peyré. How smooth is attention? In *Forty-first International Conference on Machine Learning*, 2024.

[Car96]   John Cardy. *Scaling and renormalization in statistical physics*, volume 5. Cambridge university press, 1996.

[CdCI18]   Henry Cohn and Matthew de Courcy-Ireland. The Gaussian core model in high dimensions. *Duke Mathematical Journal*, 167(13):2417 − 2455, 2018.

[CK07]   Henry Cohn and Abhinav Kumar. Universally optimal distribution of points on spheres. *Journal of the American Mathematical Society*, 20(1):99−148, 2007.

[CNQG24]   Aditya Cowsik, Tamra Nebabu, Xiao-Liang Qi, and Surya Ganguli. Geometric dynamics of signal propagation predict trainability of transformers. *arXiv preprint arXiv:2403.02579*, 2024.

[CP89]   Jack Carr and Robert L Pego. Metastable patterns in solutions of $u_t = \epsilon^2 u_{xx} - f(u)$. *Communications on Pure and Applied Mathematics*, 42(5):523−576, 1989.

[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.

[CZC+22]   Tianlong Chen, Zhenyu Zhang, Yu Cheng, Ahmed Awadallah, and Zhangyang Wang. The principle of diversity: Training stronger vision transformers calls for reducing all levels of redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12020−12030, 2022.

[DBK24]   Gbètondji JS Dovonon, Michael M Bronstein, and Matt J Kusner. Setting the record straight on transformer oversmoothing. *arXiv preprint arXiv:2401.04301*, 2024.

[DCL21]   Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning*, pages 2793−2803. PMLR, 2021.

[FH89]   Giorgio Fusco and Jack K Hale. Slow-motion manifolds, dormant instability, and singular perturbations. *Journal of Dynamics and Differential Equations*, 1:75−94, 1989.

[FW98]   MI Freidlin and AD Wentzell. *Random perturbations*. Springer, 1998.

[FZH+22]   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.

[GBLJ19]   Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[GLPR23]   Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *arXiv preprint arXiv:2312.10794*, 2023.

[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.

[GSSD19]   Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. *arXiv preprint arXiv:1909.12051*, 2019.

[GWDW23] Xiaojun Guo, Yifei Wang, Tianqi Du, and Yisen Wang. Contranorm: A contrastive learning perspective on oversmoothing and beyond. In *The Eleventh International Conference on Learning Representations*, 2023.

[HMZ+23]   Bobby He, James Martens, Guodong Zhang, Aleksandar Botev, Andrew Brock, Samuel L Smith, and Yee Whye Teh. Deep transformers without shortcuts: Modifying self-attention for faithful signal propagation. In *The Eleventh International Conference on Learning Representations*, 2023.

[JCD23]   Liwei Jiang, Yudong Chen, and Lijun Ding. Algorithmic regularization in model-free overparametrized asymmetric matrix factorization. *SIAM Journal on Mathematics of Data Science*, 5(3):723–744, 2023.

[JDB23]   Amir Joudaki, Hadi Daneshmand, and Francis Bach. On the impact of activation and normalization in obtaining isometric embeddings at initialization. *Advances in Neural Information Processing Systems*, 36:39855–39875, 2023.

[JGŞ+21]   Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

[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.

[KNS16]   Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I*, pages 795–811. Springer, 2016.

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

[LLH$^+$19]   Yiping Lu, Zhuohan Li, Di He, Zhiqing Sun, Bin Dong, Tao Qin, Liwei Wang, and Tie-Yan Liu. Understanding and improving transformer from a multi-particle dynamic system point of view. *arXiv preprint arXiv:1906.02762*, 2019.

[LXB19]   Shuyang Ling, Ruitu Xu, and Afonso S Bandeira. On the landscape of synchronization networks: A perspective from nonconvex optimization. *SIAM Journal on Optimization*, 29(3):1879–1907, 2019.

[MB24]   Andrew D McRae and Nicolas Boumal. Benign landscapes of low-dimensional relaxations for orthogonal synchronization on general graphs. *SIAM Journal on Optimization*, 34(2):1427–1454, 2024.

[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.

[NAB$^+$22]   Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022.

[NLL$^+$24]   Lorenzo Noci, Chuning Li, Mufan Li, Bobby He, Thomas Hofmann, Chris J Maddison, and Dan Roy. The shaped transformer: Attention models in the infinite depth-and-width limit. *Advances in Neural Information Processing Systems*, 36, 2024.

[OR07]   Felix Otto and Maria G Reznikoff. Slow motion of gradient flows. *Journal of Differential Equations*, 237(2):372–420, 2007.

[OV00]   Felix Otto and Cédric Villani. Generalization of an Inequality by Ta-
          lagrand and Links with the Logarithmic Sobolev Inequality. *Journal
          of Functional Analysis*, 173:361–400, 2000.

[Peg07]   Robert L Pego. Lectures on dynamics in models of coarsening and
          coagulation. *Dynamics in models of coarsening, coagulation, conden-
          sation and quantization*, 9:1–61, 2007.

[Pes24]   Scott William Pesme. Deep learning theory through the lens of di-
          agonal linear networks. Technical report, EPFL, 2024.

[PF23]   Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in
          diagonal linear networks. *Advances in Neural Information Processing
          Systems*, 36:7475–7505, 2023.

[PS20]   Mircea Petrache and Sylvia Serfaty. Crystallization for Coulomb and
          Riesz interactions as a consequence of the Cohn-Kumar conjecture.
          Proceedings of the American Mathematical Society, 148(7):3047–3057,
          2020.

[RF22]   Valentina Ros and Yan V Fyodorov. The high-d landscapes paradigm:
          spin-glasses, and beyond. *arXiv preprint arXiv:2209.07975*, 2022.

[RMC21]   Noam Razin, Asaf Maman, and Nadav Cohen. Implicit regulariza-
          tion in tensor factorization. In *International Conference on Machine
          Learning*, pages 8913–8924. PMLR, 2021.

[RZZD23] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token con-
          trast for weakly-supervised semantic segmentation. In *Proceedings
          of the IEEE/CVF Conference on Computer Vision and Pattern Recogni-
          tion*, pages 3093–3102, 2023.

[SABP22] Michael E Sander, Pierre Ablin, Mathieu Blondel, and Gabriel Peyré.
          Sinkformers: Transformers with doubly stochastic attention. In *In-
          ternational Conference on Artificial Intelligence and Statistics*, pages
          3515–3530. PMLR, 2022.

[SK97]   E.B Saffanda and BJ Kuijlaarn. Distributing many points on a sphere.
          The mathematical intelligencer, 19:5–11, 1997.

[Sti76]   Frank H Stillinger. Phase transitions in the Gaussian core system.
          The Journal of Chemical Physics, 65(10):3968–3974, 1976.

[SWJS24] Michael Scholkemper, Xinyi Wu, Ali Jadbabaie, and Michael Schaub.
          Residual connections and normalization can provably prevent over-
          smoothing in gnns. *arXiv preprint arXiv:2406.02997*, 2024.

[Tad23]   Eitan Tadmor. Swarming: hydrodynamic alignment with pressure.
          Bulletin of the American Mathematical Society, page 285–325, 2023.

[Vil21]   Cédric Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[VSP+17]   Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[WAW+24]   Xinyi Wu, Amir Ajorlou, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the role of attention masks and layernorm in transformers. *arXiv preprint arXiv:2405.18781*, 2024.

[WAWJ24]   Xinyi Wu, Amir Ajorlou, Zihui Wu, and Ali Jadbabaie. Demystifying oversmoothing in attention-based graph neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[ZLL+23]   Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning*, pages 40770–40803. PMLR, 2023.

[ZMZ+23]   Haiteng Zhao, Shuming Ma, Dongdong Zhang, Zhi-Hong Deng, and Furu Wei. Are more layers beneficial to graph transformers? In *The Eleventh International Conference on Learning Representations*, 2023.

Borjan Geshkovski
Inria & Laboratoire Jacques-Louis Lions
Sorbonne Université
4 Place Jussieu
75005 Paris, France
e-mail: borjan.geshkovski@inria.fr

Hugo Koubbi
Department of Statistics and Data Science
ENS Paris-Saclay & Yale University
219 Prospect Street
New Haven, CT 06511, United States
e-mail: hugo.koubbi@ens-paris-saclay.fr

Yury Polyanskiy
Department of EECS
Massachusetts Institute of Technology
77 Massachusetts Ave
Cambridge 02139 MA, United States
e-mail: yp@mit.edu

Philippe Rigollet
Department of Mathematics
Massachusetts Institute of Technology
77 Massachusetts Ave
Cambridge 02139 MA, United States
e-mail: rigollet@math.mit.edu