Quantifying Concentration Phenomena of Mean-Field Transformers in the Low-Temperature Regime

Albert Alcalde
Department of Mathematics
Friedrich-Alexander-University
Erlangen-Nürnberg
albert.alcalde@fau.de

Leon Bungert
Institute of Mathematics
CAIDAS
University of Würzburg
leon.bungert@uni-wuerzburg.de

Konstantin Riedl
Mathematical Institute
University of Oxford
Reuben College
konstantin.riedl@maths.ox.ac.uk

Tim Roith
CIT School
Technical University of Munich
Munich Center for Machine Learning
tim.roith@tum.de

Abstract

Transformers with self-attention modules as their core components have become an integral architecture in modern large language and foundation models. In this paper, we study the evolution of tokens in deep encoder-only transformers at inference time which is described in the large-token limit by a mean-field continuity equation. Leveraging ideas from the convergence analysis of interacting multi-particle systems, with particles corresponding to tokens, we prove that the token distribution rapidly concentrates onto the push-forward of the initial distribution under a projection map induced by the key, query, and value matrices, and remains metastable for moderate times. Specifically, we show that the Wasserstein distance of the two distributions scales like $\sqrt{\log(\beta+1)/\beta}\exp(Ct)+\exp(-ct)$ in terms of the temperature parameter $\beta^{-1}\to 0$ and inference time $t\geq 0$. For the proof, we establish Lyapunov-type estimates for the zero-temperature equation, identify its limit as $t\to\infty$, and employ a stability estimate in Wasserstein space together with a quantitative Laplace principle to couple the two equations. Our result implies that for time scales of order $\log\beta$, which correspond to the early-phase behavior most relevant for practical finite-depth models, the token distribution concentrates at the identified limiting distribution. Numerical experiments confirm this and, beyond that, complement our theory by showing that for finite β and large t the dynamics enter a different terminal phase, dominated by the spectrum of the value matrix.

1 Introduction

The transformer architecture has driven in recent years the unprecedented successes of modern machine learning, establishing itself as the fundamental backbone for large language models [7, 13, 24, 31], and foundation models [11]. Central to its design is the self-attention mechanism [8, 50], which enables the model to capture intricate, long-range contextual dependencies within the input data. A rigorous characterization of how these internal representations evolve through the network, however, remains a significant analytical challenge.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

To mathematically study how tokens $\{x_i\}_{i=1}^n \subset \mathbb{R}^d$ evolve through the self-attention modules of a deep encoder-only transformer, we follow [15, 17, 30, 46] and model the forward pass of a transformer with single-head attention and without fully-connected MLPs, in the infinite-depth limit, as a continuous-time dynamical interacting particle system of the form

$$\dot{x}_i(t) = \mathrm{P}_{x_i(t)} \left(\sum_{j=1}^n \frac{\exp\left(\beta \left\langle Q(t)x_i(t), K(t)x_j(t) \right\rangle\right)}{\sum_{k=1}^n \exp\left(\beta \left\langle Q(t)x_i(t), K(t)x_k(t) \right\rangle\right)} V(t)x_j(t) \right). \tag{1}$$

System (1) governs the evolution of tokens $\{x_i\}_{i=1}^n$, interpreted as particles, through the self-attention layers of the transformer at inference. Here, $Q(t), K(t) \in \mathbb{R}^{d_{\text{int}} \times d}$ and $V(t) \in \mathbb{R}^{d \times d}$ denote the *query*, *key*, and *value* weight matrices of the attention modules, while $\beta > 0$ is an inverse-temperature parameter [15, 20, 30]. The weight matrices have been obtained during training and are fixed for the inference dynamics, which are the focus of the present work. Here, $\mathrm{P}_x(y) \coloneqq y - \langle x, y \rangle x$ for $y \in \mathbb{R}^d$ is the orthogonal projection onto the tangent space $\mathrm{T}_x\mathbb{S}^{d-1} := \{y \in \mathbb{R}^d : \langle x, y \rangle = 0\}$ of the sphere $\mathbb{S}^{d-1}$ at a point $x \in \mathbb{S}^{d-1}$, and models layer normalization. This is motivated by the simplified root mean square normalization $x \mapsto g \odot x/\|x\|$ with gain $g \in \mathbb{R}^d$ [52], which is used, e.g., in Llama models [49]. We set $g = 1$ and refer to [32] and [17, Section 1(b)] for a recent review of modeling choices for normalizations.

Mathematical setting. Let us now introduce the large-token regime of (1), which serves as the analytical framework for this paper. As done in the largest part of the current literature [15, 17, 30, 46] and justified by memory-efficient variants of transformers such as Universal Transformers [22] and ALBERT [35], we assume in our analysis that the weight matrices Q, K, and V are identical throughout the layers, i.e., that $Q(t) \equiv Q$, $K(t) \equiv K$, and $V(t) \equiv V$ for all $t \geq 0$. For notational simplicity and without loss of generality, we henceforth write $B \coloneqq Q^\top K \in \mathbb{R}^{d \times d}$.

After positional encoding [50, Section 3.5], the transformer dynamics (1) are invariant under permutations allowing the empirical measure $\rho_t^{\beta,n} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i(t)}$ to fully characterize the dynamics, which is described by a continuity equation of the form (2) below. Our theoretical analysis will take place in the infinite-token regime, where the temporal dynamics of the token distribution $(\rho_t^\beta)_{t\geq 0} \subset \mathcal{P}(\mathbb{S}^{d-1})$ with $\rho_0^\beta = \rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$ is captured by the *transformer continuity equation*

$$\partial_t \rho_t^\beta = -\operatorname{div}\left(\rho_t^\beta \, \mathrm{P}_x \left(\int_{\mathbb{S}^{d-1}} \frac{\exp\left(\beta \left\langle x, By \right\rangle\right)}{\int_{\mathbb{S}^{d-1}} \exp\left(\beta \left\langle x, Bz \right\rangle\right) \mathrm{d}\rho_t^\beta(z)} Vy \, \mathrm{d}\rho_t^\beta(y) \right) \right), \tag{2}$$

which describes a flow in the space $\mathcal{P}(\mathbb{S}^{d-1})$ of probability measures over the sphere $\mathbb{S}^{d-1}$. Here, div and ∇ denote the manifold divergence and gradient on the sphere $\mathbb{S}^{d-1}$, respectively. It is well-known that the empirical measure $\rho_t^{\beta,n}$ defined above is a weak solution to (2), however, this equation allows for more complex solutions if the initial condition is not supported on a finite set.

As shown in the seminal work [15] under invertibility assumptions on B and V, and Assumption 3 on the initial datum ρ_0, solutions of (2) converge in the zero-temperature limit $\beta^{-1} \to 0$ to those of

$$\partial_t \rho_t = -\operatorname{div}\left(\rho_t \, \mathrm{P}_x \left(\frac{VB^\top x}{\|B^\top x\|} \right) \right). \tag{3}$$

Moreover, for large times, solutions of (3) concentrate on the top eigenspace of the matrix $VB^\top$, that is, the eigenspace associated with its largest eigenvalue, which we denote by E. Correspondingly, for such a linear subspace E of $\mathbb{R}^d$, we denote by $\mathrm{P}_E : \mathbb{R}^d \to E$ the orthogonal projection onto E and define $\Pi : \mathbb{S}^{d-1} \setminus E^\perp \to \mathbb{S}^{d-1}$ as

$$\Pi(x) \coloneqq \frac{\mathrm{P}_E x}{\|\mathrm{P}_E x\|}. \tag{4}$$

Main Contributions. In this work, we study the evolution of tokens in deep encoder-only transformers at inference time by analyzing the mean-field token distribution ρ_t^β of the transformer continuity equation (2). We quantify its concentration at the push-forward $\Pi_\sharp \rho_0$ of the initial datum under the projection map (4), which is determined by the model weights, in terms of the temperature parameter β^{-1} and inference time t. Our proof leverages tools from the analysis of interacting particle systems, in particular, stability estimates in Wasserstein space involving a quantitative Laplace principle, and Lyapunov techniques to quantify convergence speeds. Our main contributions are as follows.

- Identification of $\Pi_\sharp \rho_0$ as the stationary distribution of the zero-temperature continuity equation (3) for symmetric matrices $VB^\top$, and proof of exponential convergence of the zero-temperature solution ρ_t to the top eigenspace of $VB^\top$ in the Wasserstein distance.
- Derivation of a quantitative Wasserstein stability estimate between solutions ρ_t^β of the transformer continuity equation (2) and the zero-temperature limit ρ_t.
- Quantification of the approximation error between ρ_t^β and $\Pi_\sharp \rho_0$ in Wasserstein distance as $\sqrt{\log(\beta+1)/\beta}\exp(Ct) + \exp(-ct)$, indicating an interplay between temperature β^{-1} and time t.
- Characterization of the non-asymptotic concentration regime of the mean-field transformer dynamics up to times $t = \mathcal{O}(\log \beta)$, corresponding to the early-phase behavior most relevant for practical finite-depth models.
- Experimental identification of the terminal phase after times $t = \mathcal{O}(\log \beta)$, where concentration shifts from the top eigenspace of $VB^\top$ to an eigenspace of V.

Organization. Following the presentation and discussion of our theoretical contributions in Section 2, Section 3 puts these findings into context within the current literature. Thereafter, in Section 4, we present proof details for our statements, with technical results deferred to Appendices A and B. Section 5 then illustrates and complements our theoretical findings with numerical experiments, with additional experiments provided in Appendix C. We conclude the paper with Section 6.

2 Main results

This section is dedicated to discussing the main theoretical contributions of our paper. We first state the assumptions needed for proving our main results, before elaborating on them in more detail.

Assumption 1. *The weight matrices B and V are such that*

- *A1* *the matrix $B = Q^\top K$ is invertible,*
- *A2* *the matrix $VB^\top \in \mathbb{R}^{d\times d}$ is symmetric, with largest eigenvalue μ_1 having multiplicity k and second largest eigenvalue μ_2 satisfying $\gamma \coloneqq \mu_1 - \mu_2 > 0$ (with the convention that $\gamma = \infty$ if $k = d$). We denote by E the eigenspace of $VB^\top$ with respect to μ_1 and by $\mathrm{P}_E : \mathbb{R}^d \to E$ the orthogonal projection onto E.*

Assumption 2. *The initial datum $\rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$ is such that $\mathcal{V}_p(\rho_0) \coloneqq \int_{\mathbb{S}^{d-1}} \frac{\|(\mathrm{Id}-\mathrm{P}_E)x\|^{2p}}{\|\mathrm{P}_E x\|^{2p}}\,\mathrm{d}\rho_0(x)$ is finite for some $p \in (0,1]$.*

Assumption 3. *The initial datum $\rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$ has a continuous density f_0 with respect to the surface measure $\mathcal{H}^{d-1} \llcorner \mathbb{S}^{d-1}$ that satisfies $\ell_0 \coloneqq \min_{x\in\mathbb{S}^{d-1}} f_0(x) > 0$.*

Under Assumptions 1–3 on the weight matrices B and V, and the initial datum ρ_0, we derive in the subsequent statement quantitative concentration bounds for the transformer continuity equation (2). We denote the smallest and largest singular values of a matrix by $\sigma_{\min}(\bullet)$ and $\sigma_{\max}(\bullet)$.

Theorem 1. *Assume that the matrices B and V satisfy Assumption 1 and that the initial distribution $\rho_0 \in \mathcal{P}(\mathbb{S}^{d-1})$ satisfies Assumption 2 for some $p \in (0,1]$ as well as Assumption 3. Let $(\rho_t^\beta)_{t\in[0,T]}$ denote the weak solution to the transformer continuity equation (2). Then, for every $\beta > 0$, it holds*

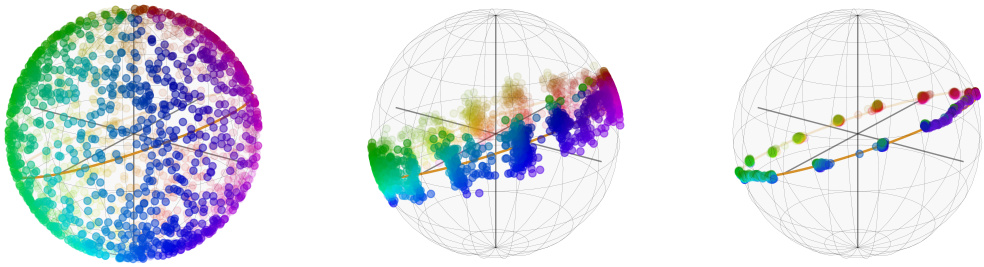
$$W_2(\rho_t^\beta, \Pi_\sharp\rho_0) \le 2\sqrt{\tfrac{\log(\beta+1)}{\beta}}\left(e^{C_1 t} - e^{C_0 t}\right) + \sqrt{2\mathcal{V}_p(\rho_0)}\exp\left(-\tfrac{p\gamma}{\sigma_{\max}(B)}t\right),$$

for constants $C_1 > C_0 > 0$ that depend only on d, $\sigma_{\min}(B)$, $\sigma_{\max}(B)$, $\sigma_{\max}(V)$, and ℓ_0.

Before discussing Theorem 1 and its assumptions in the remainder of this section, let us provide as a corollary a quantitative convergence estimate for the zero-temperature transformer continuity equation (3). It follows directly from the proof of Theorem 1 (see equation (10)) in Section 4.

Corollary 2. *Assume that the matrices B and V satisfy Assumption 1. Moreover, let the initial distribution ρ_0 be such that Assumption 2 holds for some $p \in (0,1]$. Let $(\rho_t)_{t\in[0,T]}$ denote the weak solution to the zero-temperature transformer continuity equation (3). Then, it holds*

$$W_2(\rho_t, \Pi_\sharp\rho_0) \le \sqrt{2\mathcal{V}_p(\rho_0)}\exp\left(-\tfrac{p\gamma}{\sigma_{\max}(B)}t\right).$$



(a) $t = 0$ (b) $t = 2.5$ (c) $t = 5$

Figure 1: Illustration of Theorem 1. We observe concentration of $\rho_t^{\beta,n}$ at $E \cap \mathbb{S}^{d-1}$ (orange circle) when running (1) with $n = 2,000$ tokens to simulate (2) for $B = R^\top \mathrm{diag}(5,5,1)R$ and $V = \mathrm{Id}$, where R is a rotation by $\pi/8$. We use $\beta = 30$ and a von Mises–Fisher mixture as initial density ρ_0.

To the best of our knowledge, Corollary 2 is the first result which identifies the stationary distribution of the zero-temperature transformer continuity equation (3). Note, however, that as observed in [15, Remark 3.6] there exist weight matrices B and V (which do not satisfy A2 in Assumption 1) such that only the collapse onto E but no convergence can be expected.

Discussion of the assumptions. Demanding invertibility of the product $B = Q^\top K$ of the key and query matrices in A1 of Assumption 1 is relatively common in the field, see, e.g., [15, 18, 29, 30, 45]. It is especially important when working with the zero-temperature continuity equation (3) which is not well-defined unless B is invertible. The symmetry assumption on $VB^\top$ from A2 might appear to be relatively strong. It is important to note, however, that without an assumption of this form one cannot hope to be able to characterize the stationary distribution $\Pi_\sharp \rho_0$ for (3) or prove exponential convergence to it as done in Corollary 2. Still, the quantitative stability result from Proposition 6, stating that $W_2(\rho_t^\beta, \rho_t) \leq 2\sqrt{\log(\beta+1)/\beta}(e^{C_1 t} - e^{C_0 t})$, is true without A2. As a consequence, whenever one can analyze the convergence of the zero-temperature solution ρ_t under more relaxed assumptions than A2, such results extend to ρ_t^β for large β by the triangle inequality. See also Appendix C.2 for related numerical experiments.

Moving on to Assumption 2, we note that this assumption implies that $\rho_0(E^\perp \cap \mathbb{S}^{d-1}) = 0$ since otherwise the denominator would be zero on a set of positive measure. This is necessary: Since $E^\perp \cap \mathbb{S}^{d-1}$ is invariant for (3) when $VB^\top$ is symmetric, any mass initially supported on $E^\perp$ remains there for all times and cannot flow toward E. However, Assumption 2 is stronger requiring that not too much mass accumulates close to $E^\perp \cap \mathbb{S}^{d-1}$ either. This is necessary to prove Corollary 2 with Lyapunov techniques. We suspect that it might be possible to weaken this assumption or even replace it by $\rho_0(E^\perp \cap \mathbb{S}^{d-1}) = 0$ if one is willing to worsen or sacrifice having a convergence rate altogether.

Assumption 3 is only needed for the quantitative Laplace principle from Lemma 5. Loosely speaking, for token distributions ρ_t^β with zero mass on open subsets of the sphere, the vector field in (2) is a poor approximation of the one in (3) even for large β. The authors of [15] also used Assumption 3 to show that ρ_t^β has a positive density for all finite times $t > 0$, and similar assumptions were used in [14, 19, 45]. One possible way to remove this assumption is to add noise, leading to noisy transformer models as recently considered in [10, 45, 48]. This is also exploited in the analysis of the conceptually similar consensus-based optimization model [27]. Note, however, that the exponential convergence of the solution to the zero-temperature continuity equation (3) from Corollary 2 holds without Assumption 3 and is, in particular, valid for singular initial distributions concentrated on finitely many points as long as they do not lie on the lower-dimensional sphere $E^\perp \cap \mathbb{S}^{d-1}$, see the numerical experiments for $\beta = \infty$ in Appendix C.3.

Let us remark that at first glance it might seem like Assumptions 2 and 3 contradict each other. After all, the former prevents mass from accumulating close to $E^\perp \cap \mathbb{S}^{d-1}$, while the latter demands a strictly positive density. However, this is no contradiction if the parameter p and the dimension of the eigenspace E are in a correct relationship to each other, as demonstrated by the following example.

Example 3. Since Assumption 3 implies that $\rho_0 \geq \ell_0 \, \mathcal{H}^{d-1} \llcorner \mathbb{S}^{d-1}$, it is necessary for Assumption 2 to verify whether $\int_{\mathbb{S}^{d-1}} \|(\mathrm{Id} - \mathrm{P}_E)x\|^{2p} \, \|\mathrm{P}_E x\|^{-2p} \, \mathrm{d}\mathcal{H}^{d-1} < \infty$. In fact, it suffices to check whether $\int_{\mathbb{S}^{d-1}} \|\mathrm{P}_E x\|^{-2p} \, \mathrm{d}\mathcal{H}^{d-1}(x) < \infty$. Note that, for $p > 0$, the singular set of the integrand is $S \coloneqq E^\perp \cap \mathbb{S}^{d-1}$ and equals a $(d-k-1)$-dimensional sphere. In a neighborhood of S the quantity $\|\mathrm{P}_E x\|$ scales like the distance of x to S. The integral over a tubular ε-neighborhood S_ε of S in spherical coordinates therefore scales like $\int_{S_\varepsilon} \|\mathrm{P}_E x\|^{-2p} \, \mathrm{d}\mathcal{H}^{d-1} \sim \int_0^\varepsilon \frac{1}{r^{2p}} r^{k-1} \, \mathrm{d}r = \int_0^\varepsilon r^{k-1-2p} \, \mathrm{d}r$, which is finite if and only if $k > 2p$. While for $p < \frac{1}{2}$ this condition is automatically satisfied, it can be a restriction for larger p. For instance, if $p = 1$ it requires the dimension of the top eigenspace E of $VB^\top$ to be at least three. Interestingly, the condition $\dim E \geq 3$ also appeared in [19, Theorem 3.1] for analyzing clustering phenomena of mean-field transformer models.

Discussion of the main results. Our main result, Theorem 1, encapsulates two different effects that are at play for solutions to the transformer continuity equation (2). The first one is a temperature effect stating that for large values of β (corresponding to low temperatures), solutions ρ_t^β of (2) are close to solutions ρ_t of the zero-temperature continuity equation (3) in the Wasserstein distance. In fact, Proposition 6 in Section 4.1 states that $W_2(\rho_t^\beta, \rho_t) \leq 2\sqrt{\log(\beta+1)/\beta}(e^{C_1 t} - e^{C_0 t})$, which quantifies [15, Theorem 3.2]. This estimate is not uniform in time due to the exponential factor, and we do not expect it to be either, as demonstrated by our numerical experiments in Section 5. The second effect is the exponential convergence of the zero-temperature limit (3) to a stationary distribution ρ_∞ which we show to depend explicitly on the initial datum and to have the form $\rho_\infty \coloneqq \Pi_\sharp \rho_0$. Taking into account the definition of Π in (4), ρ_∞ can be thought of as a projection of ρ_0 onto $E \cap \mathbb{S}^{d-1}$.

The time scales at which these two effects are present can be derived directly from Theorem 1 by investigating which of the two terms dominates. Until times of order $\mathcal{O}(\log \beta)$, the exponential convergence prevails. Thereafter, the approximation error due to positive temperature takes over, see also Figure 4 for a numerical illustration of this behavior.

$$W_2(\rho_t^\beta, \Pi_\sharp \rho_0) \lesssim \begin{cases} \exp\left(-\frac{p\gamma}{\sigma_{\max}(B)} t\right), & 0 \leq t \lesssim \log \beta, \\ e^{C_1 t} \sqrt{\frac{\log(\beta+1)}{\beta}}, & t \gtrsim \log \beta. \end{cases}$$

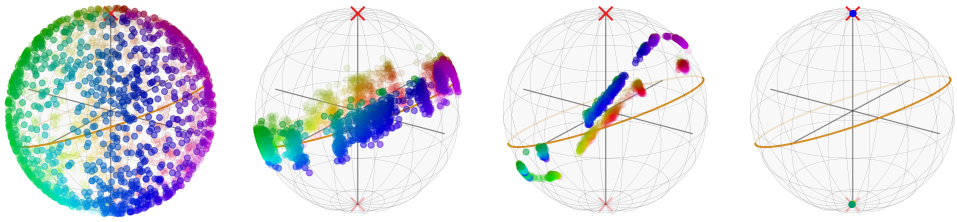
Hence, the most relevant regime for our result is that of time scales of at most $\mathcal{O}(\log \beta)$. We empirically validate this finding in Appendix C.5 by showing that the time t^β at which $W_2(\rho_t^\beta, \Pi_\sharp \rho_0)$ is minimal indeed scales like $\mathcal{O}(\log \beta)$. On such scale, the following corollary of Theorem 1 states that for any given target accuracy $\varepsilon > 0$, there exists $\beta > 0$ large enough such that after a time of order $\log \frac{1}{\varepsilon}$, the Wasserstein distance of ρ_t^β to $\Pi_\sharp \rho_0$ is below ε in a proper time interval. See Appendix A.3 for the proof of the following result.

Corollary 4. *For any $\varepsilon > 0$ and for all $\beta \gg \mathrm{poly}\left(\frac{1}{\varepsilon}\right) \log \frac{1}{\varepsilon}$ large enough it holds $W_2(\rho_t^\beta, \Pi_\sharp \rho_0) \leq \varepsilon$ for all $t \in [t_1, t_2]$ with $0 \leq t_1 < t_2$, where*

$$t_1 \coloneqq \tfrac{\sigma_{\max}(B)}{p\gamma} \max\left(\log\left(\tfrac{2\sqrt{2\mathcal{V}_p(\rho_0)}}{\varepsilon}\right), 0\right) \quad \text{and} \quad t_2 \coloneqq \tfrac{1}{C_1} \log\left(1 + \tfrac{\varepsilon}{4}\sqrt{\tfrac{\beta}{\log(\beta+1)}}\right).$$

Conjectured long-time behavior. We close this section by emphasizing that there is no reason to believe that after a time scale of order $\mathcal{O}(\log \beta)$ the token distribution ρ_t^β is still close to $\Pi_\sharp \rho_0$. In fact, we have numerical evidence suggesting convergence to the top eigenspace F of V, that is, the eigenspace associated with its largest eigenvalue, or to the dominant eigenspace F^{abs} of V associated with an eigenvalue of largest magnitude for large times in certain cases (see Figure 2). This observation is supported by the asymptotic stability results from [4], and it neither contradicts our result nor the three-phase analysis of [15]. Phases 2 and 3 described therein apply only once the distribution ρ_t^β has fully collapsed onto $E \cap \mathbb{S}^{d-1}$, and they are derived under the assumption that $V = \mathrm{Id}$. Under this assumption, the following conjecture becomes partially trivial, since $F = \mathbb{R}^d$ and hence $F \cap \mathbb{S}^{d-1} = \mathbb{S}^{d-1}$, while the convergence statement of the conjecture remains non-trivial. We provide heuristics and numerical experiments supporting the conjecture in Appendices B and C.4, respectively.

Conjecture 1. *There exist weight matrices B and $V \neq \mathrm{Id}$ satisfying Assumption 1, and a probability measure $\rho_\infty \in \mathcal{P}(\mathbb{S}^{d-1})$ supported on $F \cap \mathbb{S}^{d-1}$ or $F^{\mathrm{abs}} \cap \mathbb{S}^{d-1}$ such that $W_2(\rho_t^\beta, \rho_\infty) \to 0$ as $t \to \infty$.*



(a) $t = 0$ (b) $t = 2.5$ (c) $t = 4$ (d) $t = 10$

Figure 2: Illustration of Conjecture 1. Using the setup of Figure 1, with the same matrix $VB^\top$, but with $V = \mathrm{diag}(1, 1, 2)$ and B adapted accordingly, we observe initial concentration of $\rho_t^{\beta,n}$ at $E \cap \mathbb{S}^{d-1}$ (orange circle) before the terminal clustering in $F \cap \mathbb{S}^{d-1}$ (red crosses).

3 Related work

Let us now relate our results to previous works in the literature on the mathematical analysis of transformer models, and outline their connection to a class of interacting particle systems.

Mathematical analysis of transformers. A central body of work models transformer layers through the lens of continuous dynamical systems and partial differential equations (PDEs). This perspective was pioneered in [46] by linking attention to transport-type PDEs, and expanded in [30] by incorporating key architectural components such as layer normalization within a mean-field framework. More recent contributions such as [18] provide a unified treatment of these approaches. Together, these works yield a concise mathematical description of transformer dynamics that serves also as our modeling foundation. The authors of [15] identify the zero-temperature limit (3) and establish convergence towards it as $\beta^{-1} \to 0$. However, their result is not quantitative, i.e., no rate in β nor its interplay with inference time t is provided. The analysis further identifies additional phases of the dynamics once the solution concentrates on the top eigenspace E of $VB^\top$. In contrast, our work takes low but non-zero temperature effects into account and provides quantitative convergence guarantees. This perspective is motivated by the numerical evidence in Section 5 which suggests that $\beta^{-1} > 0$ may prevent full collapse onto the eigenspace E. Related to this, [17] also focuses on equation (2) with $\beta^{-1} > 0$ and characterizes the stationary states of (2) in the case $B = \pm V$, where the system can be interpreted as a gradient flow of the energy $\mathcal{E}(\varrho) = \iint e^{\langle x, By\rangle} \,\mathrm{d}\varrho(x)\,\mathrm{d}\varrho(y)$, either maximizing it if $B = V$ or minimizing it if $B = -V$. We discuss this case in Appendix C.3 and only emphasize here that our results complement their theory in two ways: we go beyond the gradient flow setting by allowing $B \neq V$ and, in particular, make concrete statements about the asymptotic behavior of ρ_t^β. In [39] this gradient flow interpretation was extended to multi-head attention with time-dependent parameter matrices.

Besides mean-field descriptions, a complementary line of research investigates transformer dynamics at the finite-particle level, with an emphasis on qualitative and asymptotic behavior. In particular, clustering and concentration phenomena are analyzed in [19, 30, 45], with extensions to causal attention settings in [1, 33]. The long-time behavior of transformer dynamics has also been studied, including the emergence of metastability [3, 28], multistability with several equilibria related to the spectrum of the value matrix [4], and phase transitions [21], as well as propagation-of-chaos results in large-token limits [14]. Our work differs by considering token distributions with positive densities and providing quantitative estimates of concentration phenomena in an initial non-asymptotic phase of the dynamics where the attention matrix B plays a crucial role.

Interacting particle systems. Transformers share striking similarities with classical interacting particle systems (e.g., the Kuramoto model [37, 42]), and optimization methods (e.g., gradient descent [51] or the Frank–Wolfe algorithm [3]). In particular, they are connected (as noted in [30]) to *consensus-based optimization* (CBO) [41, 44] which we formalize and exploit in what follows. In line with our setting, we consider CBO constrained to the sphere, as in [26]. The aim is to locate a global minimizer of some objective function $J : \mathbb{S}^{d-1} \to \mathbb{R}$ by evolving a particle ensemble $\{x_i\}_{i=1}^n$

that follows a coupled system of stochastic differential equations of the form

$$\dot{x}_i(t) = \mathrm{P}_{x_i(t)}\big(m_{\beta,\rho_t^{\beta,n}}\big) + \sigma \cdot \text{noise}\,. \tag{5}$$

The so-called consensus point $m_{\beta,\varrho}$, defined for any probability measure $\varrho \in \mathcal{P}(\mathbb{S}^{d-1})$, is a weighted mean to which particles with smaller values of J contribute more. In fact, particle weights can also be determined through a pairwise interaction via polarization, as proposed in [16]. Mathematically, this amounts to kernelizing the consensus point as

$$m^{\kappa}_{\beta,\varrho}(x) = \int_{\mathbb{S}^{d-1}} \frac{\kappa(x,y)\exp\left(-\beta J(y)\right)}{\int_{\mathbb{S}^{d-1}} \kappa(x,z)\exp\left(-\beta J(z)\right)\mathrm{d}\varrho(z)}\, y\,\mathrm{d}\varrho(y), \tag{6}$$

where $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is some kernel. With $\|\bullet\|_B \coloneqq \sqrt{\langle \bullet, B\,\bullet\rangle}$ being the energy-norm, we choose

$$J(x) = -\tfrac{1}{2}\,\|x\|_B^2\,, \quad \text{and} \quad \kappa(x,y) = \exp\Big(-\tfrac{\beta}{2}\,\|x-y\|_B^2\Big).$$

In the noiseless case $\sigma = 0$, equation (5) with consensus point as in (6) leads exactly to the transformer model (1) when $V \equiv \mathrm{Id}$ and $B \equiv Q^\top K$ is time-independent. This interpretation motivated us to leverage techniques from the mean-field analysis of interacting particle systems for optimization—most prominently, the quantitative Laplace principle and Lyapunov techniques—to analyze the transformer model (2). Apart from this, by embedding the self-attention dynamics (1) into the rich family of CBO models [16, 27, 43], one can hope to derive more expressive transformers by changing the kernel, the objective function, or by adding additional drift terms and noise (the latter of which leads to noisy transformers as previously investigated in, e.g., [2, 10, 25, 34, 40, 45, 47, 48]).

4 Proof details for the main results

The proof of Theorem 1 relies on two key ingredients: quantitative stability estimates for continuity equations in Wasserstein space which leverage a quantitative Laplace principle, and Lyapunov-type estimates for the Wasserstein distance to control the convergence of the zero-temperature transformer continuity equation (3). While the former allow to control the Wasserstein distance between the flows ρ_t^β and ρ_t (see Proposition 6), the latter enable us to quantify the convergence rate of ρ_t to $\Pi_\sharp\rho_0$ (see Proposition 8 in combination with Lemmas 7 and 9). Before presenting these key ingredients in Sections 4.1 and 4.2, we give the proof of Theorem 1, which demonstrates how those tools are used.

Proof of Theorem 1. Denoting by $(\rho_t^\beta)_{t\in[0,T]}$ the solution of (2), and letting $(\rho_t)_{t\in[0,T]}$ denote the solution of (3) with coinciding initial datum $\rho_0 = \rho_0^\beta$, we can bound by the triangle inequality

$$W_2(\rho_t^\beta, \Pi_\sharp\rho_0) \leq W_2(\rho_t^\beta, \rho_t) + W_2(\rho_t, \Pi_\sharp\rho_0). \tag{7}$$

The first term on the right-hand side of (7) can be estimated by employing the quantitative stability estimate in Wasserstein space from Proposition 6 for the continuity equations (2) and (3), yielding

$$W_2(\rho_t^\beta, \rho_t) \leq 2\sqrt{\tfrac{\log(\beta+1)}{\beta}}\left(e^{C_1 t} - e^{C_0 t}\right) \tag{8}$$

for all $t \in [0,T]$. For the second term on the right-hand side of (7), we first note that thanks to Proposition 8 it holds for the functional $\mathcal{V}_p$ defined in (14) that

$$\mathcal{V}_p(\rho_t) \leq \mathcal{V}_p(\rho_0)\exp\Big(-\tfrac{2p\gamma}{\sigma_{\max}(B)}t\Big). \tag{9}$$

With $\mathcal{V}_p(\rho_0) < \infty$ as of Assumption 2, we have that $\mathcal{V}_p(\rho_t) < \infty$ for all $t \in [0,T]$. This implies that $\rho_t(\mathbb{S}^{d-1} \cap E^\perp) = 0$ for all $t \in [0,T]$, since otherwise the denominator in the integrand of $\mathcal{V}_p$ would be zero on a set of positive measure. Since according to Lemma 9 it furthermore holds that $\Pi_\sharp\rho_t = \Pi_\sharp\rho_0$ for all $t \in [0,T]$, an application of Lemma 7 together with the bound (9) shows that

$$W_2(\rho_t, \Pi_\sharp\rho_0) = W_2(\rho_t, \Pi_\sharp\rho_t) \leq \sqrt{2\mathcal{V}_p(\rho_t)} \leq \sqrt{2\mathcal{V}_p(\rho_0)}\exp\Big(-\tfrac{p\gamma}{\sigma_{\max}(B)}t\Big), \tag{10}$$

verifying that $\sqrt{2\mathcal{V}_p(\rho_t)}$ is a suitable Lyapunov functional for $W_2(\rho_t, \Pi_\sharp\rho_0)$. Inserting (8) and (10) into (7) concludes the proof. □

4.1 Quantitative estimates for the Wasserstein distance between the flows ρ_t^β and ρ_t

To estimate the Wasserstein distance $W_2(\rho_t^\beta, \rho_t)$ of the solutions ρ_t^β and ρ_t of the transformer continuity equations (2) and (3) in terms of the temperature parameter β, we develop a stability estimate in Wasserstein space, which is based on a quantitative version [27] of the Laplace principle [23]. Denoting by $\varrho \in \mathcal{P}(\mathbb{S}^{d-1})$ a fully supported probability measure, for any fixed $x \in \mathbb{S}^{d-1}$ the vector

$$m_{\beta,\varrho}(x) \coloneqq \int_{\mathbb{S}^{d-1}} \frac{\exp\left(\beta \left\langle B^\top x, y\right\rangle\right)}{\int_{\mathbb{S}^{d-1}} \exp\left(\beta \left\langle B^\top x, z\right\rangle\right) \mathrm{d}\varrho(z)} y \, \mathrm{d}\varrho(y) \tag{11}$$

converges as $\beta \to \infty$ to a maximizer of the function $\mathbb{S}^{d-1} \ni y \mapsto \left\langle B^\top x, y\right\rangle$, which is given by

$$y^*(x) \coloneqq \frac{B^\top x}{\|B^\top x\|}, \qquad x \in \mathbb{S}^{d-1}, \tag{12}$$

if B is invertible. More rigorously, the next lemma holds; see Appendix A.1.1 for the proof.

Lemma 5. *Assume that the matrix B satisfies A1. Let $\varrho \in \mathcal{P}(\mathbb{S}^{d-1})$, $r > 0$, and define the spherical cap $\mathcal{B}_r(y^*(x)) \coloneqq \left\{y \in \mathbb{S}^{d-1} : \|y - y^*(x)\| \le r\right\}$. Then, for every $\beta > 0$ and every $q > 0$, it holds*

$$\left\|m_{\beta,\varrho}(x) - y^*(x)\right\| \le \sqrt{r^2 + \tfrac{2q}{\sigma_{\min}(B)}} + \tfrac{2e^{-\beta q}}{\varrho(\mathcal{B}_r(y^*(x)))}. \tag{13}$$

Using this result, we quantify the convergence $\rho_t^\beta \to \rho_t$ as $\beta \to \infty$, which has been shown earlier in [15, Theorem 3.2] under invertibility assumptions on B and V as well as Assumption 3 on the initial datum ρ_0. We show the following; see Appendix A.1.2 for the proof.

Proposition 6. *Assume that the matrix B satisfies A1. Let $(\rho_t^\beta)_{t\in[0,T]}$ and $(\rho_t)_{t\in[0,T]}$ denote weak solutions to the continuity equations* (2) *and* (3) *with coinciding initial datum $\rho_0^\beta = \rho_0$, which is assumed to satisfy Assumption 3. Then, there exist constants $C_1 > C_0 > 0$ that depend only on d, $\sigma_{\min}(B)$, $\sigma_{\max}(B)$, $\sigma_{\max}(V)$, and ℓ_0, such that for every $\beta > 0$, it holds*

$$W_2(\rho_t^\beta, \rho_t) \le 2\sqrt{\tfrac{\log(\beta+1)}{\beta}} \left(e^{C_1 t} - e^{C_0 t}\right).$$

4.2 Lyapunov-type estimates in Wasserstein space controlling the convergence of ρ_t to $\Pi_\sharp \rho_0$

To measure how well the token distribution ρ_t of (3) concentrates near the top eigenspace E of $VB^\top$, let us define for a measure $\varrho \in \mathcal{P}(\mathbb{S}^{d-1})$ and for any $p \in (0, 1]$, the energy (Lyapunov) functional

$$\mathcal{V}_p(\varrho) \coloneqq \int_{\mathbb{S}^{d-1}} R_p(x) \, \mathrm{d}\varrho(x) \quad \text{with} \quad R_p(x) \coloneqq \frac{\|(\mathrm{Id} - \mathrm{P}_E)x\|^{2p}}{\|\mathrm{P}_E x\|^{2p}}. \tag{14}$$

Here, R_p is equal to zero on E and infinity on $E^\perp$. The following auxiliary result shows that R_p can be regarded as a distance from E; see Appendix A.2.1 for the proof.

Lemma 7. *Let $E \subset \mathbb{R}^d$ be a linear subspace and let $p \in (0, 1]$. Then, for all $x \in \mathbb{S}^{d-1} \setminus E^\perp$, it holds that $\|x - \Pi(x)\|^2 \le 2R_p(x)$ and for any probability measure $\varrho \in \mathcal{P}(\mathbb{S}^{d-1})$ with $\varrho(E^\perp \cap \mathbb{S}^{d-1}) = 0$ we have*

$$W_2^2(\varrho, \Pi_\sharp \varrho) \le 2 \int_{\mathbb{S}^{d-1}} R_p(x) \, \mathrm{d}\varrho(x) = 2\mathcal{V}_p(\varrho).$$

Along the flow ρ_t of the zero-temperature equation (3), we can quantify the decay of $\mathcal{V}_p(\rho_t)$ as demonstrated in the subsequent result; see Appendix A.2.2 for the proof.

Proposition 8. *Assume that the matrices B and V satisfy Assumption 1. Moreover, let the initial distribution ρ_0 be such that Assumption 2 holds for some $p \in (0, 1]$. Let $(\rho_t)_{t\in[0,T]}$ denote the weak solution of* (3) *with initial datum ρ_0. Then, for all $t \in [0, T]$, it holds*

$$\mathcal{V}_p(\rho_t) \le \mathcal{V}_p(\rho_0) \exp\left(-\tfrac{2p\gamma}{\sigma_{\max}(B)} t\right).$$

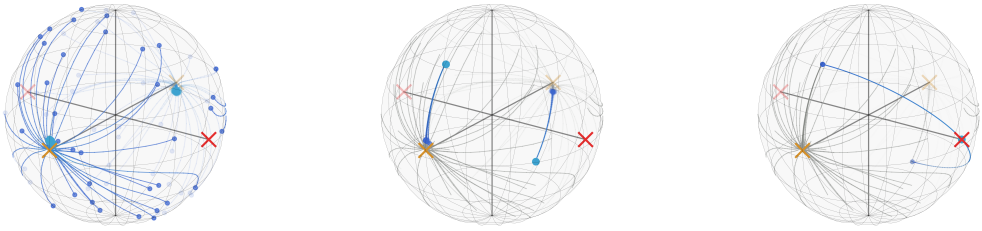
Noticing that the projection Π defined in (4) is invariant in time for a solution ρ_t of (3), as made rigorous by the next lemma, Lemma 7 justifies that $\mathcal{V}_p$ is indeed a suitable Lyapunov functional for the flow ρ_t since $W_2^2(\rho_t, \Pi_\sharp \rho_0) = W_2^2(\rho_t, \Pi_\sharp \rho_t) \leq 2\mathcal{V}_p(\rho_t)$ with $\mathcal{V}_p(\rho_t)$ decaying exponentially fast thanks to Proposition 8. The proof of Lemma 9 is given in Appendix A.2.3.

Lemma 9. *Assume that the matrices B and V satisfy Assumption 1. Moreover, let the initial distribution ρ_0 be such that Assumption 2 holds for some $p \in (0, 1]$. Let $(\rho_t)_{t\in[0,T]}$ denote the weak solution of* (3) *with initial datum ρ_0. Then it holds $\Pi_\sharp \rho_t = \Pi_\sharp \rho_0$ for all $t \in [0, T]$.*

5 Numerical experiments

We provide numerical experiments that validate Theorem 1, and investigate the long-time dynamics beyond the non-asymptotic concentration regime. In particular, we observe the predicted concentration near the top eigenspace E of $VB^\top$ up to time scales of order $\log \beta$, followed by a terminal phase in which tokens concentrate on either F or F^{abs}, consistent with Conjecture 1.

For our simulations, we use an explicit Euler discretization of (1) with timestep $\Delta t = 0.01$.[1] The remaining parameters are specified individually for each experiment. We first fix $\beta = 100$, $d = 3$,



(a) $t \in [0, 4)$ (b) $t \in [4, 9)$ (c) $t \in [9, 20)$

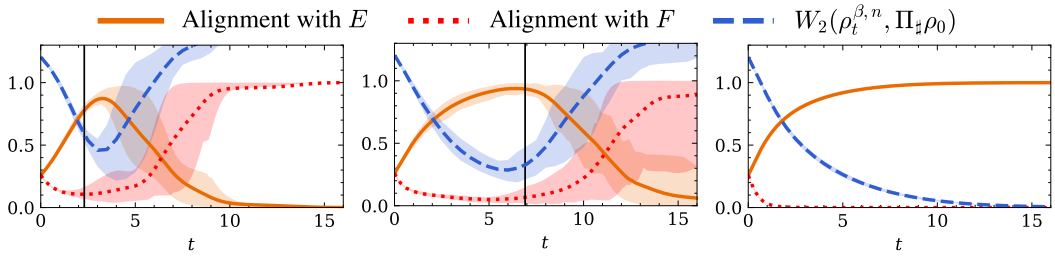
Figure 3: Dynamics of $n = 200$ tokens evolving according to (1) with $\beta = 100$. We show the top eigenspace E of $VB^\top$ (orange crosses) and the top eigenspace F of V (red crosses).

$n = 200$ tokens, and parameter matrices $V = \mathrm{diag}(-1, 1, -2)$ and $B = \mathrm{diag}(-1, -1, 1)$ such that $VB^\top = \mathrm{diag}(1, -1, -2)$. In Figure 3, we depict the dynamics until a time horizon $T \approx 20$. Initially, we observe concentration on the e_1-direction which corresponds to the top eigenspace of $VB^\top$. On a second time scale, tokens align with the e_2-direction, corresponding to F, the top eigenspace of V, in line with Conjecture 1. Figure 4 quantitatively compares the mean-field dynamics for moderate, large, and infinite inverse temperature β. For moderate β, the solution first approaches the projection $\Pi_\sharp \rho_0$ associated with the top eigenspace E of $VB^\top$, as indicated by the decrease of the Wasserstein distance and the transient growth of the alignment with E, as predicted by Theorem 1. After a time of order $\log \beta$, this metastable behavior gives way to the second time scale, on which the solution aligns instead with the top eigenspace F of V. As shown in panel (b) of Figure 4, increasing β delays this transition, while in the limiting case $\beta = \infty$, the second phase is absent by Corollary 2 and the dynamics remain aligned with E. We point to Appendix C for parameter choices of B and V that instead lead to concentration on F^{abs}, the eigenspace of V associated with an eigenvalue of largest magnitude.

6 Concluding remarks

We provided a quantitative analysis of token concentration phenomena in mean-field transformers, showing that their zero-temperature limit $\beta^{-1} \to 0$ is accurate up to time scales of order $\log \beta$ and identifying the corresponding limiting distribution as a projection of the initial distribution ρ_0 onto the top eigenspace of the matrix $VB^\top$ together with convergence rates thereto. This metastable regime corresponds to the early-phase behavior most relevant for practical finite-depth models. In addition, we provided experimental and theoretical evidence for a terminal phase in which the dynamics align with one of the eigenspaces F or F^{abs} of the value matrix V.

[1] We employ the `CBX` library [9] to run our normalized self-attention dynamics with suitable objective and kernel function. Our code can be found in the following `GitHub` repository: `https://github.com/albertalcalde/mean-field-transformer-concentration` (DOI: 10.5281/zenodo.23726584).



(a) $\beta = 10$ (b) $\beta = 10^3$ (c) $\beta = \infty$

Figure 4: Alignment of $\rho_t^{\beta,n}$ with the top eigenspaces E of $VB^\top$, and F of V for $d = 10$ (maximum alignment corresponds to a value of 1). The blue curve shows the Wasserstein distance quantified in Theorem 1, and the vertical line marks $t = \log \beta$. The diagonal matrices V, B are fixed across trials, with normally distributed entries. Curves show the mean over 20 runs with $n = 500$ tokens sampled from the uniform distribution ρ_0; shaded regions show empirical 0.10–0.90 quantile.

Extensions and limitations. There are a few limitations of our work which constitute important directions to address in future work. A natural next step is to analyze the regime of finitely many layers and tokens, i.e., time and space discretizations of (2). This includes in particular quantitative estimates for the mean-field approximation with respect to the number of tokens. Further generalizations include investigating multi-head attention, where the projection appearing in (1) is applied to the average of the outputs of multiple attention heads, and enriching the dynamics by adding noise [10, 45, 48] or a multi-layer perceptron [5]. Another limitation concerns the symmetry assumption on the matrix $VB^\top$. While this assumption enables us to characterize the stationary distribution of the zero-temperature continuity equation, it restricts the dynamics compared to the general case where no stationary distribution might exist, cf. [15]. Besides these extensions, a systematic understanding of Conjecture 1 and the quantification of the convergence in the terminal phase that we observed numerically, as well as its possible connections to the Oja flow of the value matrix V, see [4], are also of interest.

Acknowledgments and Disclosure of Funding

AA was funded by the European Union through the ModConFlex project ("Modelling and control of flexible structures interacting with fluids"; Grant Agreement No. 101073558; DOI: 10.3030/101073558). AA also thanks the University of Würzburg for its hospitality during a research stay in which a substantial part of this work was carried out. LB and TR acknowledge funding by the Federal Ministry of Research, Technology and Space of Germany (BMFTR) under grant agreement No. 16IS24072 (COMFORT). LB acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 544579844 (GeoMAR). KR was supported by the grant "DMS-EPSRC: Asymptotic Analysis of Online Training Algorithms in Machine Learning: Recurrent, Graphical, and Deep Neural Networks" (NSF DMS-2311500). TR acknowledges the support of the Munich Center for Machine Learning and the ERC Advanced Grant NEITALG, grant agreement No. 101198055.

For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.



Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. This project has received funding from the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (grant agreement No. 101198055, project acronym NEITALG).

References

[1] Á. R. Abella, J. P. Silvestre, and P. Tabuada. The asymptotic behavior of attention in transformers. *arXiv preprint arXiv:2412.02682*, 2024.

[2] A. Agazzi, G. Bruno, E. M. García, S. Saviozzi, and M. Romito. Stochastic scaling limits and synchronization by noise in deep transformer models. *arXiv preprint arXiv:2604.26898*, 2026.

[3] A. Alcalde, B. Geshkovski, and D. Ruiz-Balet. Attention's forward pass and Frank-Wolfe. In A. Agarwal, M. Dudik, S. Li, and M. Jaggi, editors, *Proceedings of the 43rd International Conference on Machine Learning, ICML 2026, Seoul, South Korea, July 6–11, 2026*, Proceedings of Machine Learning Research. PMLR, 2026.

[4] C. Altafini. Multistability of self-attention dynamics in transformers. *IEEE Transactions on Automatic Control*, 2026.

[5] A. Álvarez-López, B. Geshkovski, and D. Ruiz-Balet. Perceptrons and localization of attention's mean-field landscape. In A. Agarwal, M. Dudik, S. Li, and M. Jaggi, editors, *Proceedings of the 43rd International Conference on Machine Learning, ICML 2026, Seoul, South Korea, July 6–11, 2026*, Proceedings of Machine Learning Research. PMLR, 2026.

[6] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[7] R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[8] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*, 2015.

[9] R. Bailo, A. Barbaro, S. Gomes, K. Riedl, T. Roith, C. Totzeck, and U. Vaes. CBX: Python and Julia packages for consensus-based interacting particle methods. *J. Open Source Softw.*, 9(98):6611, 2024.

[10] K. Balasubramanian, S. Banerjee, and P. Rigollet. On the structure of stationary solutions to McKean-Vlasov equations with applications to noisy transformers. *arXiv preprint arXiv:2510.20094*, 2025.

[11] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

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

[13] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual*, 2020.

[14] G. Bruno, F. Pasqualotto, and A. Agazzi. Emergence of meta-stable clustering in mean-field transformer models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24–28, 2025*, 2025.

[15] G. Bruno, F. Pasqualotto, and A. Agazzi. A multiscale analysis of mean-field transformers in the moderate interaction regime. In D. Belgrave, C. Zhang, L. N. Montoya, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, N. Chen, I. V. M. Ruíz, and A. Loaiza-Bonilla, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2–7, 2025 / Mexico City, Mexico, November 30 – December 5, 2025*, 2025.

[16] L. Bungert, T. Roith, and P. Wacker. Polarized consensus-based dynamics for optimization and sampling. *Math. Program.*, 211(1-2):125–155, 2025.

[17] M. Burger, S. Kabri, Y. Korolev, T. Roith, and L. Weigand. Analysis of mean-field models arising from self-attention dynamics in transformer architectures with layer normalization. *Philos. Trans. Roy. Soc. A*, 383(2298):Paper No. 20240233, 48, 2025.

[18] V. Castin, P. Ablin, J. A. Carrillo, and G. Peyré. A unified perspective on the dynamics of deep transformers. *arXiv preprint arXiv:2501.18322*, 2025.

[19] S. Chen, Z. Lin, Y. Polyanskiy, and P. Rigollet. Quantitative clustering in mean-field transformer models. *arXiv preprint arXiv:2504.14697*, 2025.

[20] S. Chen, Z. Lin, Y. Polyanskiy, and P. Rigollet. Critical attention scaling in long-context transformers. In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23–27, 2026*, 2026.

[21] A. Cowsik, T. Nebabu, X. Qi, and S. Ganguli. Geometric dynamics of signal propagation predict trainability of transformers. *Phys. Rev. E*, 112(5):Paper No. 055301, 13, 2025.

[22] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and L. Kaiser. Universal transformers. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*, 2019.

[23] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 1998.

[24] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[25] M. Engel and A. Shalova. Random quadratic form on a sphere: Synchronization by common noise. *arXiv preprint arXiv:2603.06187*, 2026.

[26] M. Fornasier, H. Huang, L. Pareschi, and P. Sünnen. Consensus-based optimization on hypersurfaces: Well-posedness and mean-field limit. *Math. Models Methods Appl. Sci.*, 30(14):2725–2751, 2020.

[27] M. Fornasier, T. Klock, and K. Riedl. Consensus-Based Optimization Methods Converge Globally. *SIAM J. Optim.*, 34(3):2973–3004, 2024.

[28] B. Geshkovski, H. Koubbi, Y. Polyanskiy, and P. Rigollet. Dynamic metastability in the self-attention model. *arXiv preprint arXiv:2410.06833*, 2024.

[29] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet. The emergence of clusters in self-attention dynamics. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10–16, 2023*, 2023.

[30] B. Geshkovski, C. Letrouit, Y. Polyanskiy, and P. Rigollet. A mathematical perspective on transformers. *Bull. Amer. Math. Soc. (N.S.)*, 62(3):427–479, 2025.

[31] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024.

[32] N. Karagodin, S. Ge, Y. Polyanskiy, and P. Rigollet. Normalization in attention dynamics. In D. Belgrave, C. Zhang, L. N. Montoya, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, N. Chen, I. V. M. Ruíz, and A. Loaiza-Bonilla, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2–7, 2025 / Mexico City, Mexico, November 30 – December 5, 2025*, 2025.

[33] N. Karagodin, Y. Polyanskiy, and P. Rigollet. Clustering in causal attention masking. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10–15, 2024*, 2024.

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

[35] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020*, 2020.

[36] J. M. Lee. *Introduction to smooth manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, New York, second edition, 2013.

[37] S. Li, T. J. Maranzatto, J. Peszek, T. Teolis, S. Akkoc, K. Riedl, S. Ulukus, and N. García Trillos. On the diverse dynamical behaviors arising in deep linear transformers. *arXiv preprint arXiv:2607.18584*, 2026.

[38] J. R. Magnus and H. Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, 2019.

[39] A. Massucco, L. Del Grande, M. Carioni, C. Brune, and C.-B. Schönlieb. Multi-headed transformer architectures as time-dependent Wasserstein gradient flows. *arXiv preprint arXiv:2605.18870*, 2026.

[40] M. A. Peletier and A. Shalova. Nonlinear diffusion limit of non-local interactions on a sphere. *Nonlinear Analysis*, 275:114283, 2027.

[41] R. Pinnau, C. Totzeck, O. Tse, and S. Martin. A consensus-based model for global optimization and its mean-field limit. *Math. Models Methods Appl. Sci.*, 27(01):183–204, 2017.

[42] Y. Polyanskiy, P. Rigollet, and A. Yao. Synchronization of mean-field models on the circle. *arXiv preprint arXiv:2507.22857*, 2025.

[43] K. Riedl. Leveraging memory effects and gradient information in consensus-based optimisation: on global convergence in mean-field law. *European J. Appl. Math.*, 35(4):483–514, 2024.

[44] K. Riedl. *Mathematical Foundations of Interacting Multi-Particle Systems for Optimization*. PhD thesis, Technische Universität München, 2024.

[45] P. Rigollet. The mean-field dynamics of transformers. In *International Congress of Mathematicians 2026*, pages 389–404. SIAM, 2026.

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

[47] A. Shalova. *Noisy gradient flows: with applications in machine learning*. PhD thesis, Eindhoven University of Technology, 2025.

[48] A. Shalova and A. Schlichting. Solutions of stationary McKean–Vlasov equation on a high-dimensional sphere and other Riemannian manifolds. *Adv. Nonlinear Anal.*, 15(1):20250141, 2026.

[49] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[51] J. von Oswald, E. Niklasson, E. Randazzo, J. Sacramento, A. Mordvintsev, A. Zhmoginov, and M. Vladymyrov. Transformers learn in-context by gradient descent. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR, 2023.

[52] B. Zhang and R. Sennrich. Root mean square layer normalization. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada*, pages 12360–12371, 2019.

Supplementary material

This supplementary material is organized into the following appendices.

A Proof details for the main result, Theorem 1 **15**

A.1 Quantitative estimate for the Wasserstein distance between the flows ρ_t^β and ρ_t . . 15

A.1.1 A quantitative Laplace principle 15

A.1.2 Stability estimates in Wasserstein space 20

A.2 Convergence of the flow ρ_t to $\Pi_\sharp \rho_0$ 23

A.2.1 Auxiliary result for Lyapunov-type convergence estimates in Wasserstein space 23

A.2.2 Lyapunov-type convergence estimates in Wasserstein space 24

A.2.3 Invariance of Π under the zero-temperature equation 26

A.3 Proof of Corollary 4 28

B Heuristics supporting Conjecture 1 **28**

C Additional numerical experiments **29**

C.1 Details on the discretization 30

C.2 Non-symmetric $VB^\top$ effects 30

C.3 Gradient flow case 31

C.4 Experiments supporting Conjecture 1 33

C.5 Empirical validation of the $\mathcal{O}(\log \beta)$ time scale 33

A Proof details for the main result, Theorem 1

In this appendix, we present auxiliary technical results for our main theoretical result, Theorem 1, together with their proof details.

A.1 Quantitative estimate for the Wasserstein distance between the flows ρ_t^β and ρ_t

The objective of this section is to derive a bound on $W_2(\rho_t^\beta, \rho_t)$ in terms of the temperature parameter β to prove Proposition 6.

Our proof is based on a stability estimate in Wasserstein space, which we detail in Appendix A.1.2. It employs the quantitative Laplace principle from Lemma 5, see also [27], to control the discrepancy between the vector fields of the continuity equations (2) and (3). The proof of the quantitative Laplace principle is provided in Appendix A.1.1.

A.1.1 A quantitative Laplace principle

We first give a proof of the quantitative Laplace principle on the sphere $\mathbb{S}^{d-1}$, Lemma 5, which was originally introduced in [27] on the plane. Here we extend it to the sphere and specialize it to a certain linear function on the sphere.

Proof of Lemma 5. The proof follows the lines of [27, Proposition 4.5], adapted to the sphere $\mathbb{S}^{d-1}$.

To simplify notation, let us define for $x \in \mathbb{S}^{d-1}$ the non-negative functions

$$\mathcal{E}_x(y) \coloneqq \left\langle B^\top x, y^*(x) \right\rangle - \left\langle B^\top x, y \right\rangle = \left\| B^\top x \right\| \left(1 - \langle y^*(x), y \rangle\right), \qquad y \in \mathbb{S}^{d-1},$$

where we recall from (12) that $y^*(x) = \frac{B^\top x}{\|B^\top x\|}$. With this definition, we have the identity $e^{\beta\langle B^\top x, y\rangle} = e^{\beta\|B^\top x\|} e^{-\beta\mathcal{E}_x(y)}$. Using this, we obtain

$$m_{\beta,\varrho}(x) = \frac{\int_{\mathbb{S}^{d-1}} e^{-\beta\mathcal{E}_x(y)} y \,\mathrm{d}\varrho(y)}{\int_{\mathbb{S}^{d-1}} e^{-\beta\mathcal{E}_x(y)} \,\mathrm{d}\varrho(y)},$$

where we canceled $e^{\beta\|B^\top x\|}$ in the numerator and denominator. Using that we are on the sphere $\mathbb{S}^{d-1}$, where $\|y - y^*(x)\|^2 = \|y\|^2 - 2\langle y^*(x), y\rangle + \|y^*(x)\|^2 = 2 - 2\langle y^*(x), y\rangle$, we notice that $\mathcal{E}_x(y) = \frac{\|B^\top x\|}{2}\|y - y^*(x)\|^2$. With this we directly observe that

$$\mathcal{E}_{x,r} \coloneqq \sup_{y\in\mathcal{B}_r(y^*(x))} \mathcal{E}_x(y) \leq \frac{\|B^\top x\|}{2} r^2.$$

The quantity $Z_\beta(x) \coloneqq \int_{\mathbb{S}^{d-1}} e^{-\beta\mathcal{E}_x(y)} \,\mathrm{d}\varrho(y)$ can be lower bounded by

$$Z_\beta(x) \geq \int_{\mathcal{B}_r(y^*(x))} e^{-\beta\mathcal{E}_x(y)} \,\mathrm{d}\varrho(y) \geq \inf_{y\in\mathcal{B}_r(y^*(x))} e^{-\beta\mathcal{E}_x(y)} \varrho(\mathcal{B}_r(y^*(x))) = e^{-\beta\mathcal{E}_{x,r}} \varrho(\mathcal{B}_r(y^*(x))), \tag{15}$$

where the first inequality is due to the integrand being positive. Now let $\tilde{r} \geq r > 0$. We can decompose (13) into contributions from inside and outside of $\mathcal{B}_{\tilde{r}}(y^*(x))$. By Jensen's inequality we have

$$\begin{aligned}&\|m_{\beta,\varrho}(x) - y^*(x)\| \\ &\quad\leq \int_{\mathcal{B}_{\tilde{r}}(y^*(x))} \|y - y^*(x)\| \frac{e^{-\beta\mathcal{E}_x(y)}}{Z_\beta(x)} \,\mathrm{d}\varrho(y) + \int_{\mathcal{B}_{\tilde{r}}(y^*(x))^c} \|y - y^*(x)\| \frac{e^{-\beta\mathcal{E}_x(y)}}{Z_\beta(x)} \,\mathrm{d}\varrho(y).\end{aligned} \tag{16}$$

Inside $\mathcal{B}_{\tilde{r}}(y^*(x))$ we have $\|y - y^*(x)\| \leq \tilde{r}$, which allows to bound the first term in (16) by $\tilde{r}$ since the integrand is nonnegative and $\int_{\mathcal{B}_{\tilde{r}}(y^*(x))} \frac{e^{-\beta\mathcal{E}_x(y)}}{Z_\beta(x)} \,\mathrm{d}\varrho(y) \leq 1$. For the second term in (16), we use the lower bound (15), which gives

$$\begin{aligned}&\int_{\mathcal{B}_{\tilde{r}}(y^*(x))^c} \|y - y^*(x)\| \frac{e^{-\beta\mathcal{E}_x(y)}}{Z_\beta(x)} \,\mathrm{d}\varrho(y) \\ &\qquad\leq \frac{e^{\beta\mathcal{E}_{x,r}}}{\varrho(\mathcal{B}_r(y^*(x)))} \int_{\mathcal{B}_{\tilde{r}}(y^*(x))^c} \|y - y^*(x)\| \, e^{-\beta\mathcal{E}_x(y)} \,\mathrm{d}\varrho(y) \\ &\qquad\leq \frac{2}{\varrho(\mathcal{B}_r(y^*(x)))} \exp\left(-\beta\left(\inf_{u\in\mathcal{B}_{\tilde{r}}(y^*(x))^c} \mathcal{E}_x(u) - \mathcal{E}_{x,r}\right)\right).\end{aligned} \tag{17}$$

To obtain the last inequality, we first note that $\|y - y^*(x)\| \leq 2$ since $y, y^*(x) \in \mathbb{S}^{d-1}$. Secondly, inside $\mathcal{B}_{\tilde{r}}(y^*(x))^c$ we have $\mathcal{E}_x(y) \geq \inf_{u\in\mathcal{B}_{\tilde{r}}(y^*(x))^c} \mathcal{E}_x(u)$, or equivalently after taking the exponential $e^{-\beta\mathcal{E}_x(y)} \leq e^{-\beta\inf_{u\in\mathcal{B}_{\tilde{r}}(y^*(x))^c}\mathcal{E}_x(u)}$. And thirdly, $\varrho\big(\mathcal{B}_{\tilde{r}}(y^*(x))^c\big) \leq 1$. Combining (17) with the observation directly before allows us to conclude (16) with the bound

$$\|m_{\beta,\varrho}(x) - y^*(x)\| \leq \tilde{r} + \frac{2}{\varrho(\mathcal{B}_r(y^*(x)))} \exp\left(-\beta\left(\inf_{u\in\mathcal{B}_{\tilde{r}}(y^*(x))^c} \mathcal{E}_x(u) - \mathcal{E}_{x,r}\right)\right). \tag{18}$$

Now, with the convention that the infimum over the empty set is $+\infty$, since $\inf_{u\in\mathcal{B}_{\tilde{r}}(y^*(x))^c} \mathcal{E}_x(u) = \inf_{u\in\mathcal{B}_{\tilde{r}}(y^*(x))^c} \frac{\|B^\top x\|}{2}\|u - y^*(x)\|^2 \geq \frac{\|B^\top x\|}{2}\tilde{r}^2$, we choose $\tilde{r}$ such that

$$\inf_{u\in\mathcal{B}_{\tilde{r}}(y^*(x))^c} \mathcal{E}_x(u) - \mathcal{E}_{x,r} \geq \frac{\|B^\top x\|}{2}(\tilde{r}^2 - r^2) = q,$$

i.e., $\tilde{r} = \sqrt{\frac{2}{\|B^\top x\|}q + r^2}$. With this, (17) is bounded by $\frac{2e^{-\beta q}}{\varrho(\mathcal{B}_r(y^*(x)))}$ and we obtain for (18) the upper bound

$$\|m_{\beta,\varrho}(x) - y^*(x)\| \leq \tilde{r} + \frac{2e^{-\beta q}}{\varrho(\mathcal{B}_r(y^*(x)))}.$$

Noticing that $\|B^\top x\| \geq \sigma_{\min}(B)$ gives (13), which concludes the proof. □

Before employing Lemma 5, we state and prove a technical result which lower-bounds the mass $\rho_t^\beta\big(\mathcal{B}_r(y^*(x))\big)$, that appears when using (13) in the proof of Proposition 6, appropriately on finite time intervals $[0,T]$. In the proof of this result, a key ingredient, which might be of independent interest, is to control the divergence of the vector field $v_\beta[\rho_t^\beta](x) \coloneqq \mathrm{P}_x\big(Vm_{\beta,\rho_t^\beta}(x)\big)$ of the continuity equation (2) in terms of β, see (20). Note that, at least heuristically, at a point $x_t \in \arg\min_{x\in\mathbb{S}^{d-1}} f_t^\beta(x)$, where f_t^β denotes the density of ρ_t^β with respect to the surface measure, we have $\partial_t f_t^\beta(x_t) = -f_t^\beta(x_t)\operatorname{div} v_\beta[\rho_t^\beta](x_t)$, where we used the product rule and $\nabla f_t^\beta(x_t) = 0$. Hence, if we had a bound on that divergence, we could use Grönwall's lemma to obtain a lower bound for the minimal density $f_t^\beta(x_t)$ and hence for the mass $\rho_t^\beta\big(\mathcal{B}_r(y^*(x))\big)$. Making this formal reasoning rigorous is the content of the following lemma.

Lemma 10. *Assume that the matrix B satisfies A1. Let $(\rho_t^\beta)_{t\in[0,T]}$ denote a weak solution to the continuity equation (2) with initial datum ρ_0, which is assumed to satisfy Assumption 3. Then, for every $\beta \geq 0$, ρ_t^β admits a continuous, strictly positive density f_t^β with respect to the surface measure $\mathcal{H}^{d-1}\llcorner\mathbb{S}^{d-1}$ for every $t\in[0,T]$.*

Furthermore, let $\bar f \coloneqq 1/\mathcal{H}^{d-1}(\mathbb{S}^{d-1})$ denote the uniform density with respect to the surface measure $\mathcal{H}^{d-1}\llcorner\mathbb{S}^{d-1}$, and set $\ell_\beta(t) \coloneqq \min_{x\in\mathbb{S}^{d-1}} f_t^\beta(x)$ as well as $L_\beta(t) \coloneqq \log \bar f/\ell_\beta(t)$. Then, for every $t\in[0,T]$, it holds $L_\beta(t)\geq 0$ and for all $x\in\mathbb{S}^{d-1}$ it holds

$$\rho_t^\beta\big(\mathcal{B}_r(y^*(x))\big) \geq c_d \bar f e^{-L_\beta(t)} r^{d-1} \tag{19}$$

for all $0<r\leq 1$. Here, $c_d>0$ denotes a constant depending only on d with $c_d\bar f \leq 1$.

Moreover, there exists a constant $c>0$ depending only on d, $\sigma_{\min}(B)$, $\sigma_{\max}(B)$, and $\sigma_{\max}(V)$ such that for every $t\in[0,T]$ and $\beta\geq 1$ it holds

$$\big\|\operatorname{div} v_\beta[\rho_t^\beta]\big\|_\infty \leq c\,(1+\log\beta + L_\beta(t)) \tag{20}$$

with the function $L_\beta(t)$ satisfying

$$1+\log\beta+L_\beta(t) \leq \left(1+\log\beta+\log\frac{\bar f}{\ell_0}\right)e^{ct}. \tag{21}$$

Proof. We start by showing that ρ_t^β has a continuous density f_t^β by applying standard arguments for continuity equations.

For $t\in[0,T]$ and $x\in\mathbb{S}^{d-1}$, define the so-called tilted probability measure

$$\mathrm{d}\nu_{t,x}(y) \coloneqq \frac{\exp\big(\beta\,\langle B^\top x, y\rangle\big)}{\int_{\mathbb{S}^{d-1}}\exp\big(\beta\,\langle B^\top x, z\rangle\big)\,\mathrm{d}\rho_t^\beta(z)}\,\mathrm{d}\rho_t^\beta(y), \tag{22}$$

where we recall the notation m_{β,ρ_t^β} defined in (11), and its covariance matrix

$$\Sigma_t(x) \coloneqq \int_{\mathbb{S}^{d-1}}\big(y-m_{\beta,\rho_t^\beta}(x)\big)\big(y-m_{\beta,\rho_t^\beta}(x)\big)^\top\,\mathrm{d}\nu_{t,x}(y).$$

Since $\int_{\mathbb{S}^{d-1}}\exp\big(\beta\,\langle B^\top x,z\rangle\big)\,\mathrm{d}\rho_t^\beta(z) \geq e^{-\beta\sigma_{\max}(B)} > 0$ and $\|m_{\beta,\rho_t^\beta}(x)\|\leq 1$, differentiation under the integral and the quotient rule give the following formula for the manifold Jacobian of the vector field $v_\beta[\rho_t^\beta]$:

$$Dv_\beta[\rho_t^\beta](x) = \beta \mathrm{P}_x V\Sigma_t(x)B^\top\mathrm{P}_x - \langle x, Vm_{\beta,\rho_t^\beta}(x)\rangle \mathrm{P}_x. \tag{23}$$

Here, we used standard calculus for tangential differentiation of tangential vector fields on the sphere [12]. Moreover,

$$0\leq\|\Sigma_t(x)\|_{\mathrm{op}} \leq \operatorname{tr}\Sigma_t(x) = 1-\big\|m_{\beta,\rho_t^\beta}(x)\big\|^2 \leq 1. \tag{24}$$

The preceding identities (23) and (24), together with weak continuity of $t\mapsto\rho_t^\beta$ (see [30, Remark 3.1]) and compactness of $\mathbb{S}^{d-1}$, imply

$$t\mapsto v_\beta[\rho_t^\beta] \in C\big([0,T];C^1(\mathbb{S}^{d-1};\mathbb{R}^d)\big).$$

Let $\Phi_t : \mathbb{S}^{d-1} \to \mathbb{S}^{d-1}$ be the associated characteristic flow, defined through

$$\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t(x) = v_\beta[\rho_t^\beta]\big(\Phi_t(x)\big), \qquad \Phi_0(x) = x, \qquad x \in \mathbb{S}^{d-1},\ t > 0.$$

Since the velocity $v_\beta[\rho_t^\beta](x) \in T_x\mathbb{S}^{d-1}$ lies in the tangent space of the sphere at $x \in \mathbb{S}^{d-1}$ and the sphere is compact, standard ODE theory [36] provides the existence of a global flow of diffeomorphisms Φ_t of $\mathbb{S}^{d-1}$. We next justify that ρ_t^β can be written as pushforward of the initial datum under this flow. To see this, note that for every $\varphi \in C^1([0,T] \times \mathbb{S}^{d-1})$ the chain rule gives

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{S}^{d-1}} \varphi(t,x)\ \mathrm{d}\big((\Phi_t)_\sharp\rho_0\big)(x) &= \frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{S}^{d-1}} \varphi(t,\Phi_t(y))\,\mathrm{d}\rho_0(y)\\
&= \int_{\mathbb{S}^{d-1}} \partial_t\varphi(t,\Phi_t(y)) + \Big\langle \nabla\varphi(t,\Phi_t(y)), v_\beta[\rho_t^\beta](\Phi_t(y))\Big\rangle\,\mathrm{d}\rho_0(y)\\
&= \int_{\mathbb{S}^{d-1}} \partial_t\varphi(t,x) + \Big\langle \nabla\varphi(t,x), v_\beta[\rho_t^\beta](x)\Big\rangle\,\mathrm{d}\big((\Phi_t)_\sharp\rho_0\big)(x).
\end{aligned}$$

Hence $(\Phi_t)_\sharp\rho_0$ is a weak solution of the linear continuity equation with prescribed velocity $v_\beta[\rho_t^\beta]$ and initial datum ρ_0. Also ρ_t^β solves the same equation. Since the prescribed velocity is in particular Lipschitz in space, uniqueness for this linear continuity equation yields $\rho_t^\beta = (\Phi_t)_\sharp\rho_0$. Since ρ_0 is assumed to have a continuous density f_0, the same holds for ρ_t^β for any $t \in (0,T]$ thanks to this pushforward formula with the diffeomorphism Φ_t. We call this density f_t^β.

Our next goal is to show the lower bound for the mass. Note that, because f_t^β is a probability density, we have

$$1 = \int_{\mathbb{S}^{d-1}} \mathrm{d}\rho_t^\beta = \int_{\mathbb{S}^{d-1}} f_t^\beta\,\mathrm{d}\mathcal{H}^{d-1} \geq \ell_\beta(t)\mathcal{H}^{d-1}(\mathbb{S}^{d-1}) = \frac{\ell_\beta(t)}{\bar f}.$$

Hence, $\ell_\beta(t) \leq \bar f$ and $L_\beta(t) = \log \bar f/\ell_\beta(t) \geq 0$. Here, we use the convention that $L_\beta(t) = \infty$ if $\ell_\beta(t) = 0$, however, as we show later in (29), $\ell_\beta(t) > 0$ for all $t > 0$.

Standard arguments show the existence of a constant $c_d > 0$ such that

$$\mathcal{H}^{d-1}(\mathcal{B}_r(x)) \geq c_d r^{d-1}, \qquad 0 < r \leq 1, \quad x \in \mathbb{S}^{d-1}.$$

Decreasing c_d if necessary, we can assume that $c_d\bar f \leq 1$. Hence, for all $0 < r \leq 1$ we get

$$\rho_t^\beta\left(\mathcal{B}_r(y^*(x))\right) = \int_{\mathcal{B}_r(y^*(x))} f_t^\beta\,\mathrm{d}\mathcal{H}^{d-1} \geq c_d\ell_\beta(t)r^{d-1} = c_d\bar f e^{-L_\beta(t)}r^{d-1}, \tag{25}$$

which shows (19).

The hard part of the proof consists in showing the estimate (21) for the function $L_\beta(t)$ appearing in (19). Upper-bounding $L_\beta(t) = \log \bar f/\ell_\beta(t)$ amounts to lower-bounding $\ell_\beta(t) := \min_{\mathbb{S}^{d-1}} f_t^\beta$. To do so, we will use the pushforward formula $\rho_t^\beta = (\Phi_t)_\sharp\rho_0$. By the change-of-variables formula the densities of ρ_0 and ρ_t^β satisfy $f_t^\beta(\Phi_t(x))J_t(x) = f_0(x)$ or equivalently

$$f_t^\beta(\Phi_t(x)) = \frac{f_0(x)}{J_t(x)}, \qquad x \in \mathbb{S}^{d-1}, \tag{26}$$

where $J_t(x) := \det D\Phi_t(x) \neq 0$ denotes the manifold Jacobian determinant of the diffeomorphism Φ_t at x. By (26) and since f_0 is strictly positive on $\mathbb{S}^{d-1}$, lower-bounding the minimum of f_t^β amounts to finding a good upper-bound for the Jacobian determinant $J_t(x)$.

For this, we apply Jacobi's determinant formula for a differentiable curve of matrices $t \mapsto A(t)$ which reads $\frac{\mathrm{d}}{\mathrm{d}t}\det A(t) = \det A(t)\,\mathrm{tr}\left(A(t)^{-1}\frac{\mathrm{d}}{\mathrm{d}t}A(t)\right)$, see [38, Section 8.3]. With the choice $A(t) := D\Phi_t(x)$, we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}J_t(x) &= J_t(x)\,\mathrm{tr}\left(D\Phi_t(x)^{-1}\frac{\mathrm{d}}{\mathrm{d}t}D\Phi_t(x)\right)\\
&= J_t(x)\,\mathrm{tr}\left(Dv_\beta[\rho_t^\beta](\Phi_t(x))\right) = J_t(x)\,\mathrm{div}\,v_\beta[\rho_t^\beta](\Phi_t(x)),
\end{aligned} \tag{27}$$

where we applied the Jacobian to both sides of the ODE $\frac{\mathrm{d}}{\mathrm{d}t}\Phi_t(x) = v_\beta[\rho_t^\beta](\Phi_t(x))$, leading to $\frac{\mathrm{d}}{\mathrm{d}t}D\Phi_t(x) = Dv_\beta[\rho_t^\beta](\Phi_t(x))\, D\Phi_t(x)$ by chain rule, together with the cyclicity of the trace to cancel $D\Phi_t(x)^{-1}$ to obtain the second equality in (27). Solving the ODE (27) explicitly, using also $J_0(x) = 1$, we obtain

$$J_t(x) = \exp\left(\int_0^t \operatorname{div} v_\beta[\rho_s^\beta](\Phi_s(x))\,\mathrm{d}s\right), \qquad x \in \mathbb{S}^{d-1}. \tag{28}$$

Combining (26) and (28) we arrive at

$$f_t^\beta(\Phi_t(x)) = f_0(x)\exp\left(-\int_0^t \operatorname{div} v_\beta[\rho_s^\beta](\Phi_s(x))\,\mathrm{d}s\right), \qquad x \in \mathbb{S}^{d-1}.$$

Taking the minimum over $x \in \mathbb{S}^{d-1}$ on both sides shows

$$\ell_\beta(t) \geq \ell_0 \exp\left(-\int_0^t \left\|\operatorname{div} v_\beta[\rho_s^\beta]\right\|_\infty \mathrm{d}s\right) \tag{29}$$

and it therefore remains to obtain an estimate for $\|\operatorname{div} v_\beta[\rho_t^\beta]\|_\infty$ that is sufficiently well-behaved in terms of β. We first compute the divergence by taking the trace of the manifold Jacobian from (23) and obtain

$$\begin{aligned}\operatorname{div} v_\beta[\rho_t^\beta](x) &= \beta \operatorname{tr}\left(\mathrm{P}_x V \Sigma_t(x) B^\top \mathrm{P}_x\right) - \langle x, V m_{\beta,\rho_t^\beta}(x)\rangle \operatorname{tr}\mathrm{P}_x \\ &= \beta \operatorname{tr}\left(V\Sigma_t(x)B^\top \mathrm{P}_x\right) - (d-1)\langle x, V m_{\beta,\rho_t^\beta}(x)\rangle,\end{aligned} \tag{30}$$

where we used cyclicity of the trace, $\mathrm{P}_x^2 = \mathrm{P}_x$, and $\operatorname{tr}\mathrm{P}_x = d-1$. Consequently, taking absolute values in (30) and estimating the traces on the right hand side using $\|m_{\beta,\rho_t^\beta}(x)\| \leq 1$, we obtain

$$\left|\operatorname{div} v_\beta[\rho_t^\beta](x)\right| \leq \sigma_{\max}(V)\sigma_{\max}(B)\beta \operatorname{tr}\Sigma_t(x) + (d-1)\sigma_{\max}(V). \tag{31}$$

A naive estimate of this quantity would be to bound $\operatorname{tr}\Sigma_t(x)$ by a constant which would result in a bound for the divergence of order $\mathcal{O}(1+\beta)$, which is not sufficient for our purposes and which is why we continue with a more careful analysis.

Recalling that $y^*(x) = B^\top x/\|B^\top x\|$, the tilted measure from (22) can equivalently be written as

$$\mathrm{d}\nu_{t,x}(y) = \frac{e^{-\frac{\beta\|B^\top x\|}{2}\|y-y^*(x)\|^2}}{\int_{\mathbb{S}^{d-1}} e^{-\frac{\beta\|B^\top x\|}{2}\|z-y^*(x)\|^2}\,\mathrm{d}\rho_t^\beta(z)}\,\mathrm{d}\rho_t^\beta(y). \tag{32}$$

Note that in Section 3 we used exactly this rewriting to connect self-attention transformers to the polarized CBO method [16]. A simple computation gives the bias-variance decomposition

$$\int_{\mathbb{S}^{d-1}} \|y-y^*(x)\|^2\,\mathrm{d}\nu_{t,x}(y) = \operatorname{tr}\Sigma_t(x) + \left\|m_{\beta,\rho_t^\beta}(x) - y^*(x)\right\|^2. \tag{33}$$

Hence, the integral on the left — which we can think of as a mean-squared error — controls the term $\operatorname{tr}\Sigma_t(x)$ in (31) from above and we will continue by estimating this integral which, by using (32), takes the form

$$\int_{\mathbb{S}^{d-1}} \|y-y^*(x)\|^2\,\mathrm{d}\nu_{t,x}(y) = \frac{\int_{\mathbb{S}^{d-1}} \|y-y^*(x)\|^2 e^{-\frac{\beta\|B^\top x\|}{2}\|y-y^*(x)\|^2}\,\mathrm{d}\rho_t^\beta(y)}{\int_{\mathbb{S}^{d-1}} e^{-\frac{\beta\|B^\top x\|}{2}\|z-y^*(x)\|^2}\,\mathrm{d}\rho_t^\beta(z)}. \tag{34}$$

Fix $0 < r \leq a$. Lower-bounding the denominator on $\mathcal{B}_r(y^*(x))$, and splitting the numerator into integrals over the set $\{y \in \mathbb{S}^{d-1} : \|y-y^*(x)\| \leq a\}$ and its complement, we get from (33) and (34) that

$$\operatorname{tr}\Sigma_t(x) \leq a^2 + \frac{4e^{-\frac{\beta\|B^\top x\|}{2}(a^2-r^2)}}{\rho_t^\beta(\mathcal{B}_r(y^*(x)))}. \tag{35}$$

Note that this inequality also holds when $a > 2$, in which case the complementary region is empty.

Next we choose suitable parameters as follows. Set $r = \beta^{-1/2}$ and define

$$a^2 \coloneqq r^2 + \frac{2}{\beta \|B^\top x\|} \log \frac{4\beta}{\rho_t^\beta(\mathcal{B}_r(y^*(x)))}.$$

Since $\beta \geq 1 > \frac{1}{4}$, the logarithm is positive. With this choice, the second term in (35) equals β^{-1}, and therefore

$$\beta \operatorname{tr} \Sigma_t(x) \leq 2 + \frac{2}{\sigma_{\min}(B)} \log \frac{4\beta}{\rho_t^\beta(\mathcal{B}_{\beta^{-1/2}}(y^*(x)))}. \tag{36}$$

Using the lower mass bound (19), which we proved above, in (36) yields

$$\begin{aligned}\beta \operatorname{tr} \Sigma_t(x) &\leq 2 + \frac{2}{\sigma_{\min}(B)} \left(L_\beta(t) + \frac{d+1}{2} \log \beta + \log \frac{4}{c_d \bar{f}} \right) \\ &\leq C\big(1 + \log \beta + L_\beta(t)\big)\end{aligned}$$

for some constant $C = C(d, \sigma_{\min}(B))$. Inserting this estimate into (31) shows

$$\begin{aligned}\big\|\operatorname{div} v_\beta[\rho_t^\beta]\big\|_\infty &\leq \sigma_{\max}(V)\sigma_{\max}(B) C\big(1 + \log \beta + L_\beta(t)\big) + (d-1)\sigma_{\max}(V) \\ &\leq c\big(1 + \log \beta + L_\beta(t)\big)\end{aligned}$$

for some constant $c > 0$ depending only on $d, \sigma_{\min}(B), \sigma_{\max}(B)$, and $\sigma_{\max}(V)$. This proves (20).

Taking logarithms in (29) and using (20), we obtain

$$\log \ell_\beta(t) \geq \log \ell_0 - c \int_0^t (1 + \log \beta + L_\beta(s)) \, \mathrm{d}s.$$

Using the definition of $L_\beta(t) = \log \bar{f}/\ell_\beta(t)$ and $\ell_\beta(0) = \ell_0$, this can be rearranged to

$$1 + \log \beta + L_\beta(t) \leq 1 + \log \beta + L_\beta(0) + c \int_0^t \big(1 + \log \beta + L_\beta(s)\big) \, \mathrm{d}s.$$

Remember that $(t, x) \mapsto f_t^\beta(x)$ is continuous. Compactness of $\mathbb{S}^{d-1}$ then implies that $t \mapsto \ell_\beta(t)$ and hence $t \mapsto L_\beta(t)$ is continuous. Grönwall's inequality therefore gives

$$1 + \log \beta + L_\beta(t) \leq (1 + \log \beta + L_\beta(0)) \, e^{ct},$$

which proves (21). □

A.1.2 Stability estimates in Wasserstein space

We now have available all necessary technical tools to provide a proof of Proposition 6.

Proof of Proposition 6. Using m_{β,ρ_t^β} and y^* as defined in (11) and (12), respectively, denote the vector fields in (2) and (3) by

$$v_\beta[\rho_t^\beta](x) \coloneqq \mathrm{P}_x\big(V m_{\beta,\rho_t^\beta}(x)\big), \qquad v_\infty(x) \coloneqq \mathrm{P}_x\big(V y^*(x)\big).$$

Also denoting by $\pi_t \in \Pi(\rho_t^\beta, \rho_t)$ the optimal coupling between ρ_t^β and ρ_t for the Wasserstein distance, we differentiate the squared Wasserstein distance along the solutions to the associated continuity equations [6, Theorem 8.4.7] to obtain for almost every $t \in [0, T]$ that

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} W_2^2(\rho_t^\beta, \rho_t) &= \frac{\mathrm{d}}{\mathrm{d}t} \iint_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \|x - z\|^2 \, \mathrm{d}\pi_t(x, z) \\ &= 2 \iint_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \big\langle x - z, v_\beta[\rho_t^\beta](x) - v_\infty(z) \big\rangle \, \mathrm{d}\pi_t(x, z) \\ &\leq 2 W_2(\rho_t^\beta, \rho_t) \sqrt{\iint_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \big\| v_\beta[\rho_t^\beta](x) - v_\infty(z) \big\|^2 \, \mathrm{d}\pi_t(x, z)},\end{aligned}$$

where we used the Cauchy–Schwarz inequality to obtain the last step. By the chain rule it holds that $\frac{\mathrm{d}}{\mathrm{d}t}W_2^2(\rho_t^\beta,\rho_t) = 2W_2(\rho_t^\beta,\rho_t)\frac{\mathrm{d}}{\mathrm{d}t}W_2(\rho_t^\beta,\rho_t)$. If $W_2(\rho_t^\beta,\rho_t)=0$ for some $t>0$ we have nothing to show. Hence, we can assume that for $t>0$ this distance is positive and after dividing both sides by $2W_2(\rho_t^\beta,\rho_t)$ we get

$$\frac{\mathrm{d}}{\mathrm{d}t}W_2(\rho_t^\beta,\rho_t) \leq \sqrt{\iint_{\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}} \left\|v_\beta[\rho_t^\beta](x) - v_\infty(z)\right\|^2 \mathrm{d}\pi_t(x,z)}. \tag{37}$$

Let us now estimate the integrand in (37). For this, we decompose this term into an approximation error due to the finite inverse temperature β, and a stability term. More precisely, with the triangle inequality we have for all $x,z\in\mathbb{S}^{d-1}$ that

$$\left\|v_\beta[\rho_t^\beta](x) - v_\infty(z)\right\| \leq \left\|v_\beta[\rho_t^\beta](x) - v_\infty(x)\right\| + \left\|v_\infty(x) - v_\infty(z)\right\|. \tag{38}$$

To estimate the approximation error, i.e., the first term on the right-hand side of (38), we note after recalling $\|\mathrm{P}_x\|\leq 1$, that

$$\begin{aligned}\left\|v_\beta[\rho_t^\beta](x) - v_\infty(x)\right\| &= \left\|\mathrm{P}_x\big(Vm_{\beta,\rho_t^\beta}(x)\big) - \mathrm{P}_x\big(Vy^*(x)\big)\right\| \\ &= \left\|\mathrm{P}_x\left(V\big(m_{\beta,\rho_t^\beta}(x) - y^*(x)\big)\right)\right\| \\ &\leq \sigma_{\max}(V)\left\|m_{\beta,\rho_t^\beta}(x) - y^*(x)\right\|. \end{aligned}\tag{39}$$

We can now employ Lemma 5 to estimate for arbitrary $r>0$ and $q>0$ that

$$\left\|m_{\beta,\rho_t^\beta}(x) - y^*(x)\right\| \leq \sqrt{r^2 + \frac{2q}{\sigma_{\min}(B)}} + \frac{2e^{-\beta q}}{\rho_t^\beta\big(\mathcal{B}_r(y^*(x))\big)}. \tag{40}$$

Let us now first consider the case $\beta\geq 1$ and set $r:=\beta^{-1/2}\leq 1$. By (19) in Lemma 10 we have

$$\rho_t^\beta\big(\mathcal{B}_r(y^*(x))\big) \geq c_d\bar{f}e^{-L_\beta(t)}\beta^{-(d-1)/2}$$

for a constant $c_d>0$ depending only on the dimension and such that $c_d\bar{f}\leq 1$. Since $L_\beta(t)\geq 0$ by Lemma 10, we can choose

$$q := \frac{1}{\beta}\left(L_\beta(t) + \frac{d+1}{2}\log\beta + \log\frac{2}{c_d\bar{f}}\right) > 0$$

in (40), which leads to the following upper bounds for the summands on the right-hand side of (40)

$$\frac{2e^{-\beta q}}{\rho_t^\beta\big(\mathcal{B}_r(y^*(x))\big)} \leq \frac{1}{\beta} \quad \text{and} \quad \sqrt{r^2 + \frac{2q}{\sigma_{\min}(B)}} \leq C\sqrt{\frac{1+\log\beta + L_\beta(t)}{\beta}},$$

for a constant $C = C(d,\sigma_{\min}(B))$. Therefore, using that $\beta\geq 1$ and $L_\beta(t)\geq 0$

$$\left\|m_{\beta,\rho_t^\beta}(x) - y^*(x)\right\| \leq C\sqrt{\frac{1+\log\beta+L_\beta(t)}{\beta}} \tag{41}$$

after possibly increasing C. Leveraging (21) from Lemma 10 in the first step and $\beta\geq 1$ in the second, we have

$$1+\log\beta+L_\beta(t) \leq \left(1+\log\beta+\log\frac{\bar{f}}{\ell_0}\right)e^{ct} \leq 2\left(1+\log\frac{\bar{f}}{\ell_0}\right)\log(\beta+1)e^{ct}$$

for a constant $c = c\,(d,\sigma_{\min}(B),\sigma_{\max}(B),\sigma_{\max}(V))\geq 0$. Inserting this into (41), we conclude from (39) in the case $\beta\geq 1$ that

$$\sup_{x\in\mathbb{S}^{d-1}}\left\|v_\beta[\rho_t^\beta](x) - v_\infty(x)\right\| \leq C\sqrt{\frac{\log(\beta+1)}{\beta}}e^{ct/2}$$

for a constant $C = C(d,\sigma_{\min}(B),\sigma_{\max}(V),\ell_0)$. For the remaining case $0<\beta<1$, we use (39) to obtain

$$\sup_{x\in\mathbb{S}^{d-1}}\left\|v_\beta[\rho_t^\beta](x) - v_\infty(x)\right\| \leq \sigma_{\max}(V)\sup_{x\in\mathbb{S}^{d-1}}\left\|m_{\beta,\rho_t^\beta}(x) - y^*(x)\right\| \leq 2\sigma_{\max}(V).$$

Since $\log(\beta+1)/\beta \geq \log 2$, we can combine both cases to eventually obtain the following estimate, where $C = C(d, \sigma_{\min}(B), \sigma_{\max}(V), \ell_0)$ is a constant and we denote $C_0 \coloneqq 1 + \frac{c}{2}$:

$$\sup_{x \in \mathbb{S}^{d-1}} \left\| v_\beta[\rho_t^\beta](x) - v_\infty(x) \right\| \leq C \sqrt{\frac{\log(\beta+1)}{\beta}} e^{C_0 t} \tag{42}$$

for every $\beta > 0$. For the stability term, i.e., the second term in (38), we first notice that for all $x, z \in \mathbb{S}^{d-1}$ it holds by the triangle inequality and reverse triangle inequality that

$$\begin{aligned}
\|y^*(x) - y^*(z)\| &= \left\| \frac{B^\top x}{\|B^\top x\|} - \frac{B^\top z}{\|B^\top z\|} \right\| \leq \frac{\|B^\top x - B^\top z\|}{\|B^\top x\|} + \|B^\top z\| \left\| \frac{1}{\|B^\top x\|} - \frac{1}{\|B^\top z\|} \right\| \\
&= \frac{\|B^\top x - B^\top z\|}{\|B^\top x\|} + \left\| \frac{\|B^\top z\| - \|B^\top x\|}{\|B^\top x\|} \right\| \leq 2 \frac{\|B^\top x - B^\top z\|}{\|B^\top x\|} \\
&\leq 2 \frac{\sigma_{\max}(B)}{\sigma_{\min}(B)} \|x - z\|,
\end{aligned}$$

where we have used that B is invertible and hence $\|B^\top x\| \geq \sigma_{\min}(B)$ for all $x \in \mathbb{S}^{d-1}$. Moreover, one has for all $x, z \in \mathbb{S}^{d-1}$ that

$$\|\mathrm{P}_x - \mathrm{P}_z\| = \|xx^\top - zz^\top\| \leq \|x(x-z)^\top\| + \|(x-z)z^\top\| \leq 2\|x - z\|.$$

Using these two auxiliary estimates, and recalling that $\|\mathrm{P}_x\| \leq 1$, $\|y^*(x)\| = 1$, we obtain by the triangle inequality

$$\begin{aligned}
\left\| v_\infty(x) - v_\infty(z) \right\| &= \left\| \mathrm{P}_x \left(V y^*(x) \right) - \mathrm{P}_z \left(V y^*(z) \right) \right\| \\
&\leq \left\| \mathrm{P}_x \big(V(y^*(x) - y^*(z)) \big) \right\| + \left\| (\mathrm{P}_x - \mathrm{P}_z)(V y^*(z)) \right\| \\
&\leq \left\| \mathrm{P}_x \right\| \left\| V(y^*(x) - y^*(z)) \right\| + \left\| \mathrm{P}_x - \mathrm{P}_z \right\| \left\| V y^*(z) \right\| \\
&\leq 2\sigma_{\max}(V) \frac{\sigma_{\max}(B)}{\sigma_{\min}(B)} \|x - z\| + 2\sigma_{\max}(V) \|x - z\| \\
&= 2\sigma_{\max}(V) \left(1 + \frac{\sigma_{\max}(B)}{\sigma_{\min}(B)} \right) \|x - z\|.
\end{aligned} \tag{43}$$

Inserting (42) and (43) into (38), we obtain

$$\begin{aligned}
&\left\| v_\beta[\rho_t^\beta](x) - v_\infty(z) \right\| \\
&\qquad \leq C(d, \sigma_{\min}(B), \sigma_{\max}(B), \sigma_{\max}(V), \ell_0) \left(\sqrt{\frac{\log(\beta+1)}{\beta}} e^{C_0 t} + \|x - z\| \right).
\end{aligned}$$

Exploiting this estimate in (37) and abbreviating $C_1 \coloneqq C(d, \sigma_{\min}(B), \sigma_{\max}(B), \sigma_{\max}(V), \ell_0) > 0$ as well as $\widetilde{C}_1 \coloneqq 2C_0 + C_1$, we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} W_2(\rho_t^\beta, \rho_t) &\leq C_1 \left(W_2(\rho_t^\beta, \rho_t) + \sqrt{\frac{\log(\beta+1)}{\beta}} e^{C_0 t} \right) \\
&\leq \widetilde{C}_1 \left(W_2(\rho_t^\beta, \rho_t) + \sqrt{\frac{\log(\beta+1)}{\beta}} e^{C_0 t} \right),
\end{aligned} \tag{44}$$

where the last bound is trivial by the definition of $\widetilde{C}_1$, and we also recall that—as π_t denotes the optimal coupling between ρ_t^β and ρ_t—we have $W_2^2(\rho_t^\beta, \rho_t) = \iint_{\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}} \|x - z\|^2 \, \mathrm{d}\pi_t(x, z)$. Using the integrating factor method with integrating factor $e^{-\widetilde{C}_1 t}$, we multiply both sides of (44) by the integrating factor to observe that, after a little bit of algebra, (44) is equivalent to

$$\frac{\mathrm{d}}{\mathrm{d}t} \left(e^{-\widetilde{C}_1 t} W_2(\rho_t^\beta, \rho_t) \right) \leq \widetilde{C}_1 \sqrt{\frac{\log(\beta+1)}{\beta}} e^{(C_0 - \widetilde{C}_1)t}.$$

Since $\widetilde{C}_1 > C_0$, we can integrate this inequality in time to obtain that

$$e^{-\widetilde{C}_1 t} W_2(\rho_t^\beta, \rho_t) \leq \widetilde{C}_1 \sqrt{\frac{\log(\beta+1)}{\beta}} \int_0^t e^{(C_0-\widetilde{C}_1)s} \, \mathrm{d}s$$
$$= \frac{\widetilde{C}_1}{\widetilde{C}_1 - C_0} \sqrt{\frac{\log(\beta+1)}{\beta}} \left(1 - e^{(C_0-\widetilde{C}_1)t}\right),$$

where we used that $\rho_0^\beta = \rho_0$. Dividing now both sides by the integrating factor and using the definition of $\widetilde{C}_1$, we are left with

$$W_2(\rho_t^\beta, \rho_t) \leq \frac{\widetilde{C}_1}{\widetilde{C}_1 - C_0} \sqrt{\frac{\log(\beta+1)}{\beta}} \left(e^{\widetilde{C}_1 t} - e^{C_0 t}\right)$$
$$= \frac{2C_0 + C_1}{C_0 + C_1} \sqrt{\frac{\log(\beta+1)}{\beta}} \left(e^{(2C_0+C_1)t} - e^{C_0 t}\right)$$
$$\leq 2\sqrt{\frac{\log(\beta+1)}{\beta}} \left(e^{(2C_0+C_1)t} - e^{C_0 t}\right)$$

for all $t \in [0, T]$, which concludes the proof, upon renaming the constants. □

A.2 Convergence of the flow ρ_t to $\Pi_\sharp \rho_0$

The aim of this section is to estimate the convergence of the flow ρ_t to $\Pi_\sharp \rho_0$, i.e., to estimate $W_2(\rho_t, \Pi_\sharp \rho_0)$ by proving Proposition 8 in combination with Lemmas 7 and 9.

As a first step, we derive in Proposition 8 a time-evolution inequality for the functional $\mathcal{V}_p$ defined in (14), as we detail in Appendix A.2.2. Using $\Pi_\sharp \rho_t = \Pi_\sharp \rho_0$, as demonstrated by Lemma 9, it holds $W_2(\rho_t, \Pi_\sharp \rho_0) = W_2(\rho_t, \Pi_\sharp \rho_t)$, we verify in Lemma 7 that the functional $\mathcal{V}_p$ is indeed a suitable Lyapunov functional for the flow ρ_t since $W_2(\rho_t, \Pi_\sharp \rho_0) = W_2(\rho_t, \Pi_\sharp \rho_t) \leq \sqrt{2\mathcal{V}_p(\rho_t)}$, where $\mathcal{V}_p(\rho_t)$ decays exponentially fast thanks to Proposition 8.

A.2.1 Auxiliary result for Lyapunov-type convergence estimates in Wasserstein space

We first show that R_p as defined in (14) can be regarded as a distance from E (despite not formally being a distance).

Proof of Lemma 7. Fix $x \in \mathbb{S}^{d-1} \setminus E^\perp$. To simplify notation, let us use the orthogonal decomposition of x,

$$x = u + v, \qquad \text{with} \quad u = \mathrm{P}_E x \in E, \quad v = (\mathrm{Id} - \mathrm{P}_E)x \in E^\perp, \quad \|u\|^2 + \|v\|^2 = 1.$$

We note that $u \neq 0$ since $x \in \mathbb{S}^{d-1} \setminus E^\perp$. Using that $\Pi(x) = \frac{u}{\|u\|}$, a simple computation shows

$$\|x - \Pi(x)\|^2 = 2\,(1 - \|u\|)\,.$$

Denoting

$$q(x) \coloneqq \frac{\|(\mathrm{Id} - \mathrm{P}_E)x\|^2}{\|\mathrm{P}_E x\|^2} = \frac{\|v\|^2}{\|u\|^2},$$

and observing that $1 + q(x) = 1 + \frac{\|v\|^2}{\|u\|^2} = \frac{\|u\|^2 + \|v\|^2}{\|u\|^2} = \frac{1}{\|u\|^2}$ and hence $\|u\| = \frac{1}{\sqrt{1+q(x)}}$, we have

$$\|x - \Pi(x)\|^2 = 2\left(1 - \frac{1}{\sqrt{1+q(x)}}\right).$$

Consider now the scalar function

$$f(q) \coloneqq 1 - \frac{1}{\sqrt{1+q}}, \qquad q \geq 0.$$

First, $f(q) \leq 1$ for all $q \geq 0$. Second, we claim that $f(q) \leq q$ for all $q \geq 0$. To verify this, define $g(q) \coloneqq q - \left(1 - \frac{1}{\sqrt{1+q}}\right)$. It holds $g(0) = 0$ and $g'(q) = 1 - \frac{1}{2}(1+q)^{-3/2} \geq \frac{1}{2} > 0$ for all $q \geq 0$, showing that $g(q) \geq 0$, or equivalently $f(q) \leq q$. Hence $f(q) \leq \min\{1, q\}$. Since $p \in (0, 1]$, we have for all $q \geq 0$ that $\min\{1, q\} \leq q^p$. Indeed, if $0 \leq q \leq 1$, then $q \leq q^p$, while if $q \geq 1$, then $1 \leq q^p$.

Combining the above bounds shows $1 - \frac{1}{\sqrt{1+q(x)}} \leq q(x)^p$ and therefore

$$\|x - \Pi(x)\|^2 = 2\left(1 - \frac{1}{\sqrt{1+q(x)}}\right) \leq 2q(x)^p = 2R_p(x),$$

validating the first part of the claim.

To prove the second part of the statement, define the coupling $\pi \coloneqq (\mathrm{Id}, \Pi)_\sharp \varrho$ between ϱ and $\Pi_\sharp \varrho$. Then, by definition of the Wasserstein distance,

$$\begin{aligned} W_2^2(\varrho, \Pi_\sharp \varrho) &\leq \iint_{\mathbb{S}^{d-1} \times \mathbb{S}^{d-1}} \|x - y\|^2 \, \mathrm{d}\pi(x, y) \\ &= \int_{\mathbb{S}^{d-1}} \|x - \Pi(x)\|^2 \, \mathrm{d}\varrho(x) \leq 2 \int_{\mathbb{S}^{d-1}} R_p(x) \, \mathrm{d}\varrho(x), \end{aligned}$$

which concludes the proof. □

A.2.2 Lyapunov-type convergence estimates in Wasserstein space

We now have available all the necessary technical tools to prove Proposition 8.

Proof of Proposition 8. Let us denote by ρ_t the solution to the continuity equation (3) in the zero-temperature limit.

We first deal with the trivial case of $k = \dim E = d$ in which case $VB^\top$ equals a multiple of the identity matrix and hence the continuity equation (3) simplifies to $\partial_t \rho_t = 0$ (using that $\mathrm{P}_x(x) = 0$). Hence, trivially, $\rho_t = \rho_0$ for all $t > 0$ and $\Pi = \mathrm{Id}$ such that $\mathcal{V}_p(\rho_t) = 0$ and also $W_2(\rho_t, \Pi_\sharp \rho_t) = 0$ for all $t \geq 0$.

Ideally, we would like to use R_p as defined in (14) as test function for the weak formulation of (3). However, since R_p is infinite on $E^\perp$ and not Lipschitz continuous for $p < \frac{1}{2}$, this is not possible. We therefore smoothly approximate the p-th power as $\psi_{p,\delta}(t) \coloneqq (t + \delta)^p - \delta^p$ for $\delta > 0$ and choose a smooth cutoff function $\xi_K : \mathbb{R} \to [0, 1]$ for $K \in \mathbb{N}$ which satisfies $\xi_K(t) = 1$ for $0 \leq t \leq K$ and $\xi_K(t) = 0$ for $2K \leq t < \infty$. With these two ingredients, we define the smooth and truncated approximation of the p-th power as $\Psi_{p,\delta,K}(t) \coloneqq \int_0^t \psi'_{p,\delta}(s)\xi_K(s) \, \mathrm{d}s$ with $\Psi_{p,\delta,K}(\infty) \coloneqq \Psi_{p,\delta,K}(2K)$. Denoting $R(x) \coloneqq R_1(x) = {\|x - P_E x\|^2}/{\|P_E x\|^2}$, we can define the smooth test function

$$R_{p,\delta,K}(x) \coloneqq \Psi_{p,\delta,K}\left(R(x)\right) = \Psi_{p,\delta,K}\left(\frac{\|x - P_E x\|^2}{\|P_E x\|^2}\right),$$

which is smooth on $\mathbb{S}^{d-1}$, as it is constant in a neighborhood of $\mathbb{S}^{d-1} \cap E^\perp$. For the corresponding approximation of our Lyapunov functional $\mathcal{V}_p(\varrho)$ from (14) we have

$$\mathcal{V}_{p,\delta,K}(\varrho) \coloneqq \int_{\mathbb{S}^{d-1}} R_{p,\delta,K}(x) \, \mathrm{d}\varrho(x).$$

Then, by the chain rule and using the weak formulation of (3), it holds for almost all $t \in [0, T]$ that

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{V}_{p,\delta,K}(\rho_t) = \int_{\mathbb{S}^{d-1}} \frac{\left\langle \nabla R_{p,\delta,K}(x), \mathrm{P}_x\left(VB^\top x\right) \right\rangle}{\|B^\top x\|} \, \mathrm{d}\rho_t(x) \tag{45}$$

and it remains to estimate the integrand from above. We start by deriving an upper bound for the inner product $\langle \nabla R_{p,\delta,K}(x), \mathrm{P}_x(VB^\top x)\rangle$. To simplify notation, let us use the orthogonal decomposition of $x \in \mathbb{S}^{d-1}$,

$$x = u + v, \qquad \text{with} \quad u = \mathrm{P}_E x \in E, \quad v = (\mathrm{Id} - \mathrm{P}_E)x \in E^\perp, \quad \|u\|^2 + \|v\|^2 = 1,$$

allowing us to write $R(x) = {}^{\|v\|^2}/_{\|u\|^2}$ for all $x \in \mathbb{S}^{d-1} \setminus E^\perp$. Using the idempotency of P_E, we obtain for all such x that

$$\nabla R(x) = \frac{2v\|u\|^2 - 2u\|v\|^2}{\|u\|^4} = 2R(x)\left(\frac{v}{\|v\|^2} - \frac{u}{\|u\|^2}\right), \tag{46}$$

where the last equality holds for $v \neq 0$ and $\nabla R(x) = 0$ if $v = 0$. Furthermore, by chain rule, it holds

$$\nabla R_{p,\delta,K}(x) = \Psi'_{p,\delta,K}(R(x))\nabla R(x). \tag{47}$$

Let us now bound the quantity $\langle \nabla R_{p,\delta,K}(x), \mathrm{P}_x(VB^\top x)\rangle$. First, note that $\langle \nabla R(x), x\rangle = 0$. In the case that $v = 0$, this holds trivially since $\nabla R(x) = 0$. Otherwise, if $v \neq 0$ it holds

$$\left\langle \frac{v}{\|v\|^2} - \frac{u}{\|u\|^2}, x \right\rangle = \frac{\langle v, u\rangle}{\|v\|^2} + \frac{\|v\|^2}{\|v\|^2} - \frac{\|u\|^2}{\|u\|^2} - \frac{\langle u, v\rangle}{\|u\|^2} = 0$$

thanks to (46) and using that $\langle u, v\rangle = 0$ by orthogonality. Recalling that $\mathrm{P}_x(VB^\top x) = VB^\top x - \langle x, VB^\top x\rangle x$, we compute using (47) for all $x \in \mathbb{S}^{d-1} \setminus E^\perp$ and in the case $v \neq 0$ that

$$\begin{aligned}
\langle \nabla R(x), \mathrm{P}_x(VB^\top x)\rangle &= \langle \nabla R(x), VB^\top x\rangle - \langle x, VB^\top x\rangle \langle \nabla R(x), x\rangle = \langle \nabla R(x), VB^\top x\rangle \\
&= \langle \nabla R(x), VB^\top (u+v)\rangle = \langle \nabla R(x), VB^\top u\rangle + \langle \nabla R(x), VB^\top v\rangle \\
&= 2R(x)\left(\frac{\langle v, VB^\top u\rangle}{\|v\|^2} - \frac{\langle u, VB^\top u\rangle}{\|u\|^2} + \frac{\langle v, VB^\top v\rangle}{\|v\|^2} - \frac{\langle u, VB^\top v\rangle}{\|u\|^2}\right) \\
&= 2R(x)\left(\frac{\langle v, VB^\top v\rangle}{\|v\|^2} - \frac{\langle u, VB^\top u\rangle}{\|u\|^2}\right) \\
&\leq -2(\mu_1 - \mu_2)R(x) = -2\gamma R(x),
\end{aligned}$$

where we used (46) to obtain the fourth line and for the fifth line that by orthogonality and since E is an eigenspace of $VB^\top$, $\langle v, VB^\top u\rangle = \mu_1 \langle v, u\rangle = 0$ as well as $\langle u, VB^\top v\rangle = \langle VB^\top u, v\rangle = \mu_1 \langle u, v\rangle = 0$. To obtain the inequality in the last line, note that $\langle v, VB^\top v\rangle \leq \mu_2 \|v\|^2$ and $\langle u, VB^\top u\rangle = \mu_1 \|u\|^2$, since E is the top eigenspace of $VB^\top$. Further recall the identity $\gamma = \mu_1 - \mu_2$. The same bound holds trivially in the case $v = 0$.

Combining (47) with the above inequality yields in the case of general $p \in (0, 1]$ that

$$\begin{aligned}
\langle \nabla R_{p,\delta,K}(x), \mathrm{P}_x(VB^\top x)\rangle &= \Psi'_{p,\delta,K}(R(x)) \langle \nabla R(x), \mathrm{P}_x(VB^\top x)\rangle \\
&\leq -2\gamma \Psi'_{p,\delta,K}(R(x))R(x)
\end{aligned} \tag{48}$$

for all $x \in \mathbb{S}^{d-1} \setminus E^\perp$. If $R(x) \leq K$ a direct computation yields the Euler-type (recall $t\frac{\mathrm{d}}{\mathrm{d}t}(t^p) = pt^p$ for $t > 0$) inequality

$$\Psi'_{p,\delta,K}(R(x))R(x) \geq p\Psi_{p,\delta,K}(R(x)) = pR_{p,\delta,K}(x). \tag{49}$$

Combining (48) and (49) yields for all $x \in \mathbb{S}^{d-1}$ with $R(x) \leq K$ that

$$\langle \nabla R_{p,\delta,K}(x), \mathrm{P}_x(VB^\top x)\rangle \leq -2p\gamma R_{p,\delta,K}(x). \tag{50}$$

On the other hand, if $K < R(x) \leq 2K$, the right hand side in (48) is non-positive, and if $R(x) > 2K$, it is equal to zero. Combining these observations with (50), it holds for all $x \in \mathbb{S}^{d-1}$ that

$$\langle \nabla R_{p,\delta,K}(x), \mathrm{P}_x(VB^\top x)\rangle \leq -2p\gamma R_{p,\delta,K}(x)\mathbb{1}_{R(x)\leq K}. \tag{51}$$

Employing (51) in (45) and noticing that $\left\|B^\top x\right\| \leq \sigma_{\max}(B)\,\|x\| = \sigma_{\max}(B)$, we obtain the differential inequality

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{V}_{p,\delta,K}(\rho_t) &\leq -2p\gamma \int_{\mathbb{S}^{d-1}} \frac{R_{p,\delta,K}(x)}{\|B^\top x\|}\mathbb{1}_{R(x)\leq K}\,\mathrm{d}\rho_t(x)\\ &\leq -\frac{2p\gamma}{\sigma_{\max}(B)}\int_{\mathbb{S}^{d-1}} R_{p,\delta,K}(x)\mathbb{1}_{R(x)\leq K}\,\mathrm{d}\rho_t(x).\end{aligned}$$

Integrating this inequality yields by the fundamental theorem of calculus

$$\mathcal{V}_{p,\delta,K}(\rho_t) - \mathcal{V}_{p,\delta,K}(\rho_0) \leq -\frac{2p\gamma}{\sigma_{\max}(B)}\int_0^t \int_{\mathbb{S}^{d-1}} R_{p,\delta,K}(x)\mathbb{1}_{R(x)\leq K}\,\mathrm{d}\rho_s(x)\,\mathrm{d}s. \tag{52}$$

Passing first to the limit $K \to \infty$ and then to the limit $\delta \to 0$, and applying Beppo Levi's monotone convergence theorem at each step, we obtain

$$\mathcal{V}_p(\rho_t) - \mathcal{V}_p(\rho_0) \leq -\frac{2p\gamma}{\sigma_{\max}(B)}\int_0^t \mathcal{V}_p(\rho_s)\,\mathrm{d}s.$$

An application of Grönwall's inequality yields

$$\mathcal{V}_p(\rho_t) \leq \mathcal{V}_p(\rho_0)\exp\left(-\frac{2p\gamma}{\sigma_{\max}(B)}t\right),$$

which concludes the proof. □

A.2.3 Invariance of Π under the zero-temperature equation

Before giving the proof of Lemma 9, let us first verify the following technical result.

Lemma 11. *Assume that the matrices B and V satisfy Assumption 1. Then, for $x \in \mathbb{S}^{d-1} \setminus E^\perp$, the Jacobian $D\Pi(x)$ of Π as defined in (4) exists and equals*

$$D\Pi(x) = \frac{1}{\|\mathrm{P}_E x\|}\left(\mathrm{P}_E - \frac{\mathrm{P}_E x\,(\mathrm{P}_E x)^\top}{\|\mathrm{P}_E x\|^2}\right).$$

Moreover, it holds

$$D\Pi(x)\mathrm{P}_x\left(\frac{VB^\top x}{\|B^\top x\|}\right) = 0.$$

Proof. To obtain the first claim, we compute for $x \in \mathbb{S}^{d-1} \setminus E^\perp$ with quotient rule that

$$D\Pi(x) = \frac{\mathrm{P}_E}{\|\mathrm{P}_E x\|} - \frac{\mathrm{P}_E x\,(\mathrm{P}_E x)^\top}{\|\mathrm{P}_E x\|^3} = \frac{1}{\|\mathrm{P}_E x\|}\left(\mathrm{P}_E - \frac{\mathrm{P}_E x\,(\mathrm{P}_E x)^\top}{\|\mathrm{P}_E x\|^2}\right).$$

For the second claim, using that $\mathrm{P}_x(y) = y - \langle x, y\rangle x$ and that $\langle \mathrm{P}_E x, x\rangle = \|\mathrm{P}_E x\|^2$, we derive for arbitrary $y \in \mathbb{S}^{d-1}$ that

$$\begin{aligned}\left(\mathrm{P}_E - \frac{\mathrm{P}_E x(\mathrm{P}_E x)^\top}{\|\mathrm{P}_E x\|^2}\right)\mathrm{P}_x(y) &= \mathrm{P}_E y - \frac{\langle \mathrm{P}_E x, y\rangle}{\|\mathrm{P}_E x\|^2}\mathrm{P}_E x - \langle x, y\rangle\left(\mathrm{P}_E x - \frac{\mathrm{P}_E x\,\langle \mathrm{P}_E x, x\rangle}{\|\mathrm{P}_E x\|^2}\right)\\ &= \mathrm{P}_E y - \frac{\langle \mathrm{P}_E x, y\rangle}{\|\mathrm{P}_E x\|^2}\mathrm{P}_E x.\end{aligned} \tag{53}$$

To simplify notation, let us use the orthogonal decomposition of x,

$$x = u + v, \qquad \text{with} \quad u = \mathrm{P}_E x \in E, \quad v = (\mathrm{Id} - \mathrm{P}_E)x \in E^\perp, \quad \|u\|^2 + \|v\|^2 = 1.$$

Specifically, for $\hat{y} = \frac{VB^\top x}{\|B^\top x\|}$, we have $\hat{y} = \frac{\mu_1 u}{\|B^\top x\|} + \frac{VB^\top v}{\|B^\top x\|}$. Since $VB^\top v \in E^\perp$ (here, we crucially use the symmetry of $VB^\top$), it holds

$$\mathrm{P}_E\hat{y} = \frac{\mu_1 u}{\|B^\top x\|} = \frac{\mu_1 \mathrm{P}_E x}{\|B^\top x\|} \quad \text{and hence} \quad \langle \mathrm{P}_E x, \hat{y}\rangle = \mu_1\frac{\|\mathrm{P}_E x\|^2}{\|B^\top x\|}.$$

Using those identities in the last step, we have with (53) that

$$D\Pi(x)\mathrm{P}_x(\hat{y}) = \frac{1}{\|\mathrm{P}_E x\|}\left(\mathrm{P}_E - \frac{\mathrm{P}_E x(\mathrm{P}_E x)^\top}{\|\mathrm{P}_E x\|^2}\right)\mathrm{P}_x(\hat{y}) = \frac{1}{\|\mathrm{P}_E x\|}\left(\mathrm{P}_E\hat{y} - \frac{\langle \mathrm{P}_E x, \hat{y}\rangle}{\|\mathrm{P}_E x\|^2}\mathrm{P}_E x\right) = 0,$$

which concludes the proof. □

With this auxiliary technical tool, we can now provide a proof of Lemma 9.

Proof of Lemma 9. To keep the notation concise, let us write

$$v_\infty(x) \coloneqq \mathrm{P}_x\left(Vy^*(x)\right) \quad \text{with} \quad y^*(x) \coloneqq \frac{B^\top x}{\|B^\top x\|}$$

for the vector field of the continuity equation (3). Let $\varphi \in \mathcal{C}^\infty(\mathbb{S}^{d-1}, \mathbb{R})$ be a smooth test function and choose a smooth cut-off function $\eta \in \mathcal{C}^\infty([0,\infty),[0,1])$ satisfying

$$\eta(r) = 0 \quad \text{for } r \le 1, \qquad \eta(r) = 1 \quad \text{for } r \ge 2, \qquad |\eta'| \le 2.$$

For $\delta > 0$, define $\eta_\delta(r) \coloneqq \eta(r/\delta)$ and

$$\varphi_\delta(x) \coloneqq \eta_\delta\left(\|\mathrm{P}_E x\|\right)\varphi(\Pi(x)).$$

Clearly, $\varphi_\delta \in \mathcal{C}^\infty(\mathbb{S}^{d-1},\mathbb{R})$ and hence an admissible test function. Using the weak formulation of (3) for φ_δ, it holds

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{S}^{d-1}} \varphi_\delta(x)\,\mathrm{d}\rho_t(x) = \int_{\mathbb{S}^{d-1}} \langle \nabla\varphi_\delta(x), v_\infty(x)\rangle\,\mathrm{d}\rho_t(x),$$

where we can compute by product and chain rule

$$\nabla\varphi_\delta(x) = \eta_\delta(\|\mathrm{P}_E x\|)\nabla(\varphi\circ\Pi)(x) + \eta_\delta'(\|\mathrm{P}_E x\|)\varphi(\Pi(x))\,\nabla\|\mathrm{P}_E x\|, \qquad x \in \mathbb{S}^{d-1},$$

noting that $\nabla\|\mathrm{P}_E x\|$ exists on the support of η_δ'. Since Lemma 11 gives $D\Pi(x)v_\infty(x) = 0$ for all $x \in \mathbb{S}^{d-1}\setminus E^\perp$, we have for every such x that

$$\langle \nabla(\varphi\circ\Pi)(x), v_\infty(x)\rangle = \left\langle D\Pi(x)^\top\nabla\varphi(\Pi(x)), v_\infty(x)\right\rangle = \langle \nabla\varphi(\Pi(x)), D\Pi(x)v_\infty(x)\rangle = 0.$$

Furthermore, for $x \in \mathbb{S}^{d-1}\setminus E^\perp$ we have $\nabla\|\mathrm{P}_E x\| = \frac{\mathrm{P}_E x}{\|\mathrm{P}_E x\|}$. Therefore, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{S}^{d-1}} \varphi_\delta(x)\,\mathrm{d}\rho_t(x) = \int_{\mathbb{S}^{d-1}} \eta_\delta'\left(\|\mathrm{P}_E x\|\right)\varphi(\Pi(x))\left\langle \frac{\mathrm{P}_E x}{\|\mathrm{P}_E x\|}, v_\infty(x)\right\rangle \mathrm{d}\rho_t(x). \tag{54}$$

To simplify notation, let us use the orthogonal decomposition of x,

$$x = u + v, \qquad \text{with} \quad u = \mathrm{P}_E x \in E, \quad v = (\mathrm{Id} - \mathrm{P}_E)x \in E^\perp, \quad \|u\|^2 + \|v\|^2 = 1.$$

Using this notation and that $v_\infty(x) = \mathrm{P}_x\left(\frac{VB^\top x}{\|B^\top x\|}\right) = \frac{VB^\top x - \langle x, VB^\top x\rangle x}{\|B^\top x\|}$, we compute

$$\left\langle \frac{\mathrm{P}_E x}{\|\mathrm{P}_E x\|}, v_\infty(x)\right\rangle = \frac{\langle u, v_\infty(x)\rangle}{\|u\|} = \frac{\mu_1\|u\|^2 - \langle x, VB^\top x\rangle\|u\|^2}{\|B^\top x\|\,\|u\|} = \frac{\|u\|(\mu_1 - \langle x, VB^\top x\rangle)}{\|B^\top x\|},$$

where we used in the second step that by the invariance of E and $E^\perp$, it holds $\mathrm{P}_E(VB^\top x) = \mu_1 u$. This allows us to bound for a constant $C > 0$ depending only on B and $VB^\top$ that

$$\left|\left\langle \frac{\mathrm{P}_E x}{\|\mathrm{P}_E x\|}, v_\infty(x)\right\rangle\right| \le C\|u\| = C\|\mathrm{P}_E x\|.$$

Since η_δ' is supported in $A_\delta \coloneqq \{x \in \mathbb{S}^{d-1} : \delta \le \|\mathrm{P}_E x\| \le 2\delta\}$, we can bound (54) as

$$\left|\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{S}^{d-1}} \varphi_\delta(x)\,\mathrm{d}\rho_t(x)\right| \le C\int_{\mathbb{S}^{d-1}} \underbrace{|\eta_\delta'(\|\mathrm{P}_E x\|)|}_{\le \frac{2}{\delta}\mathbb{1}_{A_\delta}}\,|\varphi(\Pi(x))|\,\underbrace{\|\mathrm{P}_E x\|}_{\le 2\delta}\,\mathrm{d}\rho_t(x) \le C\rho_t(A_\delta), \tag{55}$$

where $C > 0$ is another constant depending solely on B, $VB^\top$, and the maximum norm $\|\varphi\|_\infty$ of $\varphi \in \mathcal{C}^\infty(\mathbb{S}^{d-1}, \mathbb{R})$ on the compact sphere $\mathbb{S}^{d-1}$. It remains to estimate $\rho_t(A_\delta)$. Since on A_δ it holds $\|\mathrm{P}_E x\|^2 \leq 4\delta^2$ and for $\delta \leq 1/4$ moreover

$$\|(\mathrm{Id} - \mathrm{P}_E)x\|^2 = \|v\|^2 = 1 - \|u\|^2 = 1 - \|\mathrm{P}_E x\|^2 \geq 1 - 4\delta^2 \geq 3/4,$$

we have for all $x \in A_\delta$ that

$$R_p(x) = \frac{\|(\mathrm{Id} - \mathrm{P}_E)x\|^{2p}}{\|\mathrm{P}_E x\|^{2p}} \geq c_p \delta^{-2p}$$

for a suitable constant $c_p > 0$. We can therefore estimate

$$\rho_t(A_\delta) = \int_{A_\delta} 1 \,\mathrm{d}\rho_t(x) \leq \frac{\delta^{2p}}{c_p} \int_{A_\delta} R_p(x) \,\mathrm{d}\rho_t(x) \leq \frac{\delta^{2p}}{c_p} \int_{\mathbb{S}^{d-1}} R_p(x) \,\mathrm{d}\rho_t(x) = \frac{\delta^{2p}}{c_p} \mathcal{V}_p(\rho_t).$$

Thanks to Proposition 8, $\sup_{t\in[0,T]} \mathcal{V}_p(\rho_t) \leq \mathcal{V}_p(\rho_0) < \infty$, and thus, as $\delta \to 0$, we get from (55) that

$$\sup_{t\in[0,T]} \left| \frac{\mathrm{d}}{\mathrm{d}t} \int_{\mathbb{S}^{d-1}} \varphi_\delta(x) \,\mathrm{d}\rho_t(x) \right| \leq C\delta^{2p} \mathcal{V}_p(\rho_0) \to 0, \tag{56}$$

where $C > 0$ now depends also on p. By the fundamental theorem of calculus,

$$\int_{\mathbb{S}^{d-1}} \varphi_\delta(x) \,\mathrm{d}\rho_t(x) - \int_{\mathbb{S}^{d-1}} \varphi_\delta(x) \,\mathrm{d}\rho_0(x) = \int_0^t \left(\frac{\mathrm{d}}{\mathrm{d}s} \int_{\mathbb{S}^{d-1}} \varphi_\delta(x) \,\mathrm{d}\rho_s(x) \right) \mathrm{d}s. \tag{57}$$

Recall that by definition of the cut-off function η it holds $|\varphi_\delta| \leq \|\varphi\|_\infty$ for any $\delta > 0$ and $\varphi_\delta \to \varphi \circ \Pi$ pointwise on $\mathbb{S}^{d-1} \setminus E^\perp$ as $\delta \to 0$. Since $\mathcal{V}_p(\rho_t) < \infty$ by Proposition 8, it holds $\rho_t(E^\perp) = 0$. Hence, $\int_{\mathbb{S}^{d-1}} (\varphi \circ \Pi)(x) \,\mathrm{d}\rho_t(x) = \int_{\mathbb{S}^{d-1}\setminus E^\perp} (\varphi \circ \Pi)(x) \,\mathrm{d}\rho_t(x) \leq \max_{z\in\mathbb{S}^{d-1}} |\varphi(z)| < \infty$ since φ is a smooth function on a compact space. We can now pass to the limit $\delta \to 0$ using the dominated convergence theorem to infer the convergence of the integrals on the left-hand side of (57) to $\int_{\mathbb{S}^{d-1}} \varphi(\Pi(x)) \,\mathrm{d}\rho_t(x) - \int_{\mathbb{S}^{d-1}} \varphi(\Pi(x)) \,\mathrm{d}\rho_0(x)$. With the right-hand side of (57) vanishing due to (56) as $\delta \to 0$, we obtain that

$$\begin{aligned} \int_{\mathbb{S}^{d-1}} \varphi(x) \,\mathrm{d}(\Pi_\sharp \rho_t)(x) &= \int_{\mathbb{S}^{d-1}} \varphi(\Pi(x)) \,\mathrm{d}\rho_t(x) \\ &= \int_{\mathbb{S}^{d-1}} \varphi(\Pi(x)) \,\mathrm{d}\rho_0(x) = \int_{\mathbb{S}^{d-1}} \varphi(x) \,\mathrm{d}(\Pi_\sharp \rho_0)(x). \end{aligned}$$

Since this holds for all smooth φ, we conclude that $\Pi_\sharp \rho_t = \Pi_\sharp \rho_0$, as desired. □

A.3 Proof of Corollary 4

Proof of Corollary 4. It is easy to verify that for $t \geq t_1$ it holds $\sqrt{2\mathcal{V}_p(\rho_0)} \exp\left(-p\gamma/\sigma_{\max}(B)t\right) \leq \frac{\varepsilon}{2}$, and that for $t \leq t_2$ it holds $2\sqrt{\log(\beta+1)/\beta}(e^{C_1 t} - e^{C_0 t}) \leq 2\sqrt{\log(\beta+1)/\beta}(e^{C_1 t} - 1) \leq \frac{\varepsilon}{2}$. By choosing β large enough (in fact, $\beta \gg \mathrm{poly}\left(\frac{1}{\varepsilon}\right) \log \frac{1}{\varepsilon}$ suffices) we have $t_1 < t_2$ and the conclusion follows from Theorem 1. □

B Heuristics supporting Conjecture 1

In this section, we provide heuristics supporting Conjecture 1. Some of them can actually be found in the recent work [4], but we include them here for completeness.

One-point dynamics. First, we consider the trivial case where the token distribution is completely collapsed. This captures the situation where the initial distribution ρ_0 is just supported on a single point and approximates the more interesting situation where at some positive time $t > 0$ the token distribution is (almost) collapsed.

We can show that from there on the distribution aligns with the top eigenspace of V. To see this note that for $\varrho = \delta_{x_0}$ with $x_0 \in \mathbb{S}^{d-1}$, we have

$$\int_{\mathbb{S}^{d-1}} \frac{\exp(\beta \langle x, By\rangle)}{\int_{\mathbb{S}^{d-1}} \exp(\beta \langle x, Bz\rangle) \,\mathrm{d}\varrho(z)} Vy \,\mathrm{d}\varrho(y) = Vx_0.$$

It is then easy to observe that the solution of (2) is given by $\rho_t^\beta = \delta_{x(t)}$ where $t \mapsto x(t)$ solves the Oja flow

$$\dot{x}(t) = \mathrm{P}_{x(t)} V x(t), \qquad x(0) = x_0.$$

It follows from [15, Lemmas A.13, A.15] (or a straightforward computation using the quotient rule) that $x(t) = \frac{y(t)}{\|y(t)\|}$ where $t \mapsto y(t)$ solves the system of linear ordinary differential equations

$$\dot{y}(t) = Vy(t).$$

Assuming for simplicity that V is symmetric and that its largest eigenvalue is simple, it follows from the theory of linear ordinary differential equations that $y(t) = \exp(Vt)y(0)$ and hence that $x(t)$ converges exponentially fast to an eigenvector v lying in the eigenspace of V that corresponds to its largest eigenvalue as $t \to \infty$ provided that $\langle x_0, v\rangle \neq 0$. As a consequence, $W_2(\rho_t^\beta, \delta_v) = \|x(t) - v\| \to 0$ as $t \to \infty$. This reasoning can be extended to more general classes of matrices V by using their Jordan normal form.

Some stationary distributions. The above shows, in particular, that δ_v is a stationary distribution of (2). Let us now investigate slightly more complicated stationary distributions of this equation. For this, let us assume that $\varrho = \frac{1}{n}\sum_{i=1}^n \delta_{v_i}$ where v_i are eigenvectors of V with eigenvalue $\lambda > 0$. Abbreviating $N(x) \coloneqq \sum_{i=1}^n \exp(\beta \langle x, Bv_i\rangle)$, we can compute the drift vector field in (2) as

$$\begin{aligned}\int_{\mathbb{S}^{d-1}} \frac{\exp(\beta \langle x, By\rangle)}{\int_{\mathbb{S}^{d-1}} \exp(\beta \langle x, Bz\rangle) \,\mathrm{d}\varrho(z)} Vy \,\mathrm{d}\varrho(y) &= \frac{1}{N(x)} \sum_{i=1}^n \exp(\beta \langle x, Bv_i\rangle) \, Vv_i \\ &= \frac{\lambda}{N(x)} \sum_{i=1}^n \exp(\beta \langle x, Bv_i\rangle) \, v_i, \qquad \forall x \in \mathbb{S}^{d-1}.\end{aligned}$$

Hence, for any smooth $\phi \in \mathcal{C}^\infty(\mathbb{S}^{d-1}, \mathbb{R})$, using that $\mathrm{P}_{v_i}(v_i) = 0$, we have

$$\begin{aligned}&\int_{\mathbb{S}^{d-1}} \left\langle \nabla\phi(x), \mathrm{P}_x \left(\int_{\mathbb{S}^{d-1}} \frac{\exp(\beta \langle x, By\rangle)}{\int_{\mathbb{S}^{d-1}} \exp(\beta \langle x, Bz\rangle) \,\mathrm{d}\varrho(z)} Vy \,\mathrm{d}\varrho(y) \right) \right\rangle \mathrm{d}\varrho(x) \\ &\quad = \frac{1}{n} \sum_{i=1}^n \left\langle \nabla\phi(v_i), \mathrm{P}_{v_i} \left(\frac{\lambda}{N(v_i)} \sum_{j=1}^n \exp(\beta \langle v_i, Bv_j\rangle) \, v_j \right) \right\rangle \\ &\quad = \frac{\lambda}{n} \sum_{i=1}^n \frac{1}{N(v_i)} \left\langle \nabla\phi(v_i), \mathrm{P}_{v_i} \left(\sum_{j \neq i} \exp(\beta \langle v_i, Bv_j\rangle) \, v_j \right) \right\rangle. \end{aligned} \tag{58}$$

If $n = 1$, meaning that $\varrho = \delta_v$ for some $v \in \mathbb{S}^{d-1}$ with $Vv = \lambda v$, then (58) equals zero. This shows that $\varrho = \delta_v$ is a stationary distribution of (2).

A slightly more complex example is $n = 2$ and $\varrho = \frac{1}{2}(\delta_v + \delta_{-v})$, for which we have $\mathrm{P}_v(-v) = -v - \langle v, -v\rangle \, v = 0$ and $\mathrm{P}_{-v}(v) = v - \langle -v, v\rangle \, (-v) = 0$. Hence, (58) equals zero and ϱ is a stationary distribution of (2). Such situation is depicted in Figure 2.

These two classes of stationary distributions concentrated on eigenvectors of V were already identified in [4, Lemma 7] and were referred to as consensus and bipartite consensus therein. In [4, Theorem 2], it was also shown that consensus on an eigenvector with respect to the largest eigenvalue of V is a stable equilibrium and that bipartite consensus can be stable under certain conditions.

C Additional numerical experiments

This appendix contains implementation details and supplementary numerical experiments. We begin by detailing the time discretizations used in our simulations, together with the initialization procedure for the particle simulations. We then present several additional experiments: examples showing effects that can occur when $VB^\top$ is not symmetric, simulations in the gradient flow case $B = \pm V$, and numerical evidence supporting Conjecture 1.

C.1 Details on the discretization

We provide pseudocode for the employed Euler discretization of (1), and the zero-temperature limit (3). Let us highlight here that Algorithm 1 is not a faithful discretization of the zero-temperature equation. Namely, formally setting $\beta = \infty$ in (1) amounts to replacing the softmax in lines 2–3 of Algorithm 1 by a hardmax function. In the finite-particle case, however, this does not recover Algorithm 2. We note that whenever we run a numerical experiment in what follows with $\beta = \infty$, this means we employ Algorithm 2.

Algorithm 1 Discretization of (1)

Require: Tokens $\{x_i\}_{i=1}^n \subset \mathbb{S}^{d-1}$, matrices $B, V \in \mathbb{R}^{d\times d}$, β, $\Delta t > 0$.
1: **for** $i = 1, \dots, n$ **do**
2: $\quad w_{ij} \leftarrow \dfrac{\exp\left(\beta \langle x_i, Bx_j\rangle\right)}{\sum_{k=1}^n \exp\left(\beta \langle x_i, Bx_k\rangle\right)}, \quad j = 1, \dots, n$
3: $\quad m_i \leftarrow \sum_{j=1}^n w_{ij}\, x_j$ ▷ softmax-weighted consensus
4: $\quad x_i \leftarrow x_i + \Delta t\, \mathrm{P}_{x_i} V m_i$
5: $\quad x_i \leftarrow x_i \,/\, \|x_i\|$ ▷ retract to $\mathbb{S}^{d-1}$
6: **end for**

Algorithm 2 Discretization of the zero-temperature limit (3)

Require: Tokens $\{x_i\}_{i=1}^n \subset \mathbb{S}^{d-1}$, matrices $B, V \in \mathbb{R}^{d\times d}$, $\Delta t > 0$
1: **for** $i = 1, \dots, n$ **do**
2: $\quad m_i \leftarrow B^\top x_i \,/\, \|B^\top x_i\|$
3: $\quad x_i \leftarrow x_i + \Delta t\, \mathrm{P}_{x_i} V m_i$
4: $\quad x_i \leftarrow x_i \,/\, \|x_i\|$ ▷ retract to $\mathbb{S}^{d-1}$
5: **end for**

Initialization. For the visualization in Figure 1, we employ a mixture of von Mises–Fisher distributions, with densities, that up to normalization are given as

$$p(x|\mu_i, \kappa_i) \sim \exp(\kappa_i \mu_i^\top x)$$

with parameters $\kappa_i \in [0, \infty)$ and $\mu_i \in \mathbb{R}^d$ for $i = 1, \dots, m$. For weights w_i with $\sum_{i=1}^m w_i = 1$, the mixture then fulfills

$$p(x|\{(\mu_i, \kappa_i)\}_{i=1}^m) \sim \sum_{i=1}^m w_i\, p(x|\mu_i, \kappa_i).$$

For our examples, we choose

$$\begin{gathered}\mu_1 = (1, -0.3, -0.2),\ \mu_2 = (0, 1, -0.3),\ \mu_3 = (-1, 1, 1),\\ \kappa_1 = 2,\ \kappa_2 = 10,\ \kappa_3 = 5, \quad w_1 = w_2 = w_3 = 1/3.\end{gathered}$$

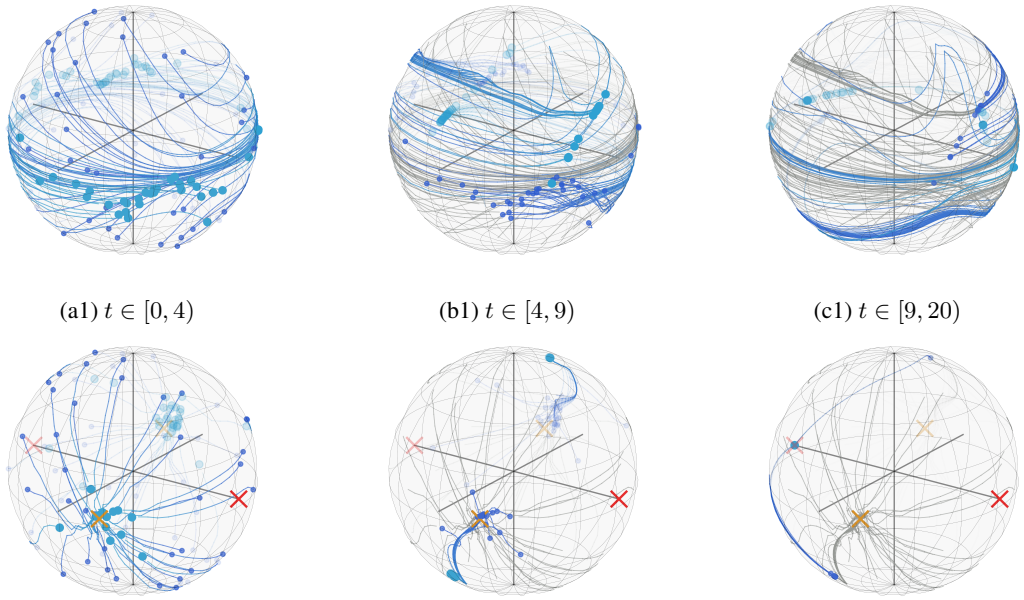
C.2 Non-symmetric $VB^\top$ effects

We illustrate the role of the symmetry assumption on $VB^\top$ through two examples. To do so, we consider the following matrices

$$\begin{aligned} V_1 &= \begin{pmatrix} -1 & 1 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}, \qquad B_1 = \mathrm{diag}(-1, -1, 1), \\ V_2 &= \mathrm{diag}(-1, 1, -2), \qquad B_2 = \begin{pmatrix} -1 & 1 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \end{aligned} \tag{59}$$

In this case, note that the eigenvalues of $V_i B_i^\top$ and V_i are not necessarily real. In fact, the eigenvalues of $V_1 B_1^\top$ and V_1 equal $\pm i$ and -2, whereas the eigenvalues of $V_2 B_2^\top$, despite being a non-symmetric

matrix, are real and equal $1 \pm \sqrt{2}$ and -2. Those of V_2 are equal to ± 1 and -2. The trajectories for these two settings are displayed in Figure 5. In the top row corresponding to V_1 and B_1, the complex eigenvalues lead to a visible rotation of the trajectories, and no clear alignment is observed. In the bottom row, however, $V_2 B_2^\top$ is still non-symmetric but has real eigenvalues, and the same two-step behavior as in the symmetric case is recovered.



(a1) $t \in [0, 4)$ (b1) $t \in [4, 9)$ (c1) $t \in [9, 20)$

(a2) $t \in [0, 1.5)$ (b2) $t \in [1.5, 6)$ (c2) $t \in [6, 20)$

Figure 5: Dynamics of $n = 200$ tokens evolving according to (1) with $\beta = 100$ for $VB^\top$ not symmetric. In the top row, corresponding to the choice V_1, B_1 in (59), we observe that trajectories rotate around the sphere and do not display clear convergence or clustering. In the bottom row, corresponding to V_2, B_2 in (59), we can see exactly the same behavior as in Figure 2: initial alignment with E, followed by a second phase of alignment with F.

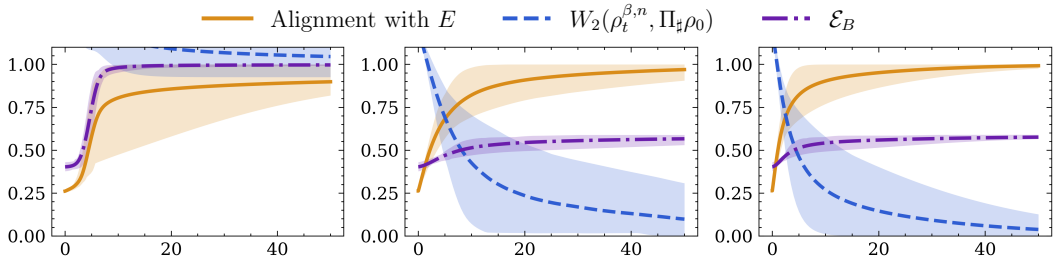
C.3 Gradient flow case

In this section, we study the alignment behavior in the gradient flow case, i.e., $B = \pm V$. In particular, as highlighted in [17], in this case the self-attention dynamics try to either maximize or minimize the energy

$$\mathcal{E}_B(\varrho) = \int_{\mathbb{S}^{d-1}} \int_{\mathbb{S}^{d-1}} e^{\beta \langle x, By \rangle} \, \mathrm{d}\varrho(x) \, \mathrm{d}\varrho(y).$$

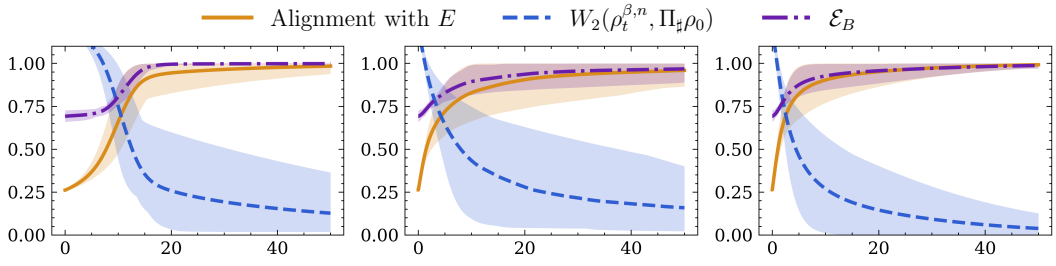
These experiments complement our theory by testing the predicted alignment behavior in the more specific gradient flow setting, where the dynamics admit an additional variational interpretation. In the following, we perform numerical experiments for four cases: maximization and minimization, each with positive and negative definite matrices. Throughout, $v_{\max}(B)$ and $v_{\min}(B)$ denote a unit eigenvector of B associated with its largest and smallest eigenvalues, respectively.

Maximization case. For $B = V$, the self-attention dynamics try to maximize the energy $\mathcal{E}_B$ and the optimal stationary state is a single Dirac at $\pm v_{\max}(B)$, i.e., $\rho^{\text{opt}} = \delta_{v_{\max}(B)}$ or $\rho^{\text{opt}} = \delta_{-v_{\max}(B)}$. However, the example in [17, Figure 2] suggests that, for large β, the stationary state supported fully on the top eigenspace of B is more likely to occur. This is consistent with Theorem 1, since in this case $VB^\top = B^2$. We display this behavior in Figure 6, where the purple line displays the evolution of $\mathcal{E}_B(\rho_t^{\beta,n})/\mathcal{E}_B(\rho^{\text{opt}})$, i.e., the energy of the ensemble normalized by the theoretical best energy. For small values of β, the iteration indeed converges towards an energetically optimal state, which is a single Dirac. However, since the initial configuration is sampled uniformly on the sphere,



(a) $\beta = 1$ (b) $\beta = 100$ (c) $\beta = \infty$

Figure 6: Alignment in the gradient flow maximization case, i.e., $B = V$, with B being a symmetric positive definite matrix. We run the experiment for 10 different random configurations. We consider $n = 100$ particles initialized independently from ρ_0, the uniform distribution on $\mathbb{S}^{d-1}$.



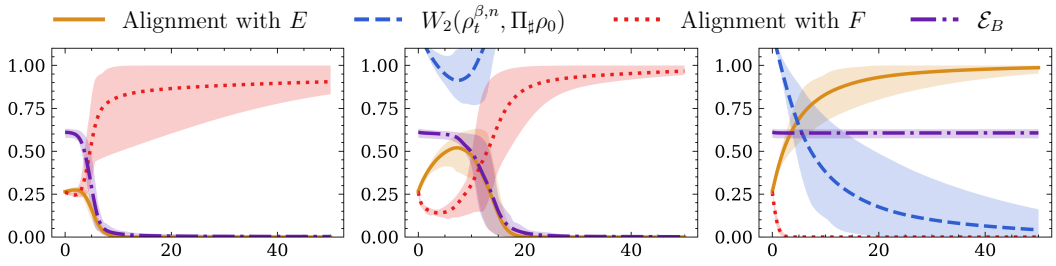
(a) $\beta = 1$ (b) $\beta = 100$ (c) $\beta = \infty$

Figure 7: Alignment in the gradient flow maximization case, i.e., $B = V$, with B being a symmetric negative definite matrix. The remaining setup is the same as in Figure 6.

$\Pi_\sharp \rho_0^n$ is likely clustered both at $v_{\max}(B)$ and $-v_{\max}(B)$, which causes the Wasserstein distance in Figure 6 to be large. For larger values of β, one instead observes convergence towards energetically suboptimal states, which are closer to $\Pi_\sharp \rho_0^n$.

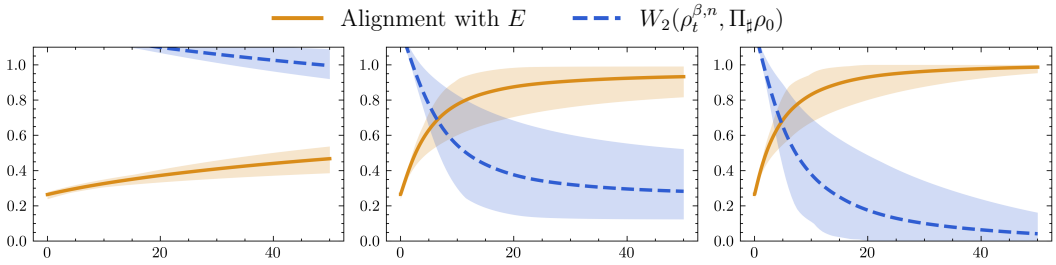
On the other hand, if B is negative semi-definite, the optimal state is $\rho^{\text{opt}} = \frac{1}{2}(\delta_{v_{\min}(B)} + \delta_{-v_{\min}(B)})$, where in this case $v_{\min}(B) = v_{\max}(B^2)$. In Figure 7, we observe that even for small values of β, we indeed obtain convergence of $\rho_t^{\beta,n}$ towards $\Pi_\sharp \rho_0^n$. Moreover, we also obtain convergence towards energetically optimal states for large values of β.

Minimization case. We now consider the case when $B = -V$, in which case the self-attention dynamics try to minimize the energy $\mathcal{E}_B$. Thus, the purple line in the following figures will now display $1 - \mathcal{E}_B(\rho^{\text{opt}})/\mathcal{E}_B(\rho_t^{\beta,n})$. We first consider the case where B is negative definite. Then, the optimal state is given as $\rho^{\text{opt}} = \delta_{v_{\min}(B)}$ and here $v_{\min}(B)$ does not coincide with $v_{\max}(VB^\top) = v_{\max}(-B^2)$, but instead $v_{\max}(V) = v_{\max}(-B) = v_{\min}(B)$. This can be observed in Figure 8, where for small β, we obtain convergence towards an energetically optimal state and alignment with F. For intermediate β, we obtain the effect highlighted in Figure 2, where we first obtain a phase of alignment with E and then align to F, which here is also a phase where $\mathcal{E}_B$ is minimized. For $\beta = \infty$, as expected, we only obtain alignment towards E and the energy of the ensemble stays constant. Since $B = -V$, this case suggests that the second phase emerges from a disagreement between the top eigenspaces of B and V. Finally, we consider the minimization case with B being symmetric positive definite. Here, the optimal states are not known analytically and we thus do not display the purple line in Figure 9. However, the experiments in [17] suggest that the optimal state has two modes at $v_{\min}(B)$ and $-v_{\min}(B)$. These modes are sharper the closer the smallest eigenvalue of B is to zero, and in fact the optimal state is $\frac{1}{2}(\delta_{v_{\min}(B)} + \delta_{-v_{\min}(B)})$ when it is exactly zero. Since in this case $v_{\min}(B) = v_{\max}(VB^\top)$, this observation is in line with Theorem 1. While in [17] this effect was observed for $\lambda_{\min}(B) \to 0$, we similarly observe it in Figure 9 for $\beta \to \infty$. We see alignment and Wasserstein convergence only for larger values of β.



(a) $\beta = 1$ (b) $\beta = 100$ (c) $\beta = \infty$

Figure 8: Alignment in the gradient flow minimization case, i.e., $B = -V$, with B being a symmetric negative definite matrix. The remaining setup is the same as in Figure 6.



(a) $\beta = 1$ (b) $\beta = 100$ (c) $\beta = \infty$

Figure 9: Alignment in the gradient flow minimization case, i.e., $B = -V$, with B being a symmetric positive definite matrix. The remaining setup is the same as in Figure 6.

C.4 Experiments supporting Conjecture 1

We provide additional numerical evidence for Conjecture 1. For generic token configurations, we conjecture that the long-time behavior of (1) is governed by an eigenspace of V: either F, associated with the largest eigenvalue of V, or F^{abs}, associated with an eigenvalue of largest magnitude. The simulations below illustrate both possibilities.

Figure 10 shows three representative random diagonal instances of V and B. For $\beta = 10$, panel (a) exhibits asymptotic alignment with F. In panel (b), $F = F^{\mathrm{abs}}$, and the dynamics align with this common eigenspace. In panel (c), $E = F^{\mathrm{abs}}$, and the limiting alignment is with this eigenspace.

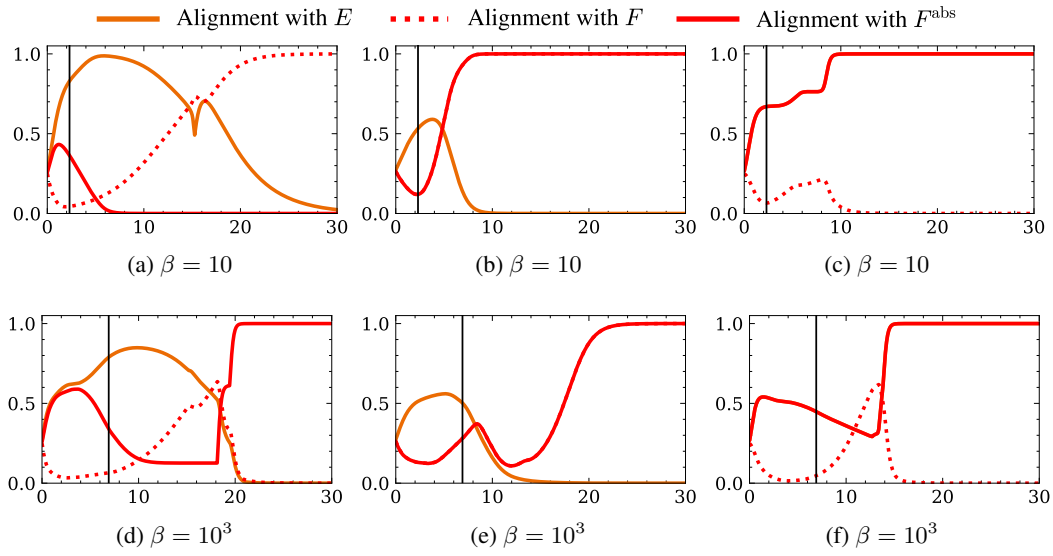
For $\beta = 10^3$, panels (e) and (f) show the same qualitative behavior as panels (b) and (c), respectively, up to the expected delay on the time scale $t \approx \log \beta$. In panel (d), however, the limiting alignment changes from F to F^{abs}. These simulations therefore support the prediction that the dynamics align with an eigenspace of V, while indicating that the selected eigenspace can be either F or F^{abs}, depending on the matrix realization, the initial configuration, and the temperature regime.

C.5 Empirical validation of the $\mathcal{O}(\log \beta)$ time scale

In this section, we confirm experimentally the interpretation of Theorem 1, see the discussion ahead of and after Corollary 4, which states that the relevant time scale for the exponential convergence to the top eigenspace E of $VB^\top$ is of order $\mathcal{O}(\log \beta)$, before the dynamics may enter a terminal phase of convergence to an eigenspace of V.

Since previous examples show that certain configurations exhibit alignment towards the top eigenspace E as their long-term behavior (e.g., when the top eigenspace F of V coincides with E), we specifically select a scenario, where the alignment breaks after a certain time frame and then record

$$t^\beta \in \operatorname*{arg\,min}_{t \in [0,T]} W_2(\rho_t^{\beta,n}, \Pi_\sharp \rho_0). \tag{60}$$



(a) $\beta = 10$ (b) $\beta = 10$ (c) $\beta = 10$

(d) $\beta = 10^3$ (e) $\beta = 10^3$ (f) $\beta = 10^3$

Figure 10: Alignment of $\rho_t^{\beta,n}$ with the top eigenspace E of $VB^\top$ and with the eigenspaces F and F^{abs} of V, for $d = 10$ (maximum alignment equals 1). The vertical line marks $t = \log \beta$. Each column uses a different pair of random diagonal matrices V and B, with independent normally distributed diagonal entries. The initial condition consists of $n = 500$ tokens sampled from the uniform distribution ρ_0 on $\mathbb{S}^{d-1}$.

Specifically, we consider the gradient flow case with $V = -B$ and B being an indefinite matrix in dimension $d = 10$. In Figure 11, we observe that the empirical times t^β recorded in this setup indeed scale like a logarithmic curve.

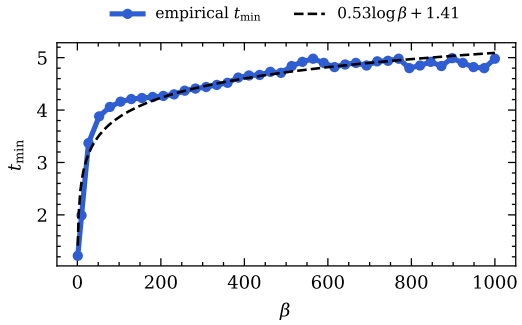


Figure 11: We plot t^β as defined in (60), i.e., the time at which the distance between $\rho_t^{\beta,n}$ and the projection of the initial condition $\Pi_\sharp \rho_0$ is minimal. For illustrative purposes, we fit a logarithmic curve to support the claimed $\mathcal{O}(\log \beta)$ scaling.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Our claims made in the abstract and the introduction accurately reflect our paper's main contributions and are fully supported by the main body of the paper.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations of our work in a dedicated section at the end of the manuscript. In particular, we explain both theoretically and experimentally how violations of our theoretical assumptions might affect the considered model.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: The paper states precisely its lemmas, propositions, corollaries, and theorems. Each statement clearly states and references its assumptions. Detailed and carefully written proofs are given for each statement, either in the main body of the paper or the appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: We disclose all information (hyperparameters, initial conditions, etc.) needed to reproduce the experimental results of the paper.

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: We provide the code associated with this paper in the GitHub repository `https://github.com/albertalcalde/mean-field-transformer-concentration`.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We provide full details about our experimental settings.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   We provide error bars to support the statistical significance of our experiments.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The numerical experiments included in this paper are merely for illustrative purposes and were run on a standard laptop on CPU, taking just a few seconds. Therefore, we do not include any hardware-related details in the paper.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: The paper focuses on theoretical analysis of a simplified transformer architecture which, as is, cannot be immediately used for real-life applications. We therefore do not expect or discuss any societal impact.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not pose such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: We do not use such assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: We do not introduce such assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not conduct research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: The research conducted in this paper does not involve LLMs as any important, original, or non-standard components. LLMs have been merely used for writing, editing, formatting, or coding purposes.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.