

Quantifying Low-Temperature Concentration in Mean-Field Transformers

Albert Alcalde (FAU), Leon Bungert (JMU), Konstantin Riedl (Oxford), and Tim Roith (TUM)

Workshop on the Mathematics of Transformers
Oxford (Aug 2026)



Funded by
the European Union

Forward dynamics of deep Transformers

We study encoder-only transformers at inference in the infinite-depth regime. We consider normalized particles or *tokens* $\{x_i(t)\}_{i \in [n]} \subset \mathbb{S}^{d-1}$.

$$\dot{x}_i(t) = P_{x_i(t)} \left(\sum_{j=1}^n \frac{\exp(\beta \langle Qx_i(t), Kx_j(t) \rangle)}{\sum_{k=1}^n \exp(\beta \langle Qx_i(t), Kx_k(t) \rangle)} Vx_j(t) \right),$$
$$x_i(0) = x_i^0 \in \mathbb{S}^{d-1}.$$

- ▶ $P_x(y) = y - \langle x, y \rangle x$ the projection onto $T_x \mathbb{S}^{d-1}$.
- ▶ $\beta > 0$ is an inverse temperature parameter, $Q, K, V \in \mathbb{R}^{d \times d}$ are trainable matrices fixed after training. We set $B = Q^\top K$.
- ▶ We omit layer-varying parameters, FFNs, multi-head attention, causal maskings and positional encodings.

Mean-field description

In the large-token limit (modern LLMs have context windows $n \sim 10^6$), the empirical measure $\rho_t^{\beta, n} = \frac{1}{n} \sum_i \delta_{x_i(t)}$ is described by a continuity equation on $\mathcal{P}(\mathbb{S}^{d-1})$:

$$\partial_t \rho_t^\beta = -\nabla \left(\rho_t^\beta P_x \left(\int_{\mathbb{S}^{d-1}} \frac{e^{\beta \langle x, \textcolor{blue}{B} y \rangle}}{\int_{\mathbb{S}^{d-1}} e^{\beta \langle x, \textcolor{blue}{B} z \rangle} d\rho_t^\beta(z)} \textcolor{blue}{V} y d\rho_t^\beta(y) \right) \right). \quad (\textcolor{red}{SA}_\beta)$$

As $\beta \rightarrow \infty$, the softmax concentrates on maximizers, and the limiting PDE is shown to be¹

$$\partial_t \rho_t = -\nabla \left(\rho_t P_x \left(\frac{\textcolor{blue}{V} \textcolor{blue}{B}^\top x}{\| \textcolor{blue}{B}^\top x \|} \right) \right). \quad (\textcolor{red}{SA}_\infty)$$

Goal: quantify how close finite- β dynamics ($\textcolor{red}{SA}_\beta$) stay to the zero temperature dynamics ($\textcolor{red}{SA}_\infty$).

¹Bruno et al. (2025): [neurIPS'25](#)

Lending tools from Consensus-Based Optimization

The aim of consensus-based optimization² (CBO) on $\mathbb{S}^{d-1}$ is to locate a global minimizer of an objective $J : \mathbb{S}^{d-1} \rightarrow \mathbb{R}$ by evolving a particle ensemble according to

$$\dot{x}_i(t) = P_{x_i(t)}(m_{\beta, \rho_t^{\beta, n}}) + \sigma \cdot \text{noise}. \quad (\text{CBO})$$

The consensus point is a weighted mean to which particles with smaller values of J contribute more.

Particle weights can also be determined through a pairwise interaction via polarization. This kernelizes the consensus point as³

$$m_{\beta, \varrho}^{\kappa}(x) = \int_{\mathbb{S}^{d-1}} \frac{\kappa(x, y) e^{-\beta J(y)}}{\int_{\mathbb{S}^{d-1}} \kappa(x, z) e^{-\beta J(z)} d\varrho(z)} y d\varrho(y). \quad (\text{polar CBO})$$

²[R. Pinnau et al. \(2017\)](#)

³[L. Bungert et al. \(2025\)](#)

Self-attention as polarized CBO

With the energy norm $\|x\|_B := \sqrt{\langle x, Bx \rangle}$, choose

$$J(x) = -\frac{1}{2} \|x\|_B^2, \quad \kappa(x, y) = \exp\left(-\frac{\beta}{2} \|x - y\|_B^2\right).$$

The polarized consensus point then becomes

$$m_{\beta, \varrho}^{\kappa}(x) = \int_{\mathbb{S}^{d-1}} \frac{e^{\beta \langle x, By \rangle}}{\int_{\mathbb{S}^{d-1}} e^{\beta \langle x, Bz \rangle} d\varrho(z)} y d\varrho(y).$$

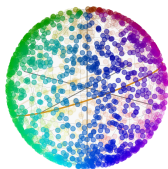
In the noiseless case $\sigma = 0$, this leads exactly to the self-attention dynamics when

$$V \equiv I, \quad B \equiv Q^\top K \quad \text{is time-independent.}$$

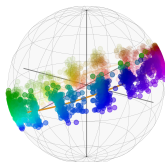
This connection motivates the analytical tools we will use to study mean-field transformers.

Concentration onto a subspace

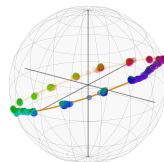
Assume $VB^\top$ is symmetric, with dominant eigenspace E .



$t = 0$



$t = 2.5$



$t = 5$

$n = 2000$ tokens, $\beta = 30$. The orange circle is $E \cap \mathbb{S}^{d-1}$.

The geometry controlling concentration

Define:

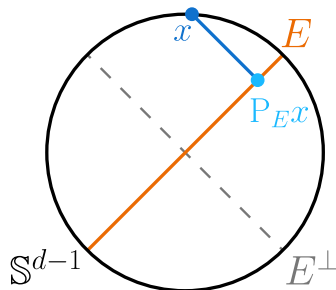
$$\Pi(x) = \frac{P_E x}{\|P_E x\|}, \quad x \notin E^\perp.$$

Here, P_E denotes the orthogonal projection onto E .

The stationary distribution is

$$\rho_\infty := \Pi_\# \rho_0,$$

i.e. the initial distribution projected onto $E \cap \mathbb{S}^{d-1}$.



Standing assumptions

Assumption A: weights

A.1 B is invertible and $VB^\top$ is symmetric.

A.2 The largest eigenvalue μ_1 of $VB^\top$ has multiplicity k , with eigenspace E . Further, the spectral gap $\gamma := \mu_1 - \mu_2 > 0$.

Assumption B: initial datum

For some $p \in (0, 1]$, $\mathcal{V}_p(\rho_0) := \int_{\mathbb{S}^{d-1}} \frac{\|(I - P_E)x\|^{2p}}{\|P_E x\|^{2p}} d\rho_0(x) < \infty$.

Assumption C: density

ρ_0 has a Lipschitz density f_0 with respect to surface measure on $\mathbb{S}^{d-1}$, and $\ell_0 := \min_{x \in \mathbb{S}^{d-1}} f_0(x) > 0$.

Main result

Theorem [ABRR'26]

Assume A–C. Let $(\rho_t^\beta)_{t \in [0, T]}$ solve the finite- β transformer continuity equation (**SA** $_\beta$). Then

$$W_2(\rho_t^\beta, \Pi_\# \rho_0) \leq 2\sqrt{\frac{\log(\beta + 1)}{\beta}}(e^{C_1 t} - e^{C_0 t}) + \sqrt{2\mathcal{V}_p(\rho_0)}e^{-\frac{p\gamma}{\sigma_{\max}(B)}t}$$

where C_0, C_1 are positive constants depending only on d, B, V and ℓ_0 .

Finite-temperature error

$$\sim \sqrt{\frac{\log \beta}{\beta}} e^{C_1 t}$$

Small for large β , but not uniformly in time.

Zero-temperature decay

$$\sim \exp\left(-\frac{p\gamma}{\sigma_{\max}(B)}t\right)$$

Drives concentration toward $E \cap \mathbb{S}^{d-1}$.

Exact zero-temperature convergence

Corollary [ABRR'26]

Assume A and B. Let $(\rho_t)_{t \in [0, T]}$ solve the zero-temperature transformer continuity equation (SA_∞). Then

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

Proof sketch of the main theorem

Step 1: finite- β approximation

Show that

$$W_2(\rho_t^\beta, \rho_t) \lesssim \sqrt{\frac{\log(\beta + 1)}{\beta}} (e^{C_1 t} - e^{C_0 t}).$$

Step 2: zero-temperature contraction

Show that

$$W_2(\rho_t, \Pi_{\#} \rho_0) \leq \sqrt{2\mathcal{V}_p(\rho_0)} e^{-p\gamma t / \sigma_{\max}(B)}.$$

The theorem follows by the triangle inequality:

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

Proof sketch: finite- β approximation I

For fixed x , the attention weights define a measure proportional to $e^{\beta \langle B^\top x, y \rangle} d\rho(y)$. On the sphere,

$$y^*(x) = \arg \max_{y \in \mathbb{S}^{d-1}} \langle B^\top x, y \rangle = \frac{B^\top x}{\|B^\top x\|},$$

where invertibility of B makes $y^*(x)$ well-defined. Setting:

$$\mathcal{A}_\beta[\rho](x) := \int_{\mathbb{S}^{d-1}} \frac{e^{\beta \langle x, By \rangle}}{\int_{\mathbb{S}^{d-1}} e^{\beta \langle x, Bz \rangle} d\rho(z)} V y d\rho(y),$$

$$\mathcal{A}_\infty(x) := V y^*(x) = \frac{V B^\top x}{\|B^\top x\|}.$$

Proof sketch: finite- β approximation II

Lemma (Quantitative Laplace principle) [ABRR'26]

Let Assumption C hold, which ensures that spherical caps around $y^*(x)$ retain positive mass. Quantitative Laplace estimates⁴ then yield, uniformly in x ,

$$\left\| \mathcal{A}_\beta[\rho_t^\beta](x) - \mathcal{A}_\infty(x) \right\| \lesssim e^{C_0 t} \sqrt{\frac{\log(\beta + 1)}{\beta}}.$$

Let ρ_t^β and ρ_t solve (SA_β) and (SA_∞) , respectively. A Grönwall-type stability estimate for continuity equations then gives

$$W_2(\rho_t^\beta, \rho_t) \leq 2 \sqrt{\frac{\log(\beta + 1)}{\beta}} (e^{C_1 t} - e^{C_0 t}).$$

⁴K. Riedl (2024)

Proof sketch: zero-temperature contraction

Lyapunov-type decay

Along the solution of (SA_∞), it holds

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

Invariance under Π and distance coupling

It holds that $\Pi_{\#}\rho_t = \Pi_{\#}\rho_0$ for all t . Further,

$$W_2^2(\rho_t, \Pi_{\#}\rho_0) = W_2^2(\rho_t, \Pi_{\#}\rho_t) \leq 2\mathcal{V}_p(\rho_t)$$

All above combined, yields

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

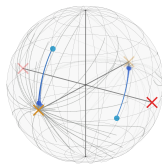
Two time scales: a metastable window

Combining both parts yields

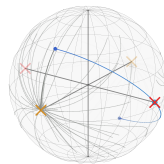
$$W_2(\rho_t^\beta, \Pi_\# \rho_0) \lesssim \begin{cases} \exp\left(-\frac{p\gamma}{\sigma_{\max}(B)}t\right), & 0 \leq t \lesssim \log \beta, \\ e^{C_1 t} \sqrt{\frac{\log(\beta+1)}{\beta}}, & t \gtrsim \log \beta. \end{cases}$$



$t \in [0, 4)$



$t \in [4, 9)$

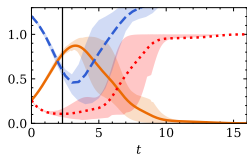


$t \in [9, 20)$

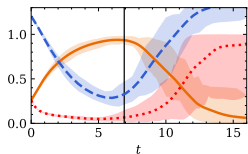
$$n = 200, \beta = 100, VB^\top = \text{diag}(1, -1, -2).$$

Larger β delays the second phase

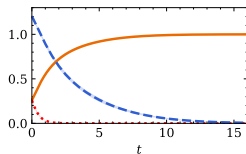
— Alignment with E ···· Alignment with F — $W_2(\rho_t^{\beta,n}, \Pi_{\#}\rho_0)$



$\beta = 10$



$\beta = 10^3$



$\beta = \infty$

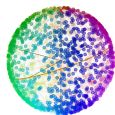
$d = 10$, $n = 500$, 20 runs. Blue curve denotes the Wasserstein error, while vertical line is $t = \log \beta$.

Conjectured long-time behavior

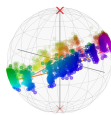
The theorem controls the non-asymptotic regime. Numerics and related work⁵ suggest a different asymptotic behavior for finite β .

Conjecture [ABRR'26]

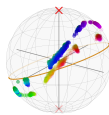
For some $B, V \neq I$, finite- β dynamics (SA_β) converge as $t \rightarrow \infty$ to a distribution supported on a dominant eigenspace of V .



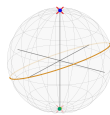
$t = 0$



$t = 2.5$



$t = 4$



$t = 10$

Initial concentration at $E \cap \mathbb{S}^{d-1}$ followed by clustering on $F \cap \mathbb{S}^{d-1}$.

⁵Altafini (2026)

Why the terminal phase should be controlled by V

Suppose that at some late time the token distribution is nearly collapsed, $\rho_t^\beta \approx \delta_{x(t)}$. Then attention has only one location to select, independently of B and β :

$$\int_{\mathbb{S}^{d-1}} \frac{e^{\beta \langle x, By \rangle}}{\int_{\mathbb{S}^{d-1}} e^{\beta \langle x, Bz \rangle} d\delta_x(z)} V y d\delta_x(y) = V x.$$

Consequently, the remaining motion is approximately the *Oja flow*

$$\dot{x} = P_x(Vx).$$

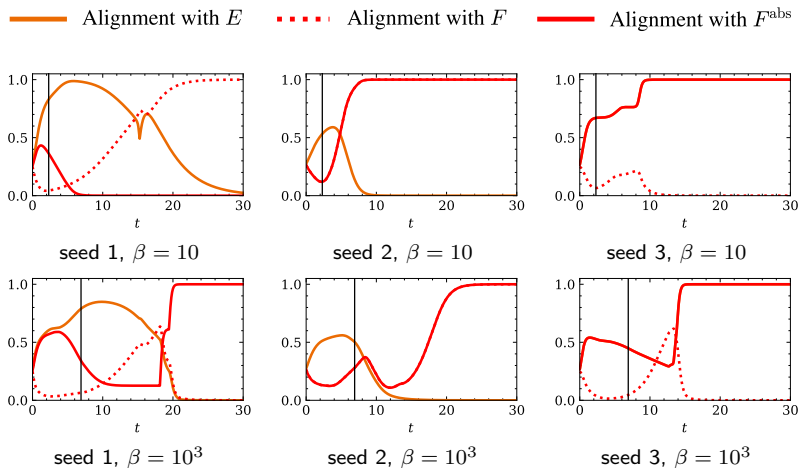
Heuristic for symmetric V

The solution satisfies

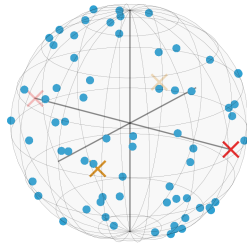
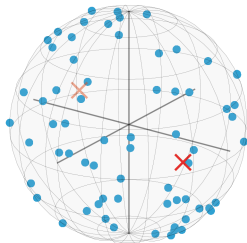
$$x(t) = \frac{e^{tV} x(0)}{\|e^{tV} x(0)\|},$$

and therefore generically aligns with a dominant eigenspace of V .

Numerical experiments: diagonal V, B



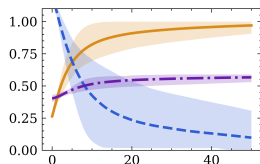
Numerical experiments: non-symmetric $VB^\top$



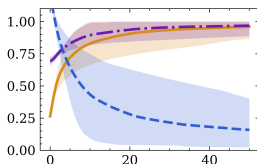
$\mu_i = \pm i, -2$: rotation, no clustering. $\mu_i = 1 \pm \sqrt{2}, -2$: two-step alignment.

Numerical experiments: gradient-flow case $B = \pm V$

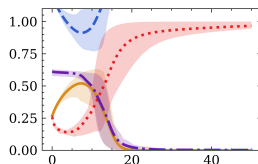
— Alignment with E - - - $W_2(\rho_t^{\beta,n}, \Pi_{\sharp}\rho_0)$ - · - $\mathcal{E}_B$



$B = V, B \succ 0$



$B = V, B \prec 0$



$B = -V, B \prec 0$

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

- ▶ When $B = V$, the flow maximizes the energy $\mathcal{E}_B$,
- ▶ When $B = -V$, the flow minimizes the energy $\mathcal{E}_B$.

The inverse-temperature β and the sign of B determine whether the dynamics follow (SA $_{\infty}$) or the energetically preferred stationary state.

Summary

- ▶ Mean-field self-attention admits a finite-temperature continuity equation and a zero-temperature limit.
- ▶ If $VB^\top$ is symmetric with spectral gap, the zero-temperature flow converges exponentially to $\Pi_{\sharp}\rho_0$.
- ▶ Therefore concentration onto $E \cap \mathbb{S}^{d-1}$ is quantitative but metastable, valid up to $t = O(\log \beta)$.
- ▶ We conjecture a later terminal phase dominated by eigenspaces of V , not $VB^\top$.

Thank you for your attention!

