

# KINETIC THEORY FOR TRANSFORMERS AND THE LOST-IN-THE-MIDDLE PHENOMENON

MITIA DUERINCKX, BORJAN GESHKOVSKI, AND STEFANO ROSSI

**ABSTRACT.** We study causal self-attention dynamics—a toy model for decoder Transformers—which we interpret as a non-exchangeable interacting particle system. Adapting cumulant expansions to the triangular causal dependency structure of the model, and appealing to non-hierarchical methods to estimate correlations using Glauber calculus, we prove a quantitative mean-field limit result and a next-order characterization of correlations. For iid uniformly distributed tokens, the limiting correlation equation can be solved in closed form and we obtain a rigorous explanation of the empirically observed *lost-in-the-middle* phenomenon: the token retrieval profile, as a function of the source position in the prompt, is U-shaped, with primacy, recency, and a unique interior minimum under an explicit smallness condition.

## 1. INTRODUCTION AND MAIN RESULTS

Large language models display several long-context effects, one of the cleanest being the dependence of the output on the position of relevant information inside the prompt. A particularly simple protocol consists in inserting a single relevant fact at a chosen location, filling the rest of the prompt with distractors, and then asking a retrieval or question-answering task whose correct answer depends only on that fact. Repeating this experiment across many prompts and across a wide class of models reveals a consistent pattern: accuracy is high near the beginning of the prompt, high again near the end, and lower in the middle. This empirically observed pattern goes by the name of *lost-in-the-middle* [30]; see Figure 1.

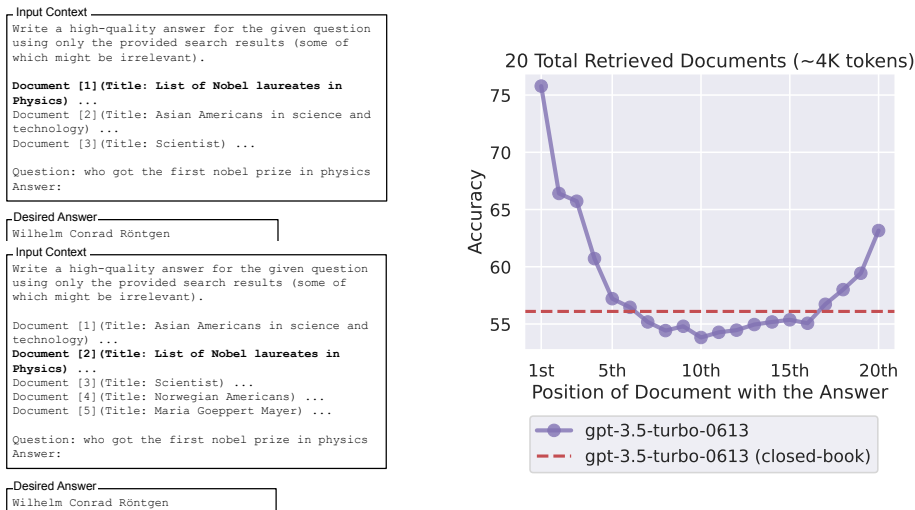


FIGURE 1. The experiment in [30] that motivates the paper. The two panels on the left show how the answer-containing document is moved through the prompt and how the amount of surrounding text is changed. The curve on the right is the empirical lost-in-the-middle profile: retrieval is strongest near the beginning and the end of the context, and weakest in the middle. Reproduced from Figures 3, 4, and 1 of [30], respectively, with permission from the authors.

A convenient framework for addressing this question from a mathematical lens is to view Transformers [38]—the neural network architecture underlying large language models—as interacting particle systems [18]. This has already proved fruitful for encoder Transformers, which are exchangeable particle systems, where empirically observed representation collapse phenomena [41, 12, 31, 11] can be rigorously proven and seen as clustering of the particles over time [17, 18, 9, 33, 16, 6]. Results of this nature have since extended to substantially more general settings [7, 29, 15, 28, 1, 19, 8].

We leverage this perspective for decoder Transformers, for which less is known. More precisely, we consider a minimal model on the one-dimensional torus  $\mathbb{T} := \mathbb{R}/2\pi\mathbb{Z}$ , structurally faithful to the ones used in the experiments of [30]. Namely, at layer  $t \geq 0$ , the  $j$ -th token embedding, for  $1 \leq j \leq N$ , evolves as

$$\frac{d}{dt}\theta_j(t) = \frac{1}{\mathfrak{X}_{N,j}} \sum_{k=1}^{j-1} e^{-\frac{\lambda}{N}(j-k)} \mathbf{w}'_{\beta}(\theta_j(t) - \theta_k(t)), \quad (1.1)$$

with normalization

$$\mathfrak{X}_{N,j} := \sum_{k=1}^{j-1} e^{-\frac{\lambda}{N}(j-k)}, \quad (1.2)$$

and interaction kernel

$$\mathbf{w}_{\beta}(\theta) := e^{\beta \cos \theta}, \quad (1.3)$$

for fixed  $\beta > 0$  and  $\lambda \in \mathbb{R}$ . This corresponds to the “USA” dynamics of [17], but in a causal form, as in decoder architectures, and it incorporates the positional encoding factor  $e^{-\lambda(j-k)/N}$  known as ALiBi [34]. In particular, causality makes the system non-exchangeable. We focus on this intentionally caricatural setting for clarity and simplicity: it is not meant to represent full Transformers, which are high-dimensional and involve trainable matrices, but models of this type have been remarkably predictive of the behavior of trained Transformers, as illustrated for instance in [18, Figure 1], and they are rigorously justified at the initialization of training [29, 6, 15]. The reduced system (1.1) already retains the two mechanisms central to the present study: early tokens are repeatedly reused by later ones, while recent tokens are favored by the positional bias. Our analysis also extends to general smooth interaction kernels and causal positional encodings, see Section 1.2.1.

To connect the dynamics with plots such as Figure 1, we use a minimal decoder, which is a stripped-down version of the retrieval task in [30]: one singles out a relevant source position and asks whether the final output token recovers it. Fix a vocabulary size  $M \geq 2$ , say  $\mathcal{V} = \{0, 1, \dots, M-1\}$ , and let

$$\vartheta_m := \frac{2\pi m}{M} \in \mathbb{T}, \quad m \in \mathcal{V}.$$

We encode an input token  $m_i$  at position  $i$  by  $\theta_i(0) = \vartheta_{m_i}$ . Given the last hidden state  $\theta_N(t)$ , the decoder returns the nearest codeword

$$\hat{m}_N(t) := \arg \min_{m \in \mathcal{V}} |\theta_N(t) - \vartheta_m|_{\mathbb{T}}.$$

Since one should think of many prompts, or equivalently of many possible initial conditions, the relevant observable is an averaged retrieval score over a statistical ensemble of prompts. For a distinguished source position  $1 \leq i_* = \lfloor \sigma_0 N \rfloor \leq N$ , with  $\sigma_0 \in (0, 1)$ , we therefore define the prediction accuracy by

$$\mathbb{P}[\hat{m}_N(t) = m_{i_*}] = \mathbb{E} \left[ \mathbf{1}_{\{|\theta_N(t) - \vartheta_{m_{i_*}}(0)|_{\mathbb{T}} \leq \frac{\pi}{M}\}} \right], \quad (1.4)$$

where the expectation is taken with respect to the random ensemble of prompts  $\{\theta_i(0)\}_i$ . For analytical convenience, we mollify the characteristic function, e.g. using a periodic Gaussian, thus defining the *soft accuracy*

$$\mathfrak{A}_N(t, \sigma_0) := \mathbb{E} \sum_{k \in \mathbb{Z}} \exp \left( -\frac{M^2}{2\pi^2} (\theta_N(t) - \vartheta_{i_*}(0) - 2\pi k)^2 \right). \quad (1.5)$$

By the Poisson summation formula, this admits the Fourier representation

$$\mathcal{A}_N(t, \sigma_0) = \frac{\sqrt{\pi/2}}{M} \sum_{n \in \mathbb{Z}} e^{-\frac{\pi^2}{2M^2} n^2} \mathbb{E} \left[ e^{in(\theta_N(t) - \theta_{i_*}(0))} \right]. \quad (1.6)$$

Our goal is to prove that this quantity satisfies the U-shape observed in [30], see Figure 1. This amounts to analyzing the large- $N$  behavior of the expectations  $\mathbb{E}[e^{in(\theta_N(t) - \theta_{i_*}(0))}]$  as the source location  $i_*$  varies.

We approach this problem using tools from kinetic theory. A well-developed framework exists for deriving continuum descriptions from interacting particle systems: in the mean-field regime, this originates in propagation of chaos and the associated McKean–Vlasov limits (see e.g. the reviews [20, 26], as well as recent advances for singular interactions, e.g. [36, 35, 5, 4, 3]). More recently, the theory has been extended, in combination with graph limit techniques, to non-exchangeable systems (see e.g. [25]). The model (1.1) considered here, however, falls outside this setting: the interaction weights do not satisfy the key boundedness requirement of [25, Assumption (3)], but instead obey

$$\sup_{1 \leq k \leq N} \sum_{j=1}^N \omega_{j,k} \simeq \log N, \quad (1.7)$$

reflecting the singular cumulative influence of small indices on later particles.

Beyond this structural issue, the retrieval observable (1.6) exhibits a more fundamental feature: its spatial dependence does not appear at the mean-field level. Capturing it therefore requires going beyond propagation of chaos and retaining the leading-order time correlations, which is the main objective of this work. For exchangeable systems, the analysis of such correlations has seen significant recent progress, with both trajectorial approaches [13] and hierarchy-based methods [32, 23, 14]. By contrast, results for genuinely non-exchangeable systems remain scarce. In the present setting (1.1), we adapt the convenient trajectorial approach of [13], based on Glauber calculus with respect to initial data. A key feature of our analysis is that the singular scaling (1.7) induces nonstandard corrections to the propagation of chaos.

**1.1. Main results.** We define the empirical measure

$$\mu_N(t, \sigma, \theta) = \frac{1}{N} \sum_{j=1}^N \delta_{(\frac{j}{N}, \theta_j(t))}(\sigma, \theta) \in L^\infty(\mathbb{R}_{\geq 0}; \mathcal{P}((0, 1) \times \mathbb{T})), \quad (1.8)$$

where the rescaled index  $\sigma = \frac{j}{N} \in (0, 1]$  plays the role of a continuous reading-cursor variable in the limit. Next, motivated by the positional encoding factor  $\frac{1}{\mathfrak{X}_{N,j}} e^{-\lambda(j-k)/N} \mathbb{1}_{k < j}$  in (1.1), we introduce the associated limiting directed graphon

$$k_\lambda(\sigma, \sigma') := \frac{\lambda e^{-\lambda(\sigma - \sigma')}}{1 - e^{-\lambda\sigma}} \mathbb{1}_{\sigma' < \sigma}, \quad (1.9)$$

which is understood by continuity as  $\sigma^{-1} \mathbb{1}_{\sigma' < \sigma}$  when  $\lambda = 0$ . Finally, denote by  $\hat{f}(n) := \int_{\mathbb{T}} e^{-in\theta} f(\theta) d\theta$  the Fourier transform on  $\mathbb{T}$ , for  $n \in \mathbb{Z}$ , and write  $\langle n \rangle := \sqrt{1 + n^2}$ .

Our first result describes the mean-field limit of the system as  $N \rightarrow \infty$ . We stress that even the qualitative convergence in part (i) does not follow from existing mean-field results for non-exchangeable systems, since the interaction weights in (1.1) fail to satisfy the boundedness assumption of [25, Assumption (3)]; see (1.7).

**Theorem 1.1.**

(i) Qualitative mean-field limit:

If initially  $\mu_N|_{t=0} \xrightarrow{*} f_\circ$  in  $\mathcal{P}((0, 1) \times \mathbb{T})$ , then we have

$$\mu_N(t) \xrightarrow{*} f(t) \quad \text{for all } t \geq 0,$$

where the limit  $f$  is the unique weak solution in  $L^\infty(\mathbb{R}_{\geq 0}; \mathcal{P}((0, 1] \times \mathbb{T}))$  of the kinetic equation

$$\begin{cases} \partial_t f(t, \sigma, \theta) = -\partial_\theta \left( f(t, \sigma, \theta) \int_0^\sigma k_\lambda(\sigma, \sigma') (w'_\beta *_\theta f)(t, \sigma', \theta) d\sigma' \right), \\ f|_{t=0} = f_\circ. \end{cases} \quad (1.10)$$

(ii) Error estimates:

Assume that the initial data  $(\theta_j^0)_{1 \leq j \leq N}$  are independent and that their distribution converges to some limit profile  $f_\circ \in C^0([0, 1]; \mathcal{P}(\mathbb{T}))$  in the following sense: for some  $\delta > 0$  and  $C, \gamma < \infty$ ,

$$\left| \mathbb{E} \left[ e^{in\theta_{\lceil N\sigma \rceil}^0} \right] - \hat{f}_\circ(\sigma, n) \right| \leq CN^{-\delta} \langle n \rangle^\gamma \quad \text{for all } \sigma \in (0, 1] \text{ and } n \in \mathbb{Z}. \quad (1.11)$$

Then, for all  $t \geq 0$ ,  $\varphi \in C^\infty([0, 1] \times \mathbb{T})$ , and  $m > \gamma \vee 2 + \frac{1}{2}$ ,

$$\mathbb{E} \left[ \left| \int_{(0,1] \times \mathbb{T}} \varphi(\mu_N(t) - f(t)) \right|^2 \right]^{\frac{1}{2}} \lesssim N^{-\delta \wedge \frac{1}{2}} e^{Ct} \|\varphi\|_{L^\infty([0,1]; W^{m,\infty}(\mathbb{T}))}. \quad (1.12)$$

We illustrate the meaning of the assumptions on the initial data. Fix a vocabulary of size  $M \geq 2$ , say  $\mathcal{V} = \{0, 1, \dots, M-1\}$ , fix an encoding  $\vartheta_m := \frac{2\pi m}{M} \in \mathbb{T}$  for  $m \in \mathcal{V}$ , and fix  $p_m \in C^1([0, 1])$  such that for all  $\sigma$  the map  $m \mapsto p_m(\sigma)$  is a probability distribution on the vocabulary,  $\sum_{m \in \mathcal{V}} p_m(\sigma) = 1$ . Say we choose this distribution so that early positions favor words such as **please**, **record**, names such as **MICHELA**, **MARCO**, middle positions favor code tokens such as **zero**,  $\dots$ , **nine**, and late positions favor words such as **city**, **PARIS**, **answer**, **no**. Now consider initial prompts constructed as follows: for each length  $N$ , sample independently  $m_i \sim p(\cdot/N)$  and set  $\theta_i(0) = \vartheta_{m_i}$ ; a typical realization may look like **please record MICHELA code zero one seven six four two city PARIS answer no**. In this setting, the variables  $\theta_i(0)$  are independent by construction and the assumption (1.11) follows from the  $C^1$ -regularity of the maps  $(p_m)_m$ , with  $f_\circ(\sigma, \theta) = \sum_{m \in \mathcal{V}} p_m(\sigma) \delta_{\vartheta(m)}(\theta)$ . In the most homogeneous version of this baseline, the initial tokens may even be taken iid and uniformly distributed on  $\mathbb{T}$ , so that (1.11) holds with  $f_\circ \equiv 1$ .

As is typical for mean-field limits of non-exchangeable systems (see e.g. [25, Eqn (5)]), the limit equation (1.10) takes the form of a “layered” McKean–Vlasov equation. In the present case, this structure reflects the causal nature of the dynamics: the macroscopic variable  $\sigma$  quantifies how much of the past is accessible to a given token, while the kernel  $k_\lambda$  encodes the positional bias. However, the mean-field limit alone does not yield useful information on the accuracy function (1.5) in the retrieval task. Indeed, the relevant signal is carried by time correlations  $\text{Cov}(e^{in\theta_N(t)}, e^{-i\theta_{i_*}(0)})$ , which vanish at the level of the mean-field description. Capturing the spatial dependence of the accuracy function therefore requires analyzing the next order in the propagation of chaos.

To this end, for a test function  $\varphi \in C^\infty(\mathbb{T})$ , we introduce the autocorrelation  $A_\varphi^N$  and the cross-correlation  $C_\varphi^N$  as follows: for  $t \geq 0$ ,  $\sigma, \sigma_0 \in (0, 1]$ , and  $\psi \in C^\infty(\mathbb{T})$ ,

$$\int \psi(\theta) A_\varphi^N(t, \sigma, \theta) d\theta := \text{Cov} \left( \psi(\theta_{\lceil N\sigma \rceil}(t)), \varphi(\theta_{\lceil N\sigma \rceil}^0) \right), \quad (1.13)$$

$$\int \psi(\theta) C_\varphi^N(t, \sigma, \theta; \sigma_0) d\theta := \text{Cov} \left( \psi(\theta_{\lceil N\sigma \rceil}(t)), \varphi(\theta_{\lceil N\sigma_0 \rceil}^0) \right) \mathbb{1}_{\lceil N\sigma_0 \rceil < \lceil N\sigma \rceil}. \quad (1.14)$$

The autocorrelation  $A_\varphi^N$  quantifies the memory of a token at position  $\sigma$  at time  $t$  with respect to its own initial value, and is of order  $O(1)$ . By contrast, the cross-correlation  $C_\varphi^N$  measures the dependence of a token at position  $\sigma$  at time  $t$  with respect to the initial value of an earlier token at a different position  $\sigma_0 < \sigma$ ; it is of order  $O(N^{-1})$ . It is precisely this cross-correlation that carries the information on the spatial dependence of the accuracy function. In the following result, we characterize the leading behavior of both  $A_\varphi^N$  and  $C_\varphi^N$ .

**Theorem 1.2.** *Assume that the initial data  $(\theta_j^0)_{1 \leq j \leq N}$  are independent and that their distribution converges to some limit profile  $f_\circ \in C^0([0, 1]; \mathcal{P}(\mathbb{T}))$  in the sense of (1.11), for some  $\delta > 0$ . Then,*

given  $\zeta \leq \delta$  with  $\zeta < 1$ , we have for all  $t \geq 0$ ,  $\sigma, \sigma_0 \in (0, 1]$ ,  $\sigma_0 < \sigma$ , and frequency  $n \in \mathbb{Z}$ ,

$$\left| \left( \hat{A}_\varphi^N - \hat{A}_\varphi \right) (t, \sigma, n) \right| \lesssim_{\zeta, \varphi} \frac{1}{N^\zeta \sigma^2} \langle n \rangle^C e^{Ct}, \quad (1.15)$$

$$\left| \left( N \hat{C}_\varphi^N - \hat{C}_\varphi \right) (t, \sigma, n; \sigma_0) \right| \lesssim_{\zeta, \varphi} \frac{1}{N^\zeta \sigma_0^2} \langle n \rangle^C e^{Ct}, \quad (1.16)$$

where  $A_\varphi, C_\varphi$  are the unique solutions of the limit equations

$$\begin{cases} \partial_t A_\varphi(t, \sigma, \theta) = -\partial_\theta \left( A_\varphi(t, \sigma, \theta) \int_0^\sigma k_\lambda(\sigma, \sigma') (w'_\beta * f)(t, \sigma', \theta) d\sigma' \right), \\ A_\varphi(0, \sigma, \theta) = f_\circ(\sigma, \theta) \left( \varphi(\theta) - \int_{\mathbb{T}} \varphi f_\circ(\sigma, \cdot) d\theta \right), \end{cases} \quad (1.17)$$

$$\begin{cases} \partial_t C_\varphi(t, \sigma, \theta; \sigma_0) = -\partial_\theta \left( f(t, \sigma, \theta) \int_{\sigma_0}^\sigma k_\lambda(\sigma, \sigma') (w'_\beta * C_\varphi)(t, \sigma', \theta; \sigma_0) d\sigma' \right) \\ \quad -\partial_\theta \left( C_\varphi(t, \sigma, \theta; \sigma_0) \int_0^\sigma k_\lambda(\sigma, \sigma') (w'_\beta * f)(t, \sigma', \theta) d\sigma' \right) \\ \quad -\partial_\theta \left( f(t, \sigma, \theta) k_\lambda(\sigma, \sigma_0) (w'_\beta * A_\varphi)(t, \sigma_0, \theta) \right), \\ C_\varphi(0, \sigma, \theta; \sigma_0) = 0. \end{cases} \quad (1.18)$$

The limiting autocorrelation  $A_\varphi$  is transported by the mean-field flow, while the cross-correlation  $C_\varphi$  evolves according to the linearization of the mean-field dynamics, with a forcing term determined by  $A_\varphi$ . This forcing injects the fluctuation carried by the source token at position  $\sigma_0$  into the dynamics of  $C_\varphi$ , and is weighted by the limiting graphon  $k_\lambda(\sigma, \sigma_0)$ , thereby encoding the spatial influence of the source point. This mechanism is ultimately responsible for the dependence of the retrieval profile on the source position  $\sigma_0$ .

The proof relies on a non-hierarchical cumulant expansion adapted to the triangular causal structure of the dynamics. More precisely, differentiating a two-point correlation in time produces third-order cumulants, so that no closed equation is available at the level of two-point correlations. To control these higher-order terms, we appeal to Glauber calculus with respect to the initial data, following the approach of [13]. In essence, quantitative bounds on first- and second-order Glauber derivatives yield sharp control of the third-cumulant remainder, allowing one to truncate the expansion at the level of two-point correlations without resorting to a BBGKY hierarchy.

Compared with previous uses of Glauber calculus to quantify corrections to mean-field limits, notably [13], several new ingredients are required. First, the analysis must be adapted to a causal, non-exchangeable decoder dynamics, where the token label persists in the limit as the macroscopic variable  $\sigma$ , and the dependency graph is triangular rather than symmetric. Second, the interaction scale at token  $j$  is of order  $j^{-1}$  rather than  $N^{-1}$ , leading to a cumulative influence of early indices on later ones and to the singular behavior (2.2). This disrupts standard propagation-of-chaos mechanisms and gives rise to nonstandard correction terms in the correlation estimates; see Lemma 3.1.

In the specific case of iid uniformly distributed prompts, thus satisfying (1.11) with  $f_\circ \equiv 1$ , the mean-field solution remains constant  $f \equiv 1$ , the autocorrelation reduces to  $A_\varphi \equiv \varphi - \int_{\mathbb{T}} \varphi$ , and the cross-correlation equation (1.18) becomes

$$\begin{cases} \partial_t C_\varphi(t, \sigma, \theta; \sigma_0) = - \left( \int_{\sigma_0}^\sigma k_\lambda(\sigma, \sigma') (w''_\beta * C_\varphi)(t, \sigma', \theta; \sigma_0) d\sigma' + k_\lambda(\sigma, \sigma_0) (w''_\beta * \varphi)(\theta) \right), \\ C_\varphi(0, \sigma, \theta; \sigma_0) = 0. \end{cases}$$

Taking the Fourier transform on  $\mathbb{T}$ , this linear equation diagonalizes and we obtain

$$\hat{C}_\varphi(t, \sigma, n; \sigma_0) = \hat{\varphi}(n) g_{a_n}(t, \sigma; \sigma_0),$$

where

$$a_n := n^2 \hat{w}_\beta(n),$$

and where for  $a \in \mathbb{R}$  we define the profile  $g_a$  as the solution of the following Volterra–Hardy equation,

$$\begin{cases} \partial_t g_a(t, \sigma; \sigma_0) - a \left( \kappa_\lambda(\sigma, \sigma_0) + \int_{\sigma_0}^{\sigma} \kappa_\lambda(\sigma, \sigma') g_a(t, \sigma'; \sigma_0) d\sigma' \right) = 0, \\ g_a(0, \sigma; \sigma_0) = 0. \end{cases} \quad (1.19)$$

By Theorem 1.2 and the Fourier representation (1.6), using this representation for limiting correlations, we deduce the following approximation for the soft accuracy:

$$\mathfrak{A}_N(t, \sigma_0) = \frac{\sqrt{\pi/2}}{M} + \frac{\sqrt{2\pi}}{MN} \mathfrak{S}_t(\sigma_0) + O\left(N^{-1-\zeta} \sigma_0^{-2} M^C e^{Ct}\right), \quad (1.20)$$

where the leading correction is given by

$$\mathfrak{S}_t(\sigma_0) := \sum_{n \geq 1} e^{-\frac{\pi^2}{2M^2} n^2} g_{a_n}(t, 1; \sigma_0). \quad (1.21)$$

We show that this quantity is indeed U-shaped; see also Figure 2.

**Theorem 1.3** (Lost in the middle). *Let  $\lambda > 0$  and assume that<sup>1</sup>*

$$t \sup_{n \geq 1} a_n \leq \min \left\{ 3 - \sqrt{3}, 2 \left( 1 - e^{-\lambda} \right) \right\}. \quad (1.22)$$

*Then the following hold.*

- (i) Primacy:  $\mathfrak{S}_t(\sigma_0) \rightarrow +\infty$  as  $\sigma_0 \downarrow 0$ .
- (ii) Recency:  $\mathfrak{S}'_t(1^-) > 0$ .
- (iii) U-shape: *The function  $\sigma_0 \mapsto \mathfrak{S}_t(\sigma_0)$  has a unique global minimum in  $(0, 1)$ .*

*Consequently, by (1.20), in the case  $f_\circ \equiv 1$ , the soft accuracy  $\sigma_0 \mapsto \mathfrak{A}_N(t, \sigma_0)$  is U-shaped with a unique interior minimum for all  $N$  large enough.*

Results on the emergence of primacy and recency biases for Transformers appear in recent works dealing with simplified or discrete-time models, but these use totally different approaches and methods—see [39, 22, 10].

The proof of Theorem 1.3 proceeds by reducing the Volterra–Hardy equation (1.19) to a Goursat problem via a change of variables, which can be solved explicitly in terms of modified Bessel functions. The crucial computation is:

**Proposition 1.4.** *For  $a > 0$ ,  $\lambda \in \mathbb{R}$ , and  $0 < \sigma_0 < \sigma \leq 1$ , the unique solution of (1.19) is given by (1.23) when  $\lambda \neq 0$  and by (1.25) when  $\lambda = 0$ :*

$$g_a(t, \sigma; \sigma_0) = \frac{\lambda e^{-\lambda(\sigma-\sigma_0)}}{1 - e^{-\lambda\sigma}} \sqrt{\frac{at}{Y(\sigma; \sigma_0)}} I_1 \left( 2\sqrt{at Y(\sigma; \sigma_0)} \right), \quad (1.23)$$

where

$$Y(\sigma; \sigma_0) := \log \frac{e^{\lambda\sigma} - 1}{e^{\lambda\sigma_0} - 1}. \quad (1.24)$$

For  $\lambda = 0$ , this reduces to

$$g_a(t, \sigma; \sigma_0) = \frac{1}{\sigma} \sqrt{\frac{at}{\log(\sigma/\sigma_0)}} I_1 \left( 2\sqrt{at \log(\sigma/\sigma_0)} \right). \quad (1.25)$$

**1.2. Extensions.** Our results extend in several natural directions. First, the specific choice of the interaction kernel  $w_\beta(\theta) = e^{\beta \cos \theta}$  plays no essential role in the analysis: it can be replaced by an arbitrary smooth and even periodic interaction kernel. Likewise, the arguments carry over to higher-dimensional phase spaces. We next briefly discuss extensions with respect to positional encodings and to the more general causal self-attention dynamics.

<sup>1</sup>As  $w_\beta$  is smooth, we have  $a_n \rightarrow 0$  as  $n \rightarrow \infty$ , so this condition is non-empty for  $t > 0$ .

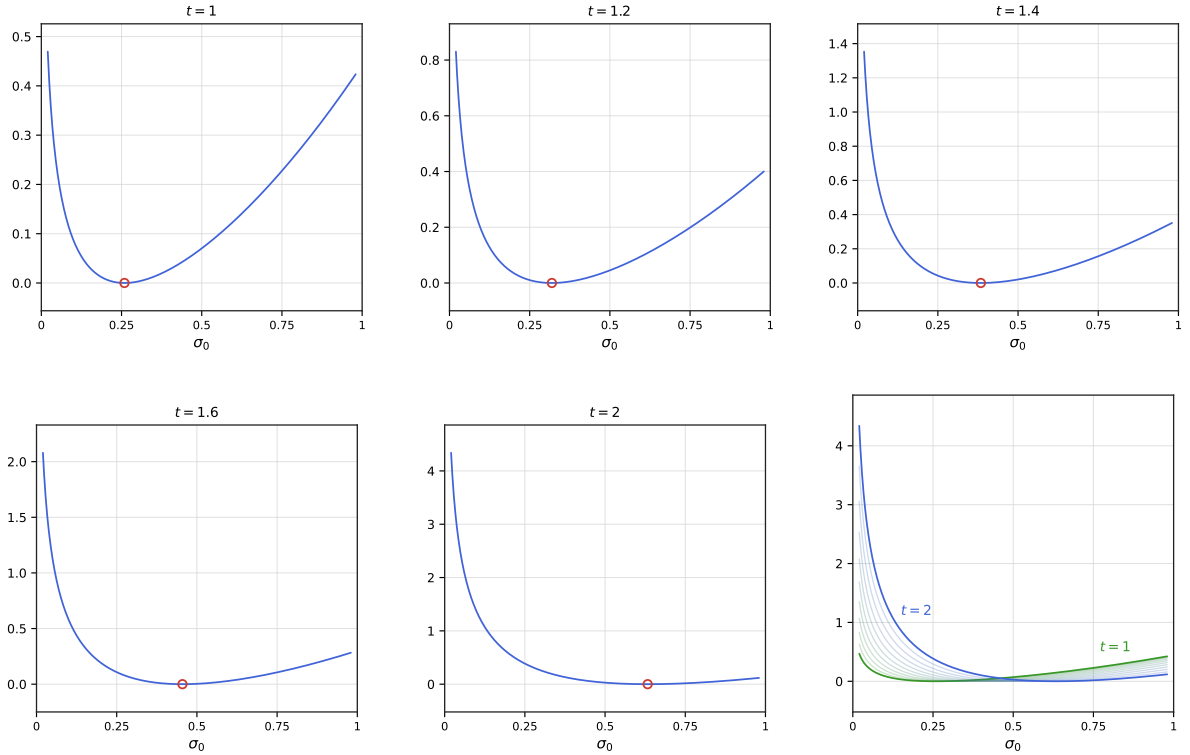


FIGURE 2. The profile predicted by Theorem 1.3 in the explicit regime  $\beta = \lambda = 1$  and  $M = 8$ . Each panel plots the centered correction  $\sigma_0 \mapsto \mathcal{S}_t(\sigma_0) - \min \mathcal{S}_t$  from (1.21).

1.2.1. *On positional encodings.* We focus on the ALiBi encoding factor for simplicity, but the proofs of Theorems 1.1 and 1.2 extend with only minor modifications to more general causal weights of the form  $b((j-k)/N)$ , provided  $b$  is at least continuous on  $[0, 1]$ . In that case, the limiting kernel  $\mathbf{k}_\lambda(\sigma, \sigma_0)$  is replaced by

$$\frac{b(\sigma, \sigma_0)}{\int_0^\sigma b(\sigma, \sigma') d\sigma'} \mathbb{1}_{\sigma_0 < \sigma}.$$

The proof of Theorem 1.3 adapts accordingly, with the parameter  $\lambda$  replaced by the local slope  $b'(0)$ .

Avoiding the resulting U-shaped profile therefore should require modifying the positional encoding so as to weaken this monotone recency mechanism or to introduce a competing phase-sensitive effect, as done in empirical works [27, 24]. RoPE-type encodings [37] provide a natural example, but fall beyond the present theory because the positional dependence then acts inside the interaction phase rather than through a scalar causal kernel, so that the Volterra–Hardy reduction no longer applies.

1.2.2. *General self-attention.* The reduced model (1.1) can be viewed as a simplification of the  $d = 2$ ,  $Q = K = V \equiv \text{Id}$  specialization of the causal spherical self-attention dynamics with ALiBi encoding:

$$\frac{d}{dt} x_j(t) = \frac{1}{\mathfrak{F}_{N,j}(t)} \sum_{k=1}^{j-1} \exp\left(\beta \langle Q(t)x_j(t), K(t)x_k(t) \rangle - \frac{\lambda}{N}(j-k)\right) \mathbf{P}_{x_j(t)}^\perp(V(t)x_k(t)), \quad (1.26)$$

where  $x_j(t) \in \mathbb{S}^{d-1}$ ,  $\mathbf{P}_x^\perp y := y - \langle x, y \rangle x$ , and with normalization

$$\mathfrak{F}_{N,j}(t) := \sum_{k=1}^{j-1} \exp\left(\beta \langle Q(t)x_j(t), K(t)x_k(t) \rangle - \frac{\lambda}{N}(j-k)\right).$$

Note that the normalization further depends on the strength of interactions, which we removed for simplicity in (1.1). A full analysis of (1.26) is left for future work.

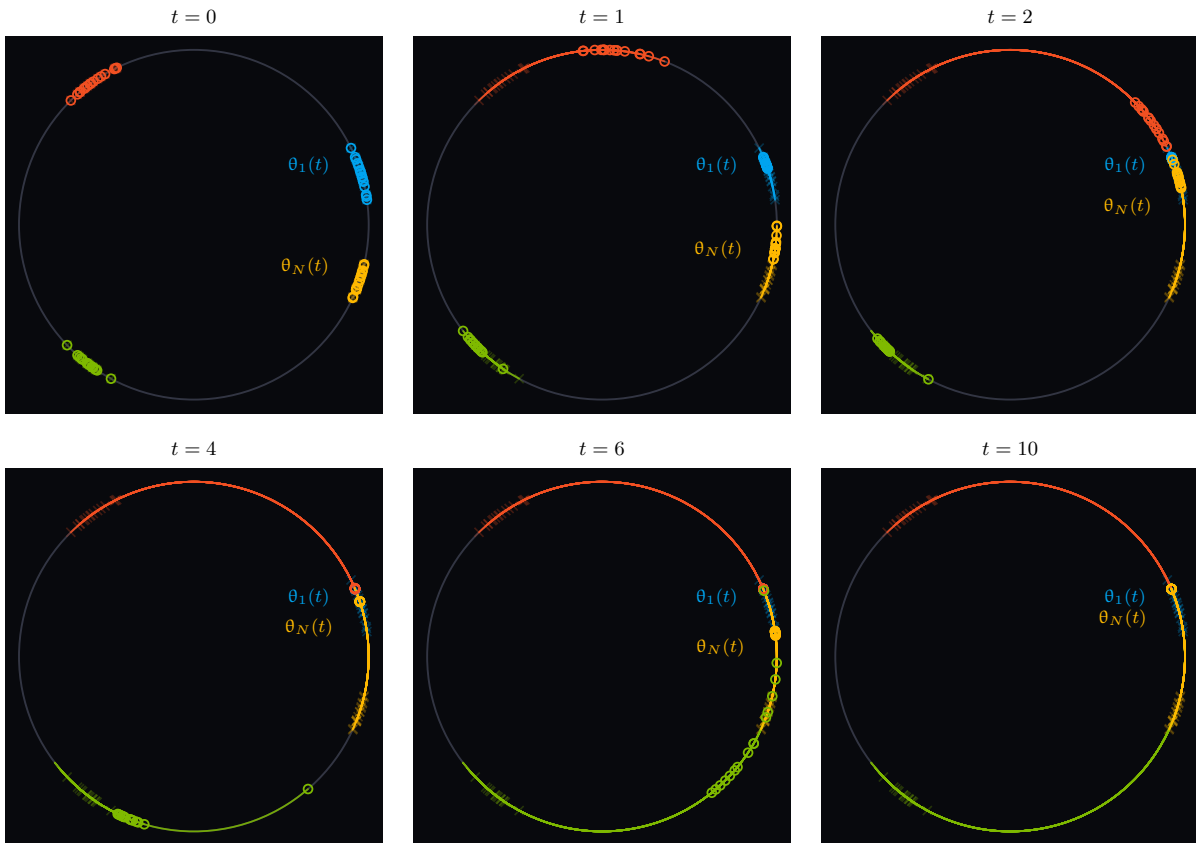


FIGURE 3. Simulation of the particle system (1.1) for  $N = 64$  and  $\beta = \lambda = 1$ , initialized near four small angular clusters. Crosses mark the initial angles  $\theta_k(0)$ , while hollow points mark the evolved angles  $\theta_j(t)$ ; the labels follow the first and last positions. The plot is read by comparing evolved angles with initial ones: for a fixed output position  $j$ , proximity of  $\theta_j(t)$  to  $\theta_k(0)$  is the trajectory-level analogue of alignment with source position  $k$ , and for  $j = N$  this is essentially the distance entering the retrieval observable. Two effects are visible. The drift of the terminal particle  $\theta_N(t)$  toward the first particle  $\theta_1(t)$  shows the primacy effect; collapse toward the first token is proved in [28], and is consistent with empirical work on attention sinks [40, 21, 2]. All the while, at short times the terminal particle remains close to the packet of nearby high-index particles initialized near  $\theta_N(0)$ ; this coherent terminal packet somewhat reflects the recency side of the U-shaped profile.

**1.3. Organization.** The paper is organized as follows. Section 2 records some elementary convergence estimates for interaction weights. Section 3 develops the cumulant estimates that form the backbone of all our convergence proofs: using Glauber calculus adapted to the triangular causal structure of the dynamics, we derive sharp bounds on covariances and third cumulants of particle trajectories. Section 4 contains the proof of Theorem 1.1 on the mean-field limit. Section 5 gives the proof of Theorem 1.2 on the characterization of correlations. Finally, Section 6 completes the proof of Theorem 1.3.

2. INTERACTION WEIGHTS AND LIMITING GRAPHON

We introduce the following short-hand notation for the non-exchangeable, directed interaction weights appearing in the model (1.1): for  $1 \leq k, j \leq N$ , set

$$\omega_{j,k} := \frac{e^{-\frac{\lambda}{N}(j-k)} \mathbf{1}_{k < j}}{\mathfrak{X}_{N,j}}, \quad (2.1)$$

where the normalization  $\mathfrak{X}_{N,j}$  ensures  $\sum_{k=1}^N \omega_{j,k} = 1$  for  $1 < j \leq N$ . A direct computation yields, uniformly for all  $1 \leq k, j \leq N$ ,

$$\omega_{j,k} = \frac{e^{\frac{\lambda}{N}k} \mathbf{1}_{k < j}}{\sum_{m=1}^{j-1} e^{\frac{\lambda}{N}m}} \simeq j^{-1} \mathbf{1}_{k < j}.$$

In particular, this implies that the weights are row-stochastic but not column-balanced, as already emphasized in (1.7):

$$\sup_{1 \leq k \leq N} \sum_{j=1}^N \omega_{j,k} \simeq \log N. \quad (2.2)$$

This logarithmic divergence reflects the cumulative influence of small indices on later particles and places the model outside the standard framework for mean-field/graph limits, namely beyond the minimal requirements of [25, Assumption (3)].

We next introduce a continuum representation of this interaction structure. Define the rescaled interpolated kernel

$$\mathsf{K}_N(\sigma, \sigma') := N \omega_{\lceil N\sigma \rceil, \lceil N\sigma' \rceil}, \quad \sigma, \sigma' \in (0, 1]. \quad (2.3)$$

This can be viewed as a continuum interaction kernel associated with the underlying directed weighted graph (a directed graphon in the sense of graph limit theory). By construction, it satisfies

$$\int_0^1 \mathsf{K}_N(\sigma, \sigma') d\sigma' = 1, \quad \sigma \in \left(\frac{1}{N}, 1\right],$$

and the pointwise uniform bound

$$\mathsf{K}_N(\sigma, \sigma') \lesssim \sigma^{-1} \mathbf{1}_{\sigma' < \sigma}.$$

It is easily checked that  $\mathsf{K}_N$  converges almost everywhere to the limiting kernel  $\mathsf{k}_\lambda$  defined in (1.9). The latter satisfies the same normalization and pointwise bound,

$$\int_0^1 \mathsf{k}_\lambda(\sigma, \sigma') d\sigma' = 1, \quad \mathsf{k}_\lambda(\sigma, \sigma') \lesssim \sigma^{-1} \mathbf{1}_{\sigma' < \sigma},$$

while the logarithmic divergence in (2.2) is reflected by the singular behavior

$$\int_0^1 \mathsf{k}_\lambda(\sigma, \sigma') d\sigma \sim |\log \sigma'|, \quad \sigma' \downarrow 0,$$

which will create technical difficulties in the analysis. Finally, we quantify the convergence  $\mathsf{K}_N \rightarrow \mathsf{k}_\lambda$ . For  $\lambda \neq 0$ , using the geometric-series identity

$$\mathfrak{X}_{N,j} = \frac{1 - e^{-\frac{\lambda}{N}(j-1)}}{e^{\frac{\lambda}{N}} - 1},$$

we obtain, uniformly for  $1 \leq k < j \leq N$ ,

$$N \omega_{j,k} = \frac{N(e^{\frac{\lambda}{N}} - 1)}{1 - e^{-\frac{\lambda}{N}(j-1)}} e^{-\frac{\lambda}{N}(j-k)} = \mathsf{k}_\lambda\left(\frac{j}{N}, \frac{k}{N}\right) + O\left(\frac{1}{N(\frac{j}{N})^2}\right).$$

For  $\lambda = 0$ , the same estimate is immediate from

$$N \omega_{j,k} = \frac{N}{j-1} = \frac{1}{j/N} + O\left(\frac{1}{N(\frac{j}{N})^2}\right).$$

Consequently, we can deduce for all  $\sigma \in (0, 1]$ ,

$$\int_0^1 |\mathsf{K}_N(\sigma, \cdot) - \mathsf{k}_\lambda(\sigma, \cdot)| \leq \frac{C}{N\sigma + 1}. \quad (2.4)$$

This convergence estimate will be used repeatedly in the sequel.

### 3. CUMULANT BOUNDS

The mean-field limit and the characterization of correlations both rely on quantitative control of the statistical dependence between particle trajectories. Provided that the initial data are independent, we can use the so-called Glauber calculus of [13] to estimate correlations. However, compared with [13], some important care is needed in the present non-exchangeable setting: the smallness of the interaction at particle  $j$  is of order  $j^{-1}$  rather than  $N^{-1}$ , which creates difficulties when  $j$  is small, as reflected for instance by (2.2). For this reason, nontrivial corrections appear in correlation estimates, see (3.3) and (3.4) below, but they are harmless when applied to macroscopic indices  $j = \lceil N\sigma \rceil$  with  $\sigma > 0$  bounded away from zero.

We start by recalling tools from Glauber calculus with respect to independent initial data. Let  $\nu_j := \text{Law}(\theta_j^0)$  and let  $(\Omega, \mathcal{F}, \mathbb{P}) := \bigotimes_{j=1}^N (\mathbb{T}, \mathcal{B}(\mathbb{T}), \nu_j)$  be the product probability space carrying the independent initial data. We identify  $\theta_j^0$  with the  $j$ -th coordinate map on  $\Omega$ . For each  $k$ , we denote by

$$\mathbb{E}_k^\circ[Y] := \mathbb{E}[Y \mid (\theta_j^0)_{j:j \neq k}]$$

the conditional expectation given all variables except  $\theta_k^0$ , or equivalently the integration with respect to  $\theta_k^0$ , and we define the Glauber derivative by

$$D_k^\circ Y := Y - \mathbb{E}_k^\circ[Y].$$

The associated Glauber Laplacian is

$$\mathcal{L}^\circ := \sum_{k=1}^N (D_k^\circ)^* D_k^\circ = \sum_{k=1}^N D_k^\circ,$$

and we write  $\mathcal{T}^\circ := (\mathcal{L}^\circ)^{-1}$  for its inverse defined on centered random variables. We shall only use the following standard properties, proved e.g. in [13, Lemmas 2.4(v) and 2.5]: for centered random variables  $Y, Y' \in L^2(\Omega)$  and all  $1 < p < \infty$ ,

$$\text{Cov}(Y, Y') = \sum_{k=1}^N \mathbb{E}[(D_k^\circ Y) \mathcal{T}^\circ(D_k^\circ Y')], \quad \|\mathcal{T}^\circ(D_k^\circ Y)\|_{L^p(\Omega)} \lesssim \|D_k^\circ Y\|_{L^p(\Omega)}, \quad (3.1)$$

hence

$$|\text{Cov}(Y, Y')| \lesssim \sum_{k=1}^N \|D_k^\circ Y\|_{L^2(\Omega)} \|D_k^\circ Y'\|_{L^2(\Omega)}. \quad (3.2)$$

Our main correlation estimates take on the following guise. For our purposes, it is enough to focus on second and third cumulants, but the proof extends easily to higher cumulants. Recall that the third joint cumulant is defined as  $\kappa_{1,1,1}(Y, Y', Y'') = \mathbb{E}[(Y - \mathbb{E}[Y])(Y' - \mathbb{E}[Y'])(Y'' - \mathbb{E}[Y''])]$ .

**Lemma 3.1.** *Assume that the initial data  $(\theta_j^0)_{1 \leq j \leq N}$  are independent. Then, for all  $\varphi \in C^2(\mathbb{T})$ ,  $t, t', t'' \in [0, T]$ , and all  $i > j > k$ ,*

$$|\text{Cov}(\varphi(\theta_i(t)), \varphi(\theta_j(t')))| \lesssim i^{-1} e^{\sqrt{CT \log(i/j)}} e^{CT} \|\varphi'\|_{L^\infty(\mathbb{T})}^2, \quad (3.3)$$

$$|\kappa_{1,1,1}(\varphi(\theta_i(t)), \varphi(\theta_j(t')), \varphi(\theta_k(t'')))| \lesssim (ij)^{-1} e^{\sqrt{CT \log(i/k)}} e^{CT} \|\varphi'\|_{L^\infty(\mathbb{T})}^2 \|\varphi'\|_{W^{1,\infty}(\mathbb{T})}. \quad (3.4)$$

*Proof.* We split the proof into five steps: we start by estimating the sensitivity of trajectories with respect to initial data in form of uniform bounds on Glauber derivatives, and we then conclude by reconstructing cumulants as sums of products of Glauber derivatives. Compared to the covariance, the third cumulant further requires bounds on second Glauber derivatives, which is postponed to the last two steps of the proof.

*Step 1.* First Glauber derivatives: proof that for all  $1 \leq i, k \leq N$  and  $t \geq 0$ , almost surely,

$$|D_k^\circ \theta_i(t)| \leq C e^{Ct} \mathbf{1}_{k=i} + C i^{-1} e^{\sqrt{Ct \log(i/k)}} e^{Ct} \mathbf{1}_{k < i}. \quad (3.5)$$

Note that for a Lipschitz function  $h$  we can bound

$$|D_k^\circ h(Y)| \leq \|h'\|_{L^\infty(\mathbb{T})} (|D_k^\circ Y| + \mathbb{E}[|D_k^\circ Y| \mid (\theta_j^\circ)_{j:j \neq k}]), \quad (3.6)$$

and thus, for any  $p \geq 1$ ,

$$\|D_k^\circ h(Y)\|_{L^p(\Omega)} \leq 2 \|h'\|_{L^\infty(\mathbb{T})} \|D_k^\circ Y\|_{L^p(\Omega)}. \quad (3.7)$$

Taking Glauber derivatives in the particle dynamics and using this latter estimate, we find

$$\frac{d}{dt} \|D_k^\circ \theta_i\|_{L^p(\Omega)} \leq 2 \|w_\beta''\|_{L^\infty(\mathbb{T})} \|D_k^\circ \theta_i\|_{L^p(\Omega)} + 2 \|w_\beta''\|_{L^\infty(\mathbb{T})} \sum_{j=1}^{i-1} \omega_{i,j} \|D_k^\circ \theta_j\|_{L^p(\Omega)}.$$

Integrating in time and recalling  $\omega_{i,j} \lesssim i^{-1}$ , this yields

$$\|D_k^\circ \theta_i(t)\|_{L^p(\Omega)} \leq e^{Ct} \|D_k^\circ \theta_i^\circ\|_{L^p(\Omega)} + i^{-1} \sum_{j=1}^{i-1} \int_0^t C e^{C(t-s)} \|D_k^\circ \theta_j(s)\|_{L^p(\Omega)} ds.$$

As initial data are independent, we note that  $D_k^\circ \theta_i^\circ = 0$  for  $k \neq i$ , and the triangular structure of the dynamics further ensures  $D_k^\circ \theta_i(t) = 0$  for all  $k > i$  and  $t \geq 0$ . The above then becomes

$$\|D_k^\circ \theta_i(t)\|_{L^p(\Omega)} \leq C e^{Ct} \mathbf{1}_{k=i} + i^{-1} \sum_{j=k}^{i-1} \int_0^t C e^{C(t-s)} \|D_k^\circ \theta_j(s)\|_{L^p(\Omega)} ds.$$

Iterating this estimate, we get for  $k < i$ ,

$$\|D_k^\circ \theta_i(t)\|_{L^p(\Omega)} \leq i^{-1} C e^{Ct} \sum_{r=0}^{i-k-1} \frac{(Ct)^r}{r!} \sum_{k < j_r < \dots < j_1 < i} (j_1 \dots j_r)^{-1},$$

and thus, after straightforward calculation,

$$\|D_k^\circ \theta_i(t)\|_{L^p(\Omega)} \leq i^{-1} C e^{Ct} \sum_{r=0}^{i-k-1} \frac{(Ct \log(i/k))^r}{r!^2} \leq i^{-1} C e^{Ct} e^{\sqrt{Ct \log(i/k)}},$$

that is, (3.5).

*Step 2.* Proof of (3.3).

Appealing to the Glauber covariance estimate (3.2), applying it to trajectories, and recalling the chain rule (3.7), we get

$$|\text{Cov}(\varphi(\theta_i(t)), \varphi(\theta_j(t')))| \lesssim \|\varphi'\|_{L^\infty(\mathbb{T})}^2 \sum_{k=1}^N \|D_k^\circ \theta_i(t)\|_{L^2(\Omega)} \|D_k^\circ \theta_j(t')\|_{L^2(\Omega)}. \quad (3.8)$$

Inserting (3.5), for  $i > j$ , we obtain

$$\begin{aligned} |\text{Cov}(\varphi(\theta_i(t)), \varphi(\theta_j(t')))| &\lesssim i^{-1} e^{\sqrt{Ct \log(i/j)}} e^{C(t+t')} \|\varphi'\|_{L^\infty(\mathbb{T})}^2 \\ &\quad + (ij)^{-1} e^{C(t+t')} \|\varphi'\|_{L^\infty(\mathbb{T})}^2 \sum_{k=1}^{j-1} e^{\sqrt{Ct \log(i/k)}} e^{\sqrt{Ct' \log(j/k)}}, \end{aligned}$$

and the claim (3.3) follows after a straightforward estimate of the last sum.

*Step 3.* Second Glauber derivatives: proof that for all  $1 \leq k < j < i \leq N$  and  $t \geq 0$ ,

$$|D_j^\circ D_k^\circ \theta_i(t)| \leq C(ij)^{-1} e^{\sqrt{Ct \log(i/k)}} e^{Ct}, \quad (3.9)$$

Arguing similarly as for (3.7), we find the following second-order chain rule: for a smooth function  $h$ , for all  $j \neq k$  and  $2 \leq p \leq \infty$ ,

$$\|D_j^\circ D_k^\circ h(Y)\|_{L^p(\Omega)} \leq 2\|h'\|_{L^\infty(\mathbb{T})} \|D_j^\circ D_k^\circ Y\|_{L^p(\Omega)} + 4\|h''\|_{L^\infty(\mathbb{T})} \|D_j^\circ Y\|_{L^p(\Omega)} \|D_k^\circ Y\|_{L^p(\Omega)}. \quad (3.10)$$

For  $k < j < i$ , taking second Glauber derivatives in the particle dynamics, using this estimate, and recalling  $\omega_{i,j} \lesssim i^{-1}$ , we find

$$\begin{aligned} \frac{d}{dt} \|D_j^\circ D_k^\circ \theta_i\|_{L^p(\Omega)} &\leq C \|D_j^\circ D_k^\circ \theta_i\|_{L^p(\Omega)} + C i^{-1} \sum_{l=1}^{i-1} \|D_j^\circ D_k^\circ \theta_l\|_{L^p(\Omega)} \\ &\quad + C i^{-1} \sum_{l=1}^{i-1} \|D_j^\circ (\theta_i - \theta_l)\|_{L^p(\Omega)} \|D_k^\circ (\theta_i - \theta_l)\|_{L^p(\Omega)}. \end{aligned}$$

Using (3.5) to estimate the last sum, we get

$$\begin{aligned} \frac{d}{dt} \|D_j^\circ D_k^\circ \theta_i\|_{L^p(\Omega)} &\leq C \|D_j^\circ D_k^\circ \theta_i\|_{L^p(\Omega)} + C i^{-1} \sum_{l=1}^{i-1} \|D_j^\circ D_k^\circ \theta_l\|_{L^p(\Omega)} \\ &\quad + C i^{-2} e^{\sqrt{Ct \log(i/k)}} e^{Ct} + C(ij)^{-1} e^{\sqrt{Ct \log(j/k)}} e^{Ct} \end{aligned}$$

and thus, after time integration,

$$\begin{aligned} \|D_j^\circ D_k^\circ \theta_i(t)\|_{L^p(\Omega)} &\leq C i^{-2} e^{\sqrt{Ct \log(i/k)}} e^{Ct} + C(ij)^{-1} e^{\sqrt{Ct \log(j/k)}} e^{Ct} \\ &\quad + i^{-1} \sum_{l=1}^{i-1} \int_0^t C e^{C(t-s)} \|D_j^\circ D_k^\circ \theta_l(s)\|_{L^p(\Omega)} ds. \end{aligned}$$

Note that the first term is bounded by the second one by a simple monotonicity argument. Further using (3.5) in the form

$$\|D_j^\circ D_k^\circ \theta_l(t)\|_{L^p(\Omega)} \leq \begin{cases} j^{-1} e^{\sqrt{Ct \log(j/k)}} e^{Ct} & : l = j > k, \\ 0 & : l = k < j, \end{cases}$$

the above can be reorganized more simply as

$$\|D_j^\circ D_k^\circ \theta_i(t)\|_{L^p(\Omega)} \leq C(ij)^{-1} e^{\sqrt{Ct \log(j/k)}} e^{Ct} + i^{-1} \sum_{j < l < i} \int_0^t C e^{C(t-s)} \|D_j^\circ D_k^\circ \theta_l(s)\|_{L^p(\Omega)} ds.$$

Now iterating this estimate, and evaluating the resulting series similarly as in Step 1, we find

$$\|D_j^\circ D_k^\circ \theta_i(t)\|_{L^p(\Omega)} \leq C(ij)^{-1} e^{\sqrt{Ct \log(j/k)}} e^{\sqrt{Ct \log(i/j)}} e^{Ct},$$

which is equivalent to the claim (3.9).

*Step 4.* Third-order Poincaré inequality for third joint cumulants: proof that for all random variables  $Y, Y', Y'' \in L^3(\Omega)$  we have

$$\begin{aligned} |\kappa_{1,1,1}[Y, Y', Y'']| &\leq \sum_{j,k=1}^N \left( \|D_j^\circ D_k^\circ Y\|_{L^3(\Omega)} \|D_j^\circ Y'\|_{L^3(\Omega)} \|D_k^\circ Y''\|_{L^3(\Omega)} \right. \\ &\quad \left. + \|D_j^\circ Y\|_{L^3(\Omega)} \|D_j^\circ D_k^\circ Y'\|_{L^3(\Omega)} \|D_k^\circ Y''\|_{L^3(\Omega)} \right. \\ &\quad \left. + \|D_j^\circ Y\|_{L^3(\Omega)} \|D_k^\circ Y'\|_{L^3(\Omega)} \|D_j^\circ D_k^\circ Y''\|_{L^3(\Omega)} \right). \quad (3.11) \end{aligned}$$

Although such Glauber estimates for *joint* cumulants were not covered in [13], they are easily deduced with the same toolkit developed in [13], as we briefly explain. Without loss of generality, we may assume that  $Y, Y', Y''$  have vanishing expectation. Using (3.1), we can write

$$\kappa_{1,1,1}[Y, Y', Y''] = \mathbb{E}[YY'Y''] = \text{Cov}(YY', Y'') = \sum_{j=1}^N \mathbb{E}[D_j^\circ(YY')\mathcal{T}^\circ(D_j^\circ Y'')].$$

Now note that we have the approximate chain rule

$$|D_j^\circ(YY') - Y(D_j^\circ Y') - Y'(D_j^\circ Y)| \leq |D_j^\circ Y||D_j^\circ Y'| + \mathbb{E}[|D_j^\circ Y||D_j^\circ Y'| | (\theta_l^\circ)_{l \neq j}].$$

Inserting this into the above and recalling the boundedness of the inverse Glauber Laplacian, cf. (3.1), we are led to

$$\begin{aligned} |\kappa_{1,1,1}[Y, Y', Y'']| &\leq \sum_{j=1}^N (|\mathbb{E}[Y(D_j^\circ Y')\mathcal{T}^\circ(D_j^\circ Y'')]| + |\mathbb{E}[Y'(D_j^\circ Y)\mathcal{T}^\circ(D_j^\circ Y'')]|) \\ &\quad + \sum_{j=1}^N \|D_j^\circ Y\|_{L^3(\Omega)} \|D_j^\circ Y'\|_{L^3(\Omega)} \|D_j^\circ Y''\|_{L^3(\Omega)}. \end{aligned} \quad (3.12)$$

It remains to estimate the first two terms. Using (3.1) again, we can write

$$\mathbb{E}[Y(D_j^\circ Y')\mathcal{T}^\circ(D_j^\circ Y'')] = \text{Cov}(Y, (D_j^\circ Y')\mathcal{T}^\circ(D_j^\circ Y'')) = \sum_{k=1}^N \mathbb{E}[(D_k^\circ Y)\mathcal{T}^\circ D_k^\circ((D_j^\circ Y')\mathcal{T}^\circ(D_j^\circ Y''))],$$

and thus, by the approximate chain rule and the boundedness of the inverse Glauber Laplacian,

$$\begin{aligned} |\mathbb{E}[Y(D_j^\circ Y')\mathcal{T}^\circ(D_j^\circ Y'')]| &\lesssim \sum_{k=1}^N (\|D_k^\circ Y\|_{L^3(\Omega)} \|D_k^\circ D_j^\circ Y'\|_{L^3(\Omega)} \|D_j^\circ Y''\|_{L^3(\Omega)} \\ &\quad + \|D_k^\circ Y\|_{L^3(\Omega)} \|D_j^\circ Y'\|_{L^3(\Omega)} \|D_k^\circ D_j^\circ Y''\|_{L^3(\Omega)}). \end{aligned}$$

Inserting this into (3.12), and noting that for  $j = k$  we have  $D_j^\circ D_k^\circ = D_j^\circ$ , the claim (3.11) follows.

*Step 5. Proof of (3.4).*

Applying (3.11) to trajectories, recalling the chain rules (3.7) and (3.10), and inserting the Glauber estimates (3.5) and (3.9), we get for all  $k < j < i$  and  $t, t', t'' \geq 0$ ,

$$\begin{aligned} |\kappa_{1,1,1}[\varphi(\theta_i(t)), \varphi(\theta_j(t')), \varphi(\theta_k(t''))]| &\leq \|\varphi'\|_{L^\infty(\mathbb{T})}^2 \|\varphi'\|_{W^{1,\infty}(\mathbb{T})} (ij)^{-1} e^{\sqrt{CT \log(i/k)}} e^{CT} \\ &\quad \times \left( 1 + \sum_{l:k < l < j} l^{-1} + k^{-1} \sum_{l:l < k} e^{\sqrt{CT \log(k/l)}} + k^{-1} \sum_{l:l < j} l^{-1} \sum_{m:m < k \wedge l} e^{\sqrt{CT \log(k/m)}} \right) \end{aligned}$$

where we have set for abbreviation  $T := t + t' + t''$ . The sums in the last parenthesis can be bounded by  $\log(j/k) \leq \log(i/k)$  and the claim (3.4) follows.  $\square$

#### 4. MEAN-FIELD LIMIT

This section proves Theorem 1.1. We split the argument into four steps. The qualitative mean-field limit is proven in Step 1 and follows from standard compactness arguments together with a uniqueness result for the mean-field equation (1.10). The error estimates are more delicate and occupy the remaining three steps of the proof. They are based on the covariance estimate of Section 3 and on some stability analysis for the mean-field equation.

Next to the empirical measure (1.8), we shall also need to work with the marginal distribution of the individual particles: for  $t \geq 0$  and  $\sigma \in (0, 1]$ , let  $f_N(t, \sigma, \cdot) \in \mathcal{P}(\mathbb{T})$  denote the probability density of  $\theta_{\lceil N\sigma \rceil}(t)$ ,

$$\int_{\mathbb{T}} \psi(\theta) f_N(t, \sigma, d\theta) = \mathbb{E}[\psi(\theta_{\lceil N\sigma \rceil}(t))], \quad \psi \in C^\infty(\mathbb{T}). \quad (4.1)$$

*Proof of Theorem 1.1.* We proceed in four steps.

*Step 1.* Qualitative mean-field limit.

In terms of the empirical measure (1.8) and the kernel  $K_N$  defined in (2.3), the particle dynamics (1.1) yields in the weak sense,

$$\partial_t \mu_N(t, \sigma, \theta) = -\partial_\theta \left( \mu_N(t, \sigma, \theta) \int_0^\sigma \int_{\mathbb{T}} K_N(\sigma, \sigma') w'_\beta(\theta - \theta') \mu_N(t, d\sigma', d\theta') \right).$$

By weak compactness, up to a subsequence, we have  $\mu_N(t) \xrightarrow{*} f(t)$  for all  $t \geq 0$ . Passing to the limit in the weak formulation of the above equation, using the graphon convergence  $K_N \rightarrow k_\lambda$ , cf. (2.4), we deduce that the limit point  $f$  is necessarily a weak solution of (1.10). In addition, noting that  $\int_{\mathbb{T}} \mu_N(t, \sigma, d\theta) = \frac{1}{N} \sum_{i=1}^N \delta_{i/N}(\sigma) \xrightarrow{*} 1$ , we deduce  $f \in L_{\text{loc}}^\infty(\mathbb{R}_{\geq 0} \times [0, 1]; \mathcal{P}(\mathbb{T}))$ .

It remains to show that a weak solution of (1.10) in  $L_{\text{loc}}^\infty(\mathbb{R}_{\geq 0} \times [0, 1]; \mathcal{P}(\mathbb{T}))$  is necessarily unique. Let  $f, f'$  be two such solutions with identical initial data. For almost all  $\sigma$ , as  $f(\cdot, \sigma, \cdot)$  and  $f'(\cdot, \sigma, \cdot)$  satisfy transport equations with velocity fields  $V_\sigma$  and  $V'_\sigma$ , which are given by

$$V_\sigma(t, \theta) = - \int_0^\sigma k_\lambda(\sigma, \tau) w'_\beta *_\theta f(t, \tau, \theta) d\tau$$

and correspondingly for  $V'_\sigma$ . A standard calculation yields

$$\partial_t^+ W_1(f(t, \sigma), f'(t, \sigma)) \leq \|\nabla V'_\sigma(t)\|_{L^\infty(\mathbb{T})} W_1(f(t, \sigma), f'(t, \sigma)) + \int_{\mathbb{T}} |V_\sigma(t) - V'_\sigma(t)| f(t, \sigma),$$

where  $W_1$  denotes the usual 1-Wasserstein distance on  $\mathcal{P}(\mathbb{T})$ . Inserting the definition of  $V_\sigma, V'_\sigma$ , we deduce

$$\partial_t^+ W_1(f(t, \sigma), f'(t, \sigma)) \leq \|w''_\beta\|_{L^\infty(\mathbb{T})} \left( W_1(f(t, \sigma), f'(t, \sigma)) + \int_0^\sigma k_\lambda(\sigma, \tau) W_1(f(t, \tau), f'(t, \tau)) d\tau \right),$$

and thus,

$$\partial_t^+ \left( \sup_{\sigma \in [0, 1]} W_1(f(t, \sigma), f'(t, \sigma)) \right) \leq 2 \|w''_\beta\|_{L^\infty(\mathbb{T})} \sup_{\sigma \in [0, 1]} W_1(f(t, \sigma), f'(t, \sigma)).$$

By Grönwall's inequality, this entails  $f = f'$ , which concludes the proof of (i).

*Step 2.* Approximate equation for marginal distribution: for  $f_N$  defined in (4.1), we have for all  $t \geq 0$ ,  $\sigma \in (0, 1]$ , and  $n \in \mathbb{Z}$ ,

$$\left| \partial_t \hat{f}_N(t, \sigma, n) - \sum_{\xi \in \mathbb{Z}} n \xi \hat{w}_\beta(\xi) \hat{f}_N(t, \sigma, n - \xi) \int_0^\sigma k_\lambda(\sigma, \tau) \hat{f}_N(t, \tau, \xi) d\tau \right| \leq \frac{C}{N\sigma + 1} e^{Ct} \langle n \rangle^2. \quad (4.2)$$

With the short-hand notation (2.1), the particle dynamics (1.1) yields

$$\frac{d}{dt} \mathbb{E} \left[ e^{-in\theta_j} \right] = -in \sum_{k=1}^{j-1} \omega_{j,k} \mathbb{E} \left[ e^{-in\theta_j} w'_\beta(\theta_j - \theta_k) \right],$$

and thus, using Fourier decomposition for the interaction  $w'_\beta$ ,

$$\frac{d}{dt} \mathbb{E} \left[ e^{-in\theta_j} \right] = \sum_{\xi \in \mathbb{Z}} n \xi \hat{w}_\beta(\xi) \sum_{k=1}^{j-1} \omega_{j,k} \mathbb{E} \left[ e^{-i(n-\xi)\theta_j} e^{-i\xi\theta_k} \right].$$

Using the smallness of correlations obtained in Lemma 3.1, and recalling  $\omega_{j,k} \lesssim j^{-1}$ , we deduce

$$\left| \frac{d}{dt} \mathbb{E} \left[ e^{-in\theta_j} \right] - \sum_{\xi \in \mathbb{Z}} n \xi \hat{w}_\beta(\xi) \sum_{k=1}^{j-1} \omega_{j,k} \mathbb{E} \left[ e^{-i(n-\xi)\theta_j} \right] \mathbb{E} \left[ e^{-i\xi\theta_k} \right] \right|$$

$$\leq Cj^{-1} \sum_{k=1}^{j-1} \omega_{j,k} e^{\sqrt{Ct \log(j/k)}} e^{Ct \langle n \rangle^2} \leq Cj^{-1} e^{Ct \langle n \rangle^2}.$$

For  $\sigma \in (0, 1]$  and  $j = \lceil N\sigma \rceil$ , in terms of Fourier modes  $\hat{f}_N(t, \sigma, n) = \mathbb{E}[e^{-in\theta \lceil N\sigma \rceil(t)}]$  and of the kernel  $K_N$  defined in (2.3), this reads

$$\left| \partial_t \hat{f}_N(t, \sigma, n) - \sum_{\xi \in \mathbb{Z}} n \xi \widehat{w}_\beta(\xi) \hat{f}_N(t, \sigma, n - \xi) \int_0^1 K_N(\sigma, \sigma') \hat{f}_N(t, \sigma', \xi) d\sigma' \right| \leq \frac{C}{N\sigma + 1} e^{Ct \langle n \rangle^2}.$$

Using the graphon convergence (2.4) together with the trivial bound  $|\hat{f}_N| \leq 1$ , this concludes the proof of (4.2).

*Step 3. Stability for limit equation.*

Based on the approximate equation (4.2), we aim to deduce an error estimate between the marginal distribution  $f_N$  and the weak solution  $f$  of the limit equation (1.10). First note that the latter reads as follows in Fourier space,

$$\partial_t \hat{f}(t, \sigma, n) - \sum_{\xi \in \mathbb{Z}} n \xi \widehat{w}_\beta(\xi) \hat{f}(t, \sigma, n - \xi) \int_0^\sigma k_\lambda(\sigma, \tau) \hat{f}(t, \tau, \xi) d\tau = 0.$$

Hence, comparing this with (4.2), we find for the error  $g_N := f_N - f$ ,

$$\begin{aligned} \partial_t \hat{g}_N(t, \sigma, n) - \sum_{\xi \in \mathbb{Z}} n \xi \widehat{w}_\beta(\xi) \hat{f}(t, \sigma, n - \xi) \int_0^\sigma k_\lambda(\sigma, \tau) \hat{g}_N(t, \tau, \xi) d\tau \\ - \sum_{\xi \in \mathbb{Z}} n \xi \widehat{w}_\beta(\xi) \hat{g}_N(t, \sigma, n - \xi) \int_0^\sigma k_\lambda(\sigma, \tau) \hat{f}_N(t, \tau, \xi) d\tau = \hat{r}_N(t, \sigma, n), \end{aligned} \quad (4.3)$$

where the remainder term  $r_N$  satisfies pointwise, for all  $t \geq 0$ ,  $\sigma \in (0, 1]$ , and  $n \in \mathbb{Z}$ ,

$$|\hat{r}_N(t, \sigma, n)| \leq \frac{C}{N\sigma + 1} e^{Ct \langle n \rangle^2}. \quad (4.4)$$

Due to the loss of a derivative in  $\theta$  (that is, of a factor  $n$ ) in (4.3), we cannot prove an  $L^\infty$  stability estimate for Fourier modes  $\hat{g}_N(t, \sigma, n)$ , and we shall rather work in a weighted  $L^2$  setting. Testing (4.3) with  $\langle n \rangle^{-\alpha} \hat{g}_N(t, \sigma, n)$ , for some exponent  $\alpha \geq 0$ , we find

$$\begin{aligned} \partial_t \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \sigma, n)|^2 &\leq 2 \left| \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} \hat{g}_N(t, \sigma, -n) \hat{r}_N(t, \sigma, n) \right| \\ &\quad - 2 \left| \sum_{n, \xi \in \mathbb{Z}} \langle n \rangle^{-\alpha} n \xi \widehat{w}_\beta(\xi) \hat{g}_N(t, \sigma, -n) \hat{f}(t, \sigma, n - \xi) \int_0^\sigma k_\lambda(\sigma, \tau) \hat{g}_N(t, \tau, \xi) d\tau \right| \\ &\quad - 2 \left| \sum_{n, \xi \in \mathbb{Z}} \langle n \rangle^{-\alpha} n \xi \widehat{w}_\beta(\xi) \hat{g}_N(t, \sigma, -n) \hat{g}_N(t, \sigma, n - \xi) \int_0^\sigma k_\lambda(\sigma, \tau) \hat{f}_N(t, \tau, \xi) d\tau \right|. \end{aligned}$$

The sum over  $n$  in the last term can be reorganized as follows, using the symmetry  $n \leftrightarrow \xi - n$ ,

$$\begin{aligned} &\sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} n \hat{g}_N(t, \sigma, -n) \hat{g}_N(t, \sigma, n - \xi) \\ &= \frac{1}{2} \sum_{n \in \mathbb{Z}} (\langle n - \xi \rangle^{-\alpha} \xi + (\langle n \rangle^{-\alpha} - \langle n - \xi \rangle^{-\alpha}) n) \hat{g}_N(t, \sigma, -n) \hat{g}_N(t, \sigma, n - \xi) \\ &\lesssim \langle \xi \rangle^{\alpha+2} \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \sigma, n)| |\hat{g}_N(t, \sigma, n - \xi)|. \end{aligned} \quad (4.5)$$

Further using  $k_\lambda(\sigma, \tau) \lesssim \sigma^{-1}$ , the above then becomes

$$\begin{aligned} \partial_t \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \sigma, n)|^2 \right)^{\frac{1}{2}} &\lesssim \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{r}_N(t, \sigma, n)|^2 \right)^{\frac{1}{2}} \\ &+ \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{2-\alpha} |\hat{f}(t, \sigma, n)|^2 \right)^{\frac{1}{2}} \frac{1}{\sigma} \int_0^\sigma \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \tau, n)|^2 \right)^{\frac{1}{2}} d\tau \\ &+ \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \sigma, n)|^2 \right)^{\frac{1}{2}} \frac{1}{\sigma} \int_0^\sigma \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{f}_N(t, \tau, n)|^2 \right)^{\frac{1}{2}} d\tau. \end{aligned}$$

Using the pointwise bound (4.4) on the remainder  $r_N$ , together with the trivial bounds  $|\hat{f}|, |\hat{f}_N| \leq 1$ , and choosing  $\alpha > 5$ , we are led to

$$\begin{aligned} \partial_t \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \sigma, n)|^2 \right)^{\frac{1}{2}} &\lesssim \frac{1}{N\sigma + 1} e^{Ct} + \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \sigma, n)|^2 \right)^{\frac{1}{2}} + \frac{1}{\sigma} \int_0^\sigma \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \tau, n)|^2 \right)^{\frac{1}{2}} d\tau. \quad (4.6) \end{aligned}$$

Recall the Hardy inequality: for every  $h \in L^p(0, 1)$  and  $1 < p < \infty$ ,

$$\left( \int_0^1 \left| \frac{1}{\sigma} \int_0^\sigma h(\tau) d\tau \right|^p d\sigma \right)^{\frac{1}{p}} \leq \frac{p}{p-1} \left( \int_0^1 |h(\sigma)|^p d\sigma \right)^{\frac{1}{p}}.$$

Using this, we get

$$\partial_t \left\| \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \cdot, n)|^2 \right)^{\frac{1}{2}} \right\|_{L^p(0,1)} \lesssim N^{-\frac{1}{p}} e^{Ct} + \left\| \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \cdot, n)|^2 \right)^{\frac{1}{2}} \right\|_{L^p(0,1)},$$

and thus, by Grönwall's inequality, recalling the initial assumption (1.11) and choosing  $\alpha > 2\gamma + 1$ ,

$$\left\| \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \cdot, n)|^2 \right)^{\frac{1}{2}} \right\|_{L^p(0,1)} \lesssim N^{-\delta \wedge \frac{1}{p}} e^{Ct}. \quad (4.7)$$

*Step 4. Conclusion: proof of error estimates.*

We start by noticing the link between the empirical measure  $\mu_N$  and the marginal distribution  $f_N$ . On the one hand, we note that  $f_N$  approximates the expectation of the empirical measure: by definition of  $\mu_N$  and  $f_N$ , we find for any  $\varphi \in C^\infty((0, 1) \times \mathbb{T})$ ,

$$\mathbb{E} \left[ \int_{(0,1) \times \mathbb{T}} \varphi \mu_N(t) \right] - \int_{(0,1) \times \mathbb{T}} \varphi f_N(t) = \int_0^1 \mathbb{E} \left[ \varphi \left( \frac{[N\sigma]}{N}, \theta_{[N\sigma]}(t) \right) - \varphi(\sigma, \theta_{[N\sigma]}(t)) \right] d\sigma,$$

and thus,

$$\left| \mathbb{E} \left[ \int_{(0,1) \times \mathbb{T}} \varphi \mu_N(t) \right] - \int_{(0,1) \times \mathbb{T}} \varphi f_N(t) \right| \lesssim N^{-1} \|\partial_\sigma \varphi\|_{L^\infty((0,1) \times \mathbb{T})}.$$

On the other hand, the variance of the empirical measure can be expanded as

$$\text{Var} \left[ \int_{(0,1) \times \mathbb{T}} \varphi \mu_N(t) \right]$$

$$= 2N^{-2} \sum_{i>j}^N \text{Cov} \left( \varphi \left( \frac{i}{N}, \theta_i(t) \right), \varphi \left( \frac{j}{N}, \theta_j(t) \right) \right) + N^{-2} \sum_{i=1}^N \text{Var} \left[ \varphi \left( \frac{i}{N}, \theta_i(t) \right) \right],$$

and thus, by the smallness of correlations obtained in Lemma 3.1,

$$\begin{aligned} \text{Var} \left[ \int_{(0,1] \times \mathbb{T}} \varphi \mu_N(t) \right] &\leq CN^{-2} \sum_{i>j}^N i^{-1} e^{\sqrt{Ct \log(i/j)}} e^{Ct} \|\varphi\|_{L^\infty((0,1]; W^{1,\infty}(\mathbb{T}))}^2 + N^{-1} \|\varphi\|_{L^\infty((0,1] \times \mathbb{T})}^2 \\ &\leq CN^{-1} e^{Ct} \|\varphi\|_{L^\infty((0,1]; W^{1,\infty}(\mathbb{T}))}^2. \end{aligned}$$

Combining these two estimates, we deduce

$$\mathbb{E} \left[ \left| \int_{(0,1] \times \mathbb{T}} \varphi(\mu_N(t) - f(t)) \right|^2 \right] \lesssim \left| \int_{(0,1] \times \mathbb{T}} \varphi(f_N(t) - f(t)) \right|^2 + N^{-1} e^{Ct} \|\varphi\|_{W^{1,\infty}((0,1] \times \mathbb{T})}^2.$$

Finally, combining this with the bound (4.7) on  $g_N = f_N - f$ , choosing e.g.  $p = 2$ , the conclusion (1.12) follows.  $\square$

**Remark 4.1** (Convergence of marginal distribution). For future reference, we show that the error bound (4.7) for  $g_N = f_N - f$  can be upgraded to a pointwise estimate. Indeed, using (4.7) to bound the last right-hand side term in (4.6), for  $\alpha > (2\gamma) \vee 4 + 1$  and  $1 < p < \infty$ , we find

$$\begin{aligned} &\partial_t \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \sigma, n)|^2 \right)^{\frac{1}{2}} \\ &\lesssim \frac{1}{N\sigma + 1} e^{Ct} + \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \sigma, n)|^2 \right)^{\frac{1}{2}} + \sigma^{-\frac{1}{p}} \left\| \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \cdot, n)|^2 \right)^{\frac{1}{2}} \right\|_{L^p((0,1))} \\ &\lesssim N^{-\delta \wedge \frac{1}{p}} \sigma^{-\frac{1}{p}} e^{Ct} + \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \sigma, n)|^2 \right)^{\frac{1}{2}}, \end{aligned}$$

and thus, by Grönwall's inequality with the initial assumption (1.11),

$$\left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |\hat{g}_N(t, \sigma, n)|^2 \right)^{\frac{1}{2}} \lesssim N^{-\delta \wedge \frac{1}{p}} \sigma^{-\frac{1}{p}} e^{Ct}.$$

Hence, for any  $\zeta \leq \delta$  with  $\zeta < 1$ , we have for all  $t \geq 0$ ,  $\sigma \in (0, 1]$ , and  $n \in \mathbb{Z}$ ,

$$|\hat{f}_N(t, \sigma, n) - \hat{f}(t, \sigma, n)| \lesssim (N\sigma)^{-\zeta} e^{Ct} \langle n \rangle^C. \quad (4.8)$$

## 5. CHARACTERIZATION OF CORRELATIONS

This section proves Theorem 1.2. The argument proceeds by using the cumulant estimates of Section 3 to derive approximate closed equations for autocorrelations and cross-correlations. The conclusion then follows from a suitable stability analysis for the latter, similarly as in Step 3 of the proof of Theorem 1.1.

*Proof of Theorem 1.2.* We split the argument into three steps.

*Step 1.* Approximate equations for time correlations:

— autocorrelations: for all  $t \geq 0$ ,  $\sigma \in (0, 1]$ , and  $n \in \mathbb{Z}$ ,

$$\left| \partial_t \hat{A}_\varphi^N(t, \sigma, n) - \sum_{\xi \in \mathbb{Z}} n \xi \hat{w}_\beta(\xi) \hat{A}_\varphi^N(t, \sigma, n - \xi) \int_0^\sigma k_\lambda(\sigma, \tau) \hat{f}_N(t, \tau, \xi) d\tau \right|$$

$$\lesssim \frac{1}{N\sigma^2} e^{Ct} \langle n \rangle^3 \|\varphi\|_{W^{2,\infty}(\mathbb{T})}; \quad (5.1)$$

— cross-correlations: for all  $t \geq 0$ ,  $\sigma, \sigma_0 \in (0, 1]$ , and  $n \in \mathbb{Z}$ ,

$$\begin{aligned} & \left| \partial_t (N\hat{C}_\varphi^N)(t, \sigma, n; \sigma_0) - \sum_{\xi \in \mathbb{Z}} n\xi \widehat{w}_\beta(\xi) \hat{f}_N(t, \sigma, n - \xi) \int_{\sigma_0}^\sigma k_\lambda(\sigma, \tau) (N\hat{C}_\varphi^N)(t, \tau, \xi; \sigma_0) d\tau \right. \\ & \quad - \sum_{\xi \in \mathbb{Z}} n\xi \widehat{w}_\beta(\xi) (N\hat{C}_\varphi^N)(t, \sigma, n - \xi; \sigma_0) \int_0^\sigma k_\lambda(\sigma, \tau) \hat{f}_N(t, \tau, \xi) d\tau \\ & \quad \left. - \sum_{\xi \in \mathbb{Z}} n\xi \widehat{w}_\beta(\xi) \hat{f}_N(t, \sigma, n - \xi) k_\lambda(\sigma, \sigma_0) \hat{A}_\varphi^N(t, \sigma_0, \xi) \right| \\ & \lesssim \frac{1}{N\sigma^2} e^{\sqrt{Ct \log(\sigma/\sigma_0)}} e^{Ct} \langle n \rangle^3 \|\varphi\|_{W^{2,\infty}(\mathbb{T})}. \quad (5.2) \end{aligned}$$

Let  $\ell \leq j$  and  $n \in \mathbb{Z}$  be fixed. By the particle dynamics (1.1), using Fourier decomposition for the interaction  $w'_\beta$ , we can compute

$$\begin{aligned} \frac{d}{dt} \text{Cov} \left( e^{in\theta_j}, \varphi(\theta_\ell^0) \right) &= in \sum_{k=1}^{j-1} \omega_{j,k} \text{Cov} \left( e^{in\theta_j} w'_\beta(\theta_j - \theta_k), \varphi(\theta_\ell^0) \right) \\ &= - \sum_{\xi \in \mathbb{Z}} n\xi \widehat{w}_\beta(\xi) \sum_{k=1}^{j-1} \omega_{j,k} \text{Cov} \left( e^{i(n+\xi)\theta_j} e^{-i\xi\theta_k}, \varphi(\theta_\ell^0) \right). \quad (5.3) \end{aligned}$$

By the law of total cumulance, we may expand, for  $k \neq \ell$ ,

$$\begin{aligned} & \text{Cov} \left( e^{i(n+\xi)\theta_j} e^{-i\xi\theta_k}, \varphi(\theta_\ell^0) \right) \\ &= \mathbb{E} \left[ e^{i(n+\xi)\theta_j} e^{-i\xi\theta_k} \varphi(\theta_\ell^0) \right] - \mathbb{E} \left[ e^{i(n+\xi)\theta_j} e^{-i\xi\theta_k} \right] \mathbb{E} \left[ \varphi(\theta_\ell^0) \right] \\ &= \mathbb{E} \left[ e^{i(n+\xi)\theta_j} \right] \text{Cov} \left( e^{-i\xi\theta_k}, \varphi(\theta_\ell^0) \right) + \mathbb{E} \left[ e^{-i\xi\theta_k} \right] \text{Cov} \left( e^{i(n+\xi)\theta_j}, \varphi(\theta_\ell^0) \right) \\ &+ \kappa_{1,1,1} \left( e^{i(n+\xi)\theta_j}, e^{-i\xi\theta_k}, \varphi(\theta_\ell^0) \right), \end{aligned}$$

and thus, appealing to the cumulant estimate of Lemma 3.1,

$$\begin{aligned} & \left| \text{Cov} \left( e^{i(n+\xi)\theta_j} e^{-i\xi\theta_k}, \varphi(\theta_\ell^0) \right) \right. \\ & \quad \left. - \mathbb{E} \left[ e^{i(n+\xi)\theta_j} \right] \text{Cov} \left( e^{-i\xi\theta_k}, \varphi(\theta_\ell^0) \right) - \mathbb{E} \left[ e^{-i\xi\theta_k} \right] \text{Cov} \left( e^{i(n+\xi)\theta_j}, \varphi(\theta_\ell^0) \right) \right| \\ & \lesssim j^{-1} (k \vee \ell)^{-1} e^{\sqrt{Ct \log(j/(k \wedge \ell))}} e^{Ct} \langle \xi \rangle^3 \langle n \rangle^2 \|\varphi\|_{W^{2,\infty}(\mathbb{T})}. \end{aligned}$$

Similarly, for  $k = \ell$ , we find

$$\begin{aligned} & \left| \text{Cov} \left( e^{i(n+\xi)\theta_j} e^{-i\xi\theta_\ell}, \varphi(\theta_\ell^0) \right) - \mathbb{E} \left[ e^{i(n+\xi)\theta_j} \right] \text{Cov} \left( e^{-i\xi\theta_\ell}, \varphi(\theta_\ell^0) \right) \right| \\ & \lesssim j^{-1} e^{\sqrt{Ct \log(j/\ell)}} e^{Ct} \langle \xi \rangle^2 \langle n \rangle \|\varphi\|_{W^{1,\infty}(\mathbb{T})}. \end{aligned}$$

Inserting these estimates into (5.3), and recalling that  $\text{Cov} \left( e^{-i\xi\theta_k}, \varphi(\theta_\ell^0) \right) = 0$  for  $k < \ell$  by the triangular structure of the dynamics, we deduce for all  $t \geq 0$ ,

$$\left| \frac{d}{dt} \text{Cov} \left( e^{in\theta_j}, \varphi(\theta_\ell^0) \right) - \sum_{\xi \in \mathbb{Z}} n\xi \widehat{w}_\beta(\xi) \mathbb{E} \left[ e^{i(n+\xi)\theta_j} \right] \sum_{\ell < k < j} \omega_{j,k} \text{Cov} \left( e^{-i\xi\theta_k}, \varphi(\theta_\ell^0) \right) \right|$$

$$\begin{aligned}
 & - \sum_{\xi \in \mathbb{Z}} n \xi \widehat{w}_\beta(\xi) \operatorname{Cov} \left( e^{i(n+\xi)\theta_j}, \varphi(\theta_\ell^0) \right) \sum_{\substack{1 \leq k < j \\ k \neq \ell}} \omega_{j,k} \mathbb{E} \left[ e^{-i\xi\theta_k} \right] \\
 & - \mathbb{1}_{\ell < j} \sum_{\xi \in \mathbb{Z}} n \xi \widehat{w}_\beta(\xi) \omega_{j,\ell} \mathbb{E} \left[ e^{i(n+\xi)\theta_j} \right] \operatorname{Cov} \left( e^{-i\xi\theta_\ell}, \varphi(\theta_\ell^0) \right) \Big| \lesssim j^{-2} e^{\sqrt{Ct \log(j/\ell)}} e^{Ct} \langle n \rangle^3 \|\varphi\|_{W^{2,\infty}(\mathbb{T})}.
 \end{aligned}$$

Separately considering the cases  $\ell < j$  and  $\ell = j$ , recalling the covariance estimate of Lemma 3.1, using the graphon convergence (2.4), and recalling the definition of correlation functions  $A_\varphi^N, C_\varphi^N$ , cf. (1.13)–(1.14), and of the marginal distribution  $f_N$ , cf. (4.1), this implies the claimed approximate equations (5.1)–(5.2).

*Step 2.* Error estimate for autocorrelations.

Using the mean-field error estimate (4.8) for the marginal distribution  $f_N$ , the approximate equation (5.1) becomes, for  $\zeta \leq \delta$  with  $\zeta < 1$ ,

$$\left| \partial_t \hat{A}_\varphi^N(t, \sigma, n) - \sum_{\xi \in \mathbb{Z}} n \xi \widehat{w}_\beta(\xi) \hat{A}_\varphi^N(t, \sigma, n - \xi) \int_0^\sigma \kappa_\lambda(\sigma, \tau) \hat{f}(t, \tau, \xi) d\tau \right| \lesssim N^{-\zeta} \sigma^{-2} e^{Ct} \langle n \rangle^C \|\varphi\|_{W^{2,\infty}(\mathbb{T})}.$$

Comparing with the solution  $A_\varphi$  of the limit equation (1.17), and using symmetry in  $n$  as in (4.5), we obtain for  $\alpha \gg 1$ ,

$$\partial_t \left( \sum_n \langle n \rangle^{-\alpha} |(\hat{A}_\varphi^N - \hat{A}_\varphi)(t, \sigma, n)|^2 \right)^{\frac{1}{2}} \lesssim \left( \sum_n \langle n \rangle^{-\alpha} |(\hat{A}_\varphi^N - \hat{A}_\varphi)(t, \sigma, n)|^2 \right)^{\frac{1}{2}} + N^{-\zeta} \sigma^{-2} e^{Ct} \|\varphi\|_{W^{2,\infty}(\mathbb{T})},$$

and thus, by Grönwall's inequality,

$$\left( \sum_n \langle n \rangle^{-\alpha} |(\hat{A}_\varphi^N - \hat{A}_\varphi)(t, \sigma, n)|^2 \right)^{\frac{1}{2}} \lesssim e^{Ct} \left( \sum_n \langle n \rangle^{-\alpha} |(\hat{A}_\varphi^N - \hat{A}_\varphi)(0, \sigma, n)|^2 \right)^{\frac{1}{2}} + N^{-\zeta} \sigma^{-2} e^{Ct} \|\varphi\|_{W^{2,\infty}(\mathbb{T})}. \quad (5.4)$$

By definition (1.13), we can write the initial autocorrelation in terms of the marginal distribution,

$$\hat{A}_\varphi^N(0, \sigma, n) = \int_{\mathbb{T}} e^{-in\theta} f_N(0, \sigma, \theta) \left( \varphi(\theta) - \int_{\mathbb{T}} \varphi f_N(0, \sigma, \cdot) \right) d\theta.$$

Comparing this with the initial condition for  $A_\varphi$  in (1.17) and computing the Fourier coefficient, we obtain for  $\sigma \in (0, 1]$ ,

$$\begin{aligned}
 (\hat{A}_\varphi^N - \hat{A}_\varphi)(0, \sigma, n) &= \sum_{\xi \in \mathbb{Z}} \left( \hat{\varphi}(\xi) (\hat{f}_N - \hat{f})(0, \sigma, n - \xi) - \hat{\varphi}(\xi) \hat{f}_N(0, \sigma, -\xi) (\hat{f}_N - \hat{f})(0, \sigma, n) \right. \\
 & \quad \left. - \hat{\varphi}(\xi) \hat{f}(0, \sigma, n) (\hat{f}_N - \hat{f})(0, \sigma, -\xi) \right),
 \end{aligned}$$

and thus, by the initial convergence assumption (1.11),

$$|(\hat{A}_\varphi^N - \hat{A}_\varphi)(0, \sigma, n)| \lesssim N^{-\delta} \langle n \rangle^\gamma \sum_{\xi \in \mathbb{Z}} \langle \xi \rangle^\gamma |\hat{\varphi}(\xi)|.$$

Inserting this into (5.4), the conclusion (1.15) follows.

*Step 3.* Error estimate for cross-correlations.

Using the mean-field error estimate (4.8) for the marginal distribution  $f_N$ , as well as (1.15) for autocorrelations, and recalling the correlation estimates of Lemma 3.1, we can replace  $\hat{f}_N, \hat{A}_\varphi^N$  by  $\hat{f}, \hat{A}_\varphi$  in the approximate equation (5.2): for  $\zeta \leq \delta$  with  $\zeta < 1$ ,

$$\left| \partial_t (N \hat{C}_\varphi^N)(t, \sigma, n; \sigma_0) - \sum_{\xi \in \mathbb{Z}} n \xi \widehat{w}_\beta(\xi) \hat{f}(t, \sigma, n - \xi) \int_{\sigma_0}^\sigma \kappa_\lambda(\sigma, \tau) (N \hat{C}_\varphi^N)(t, \tau, \xi; \sigma_0) d\tau \right|$$

$$\begin{aligned}
& - \sum_{\xi \in \mathbb{Z}} n \xi \widehat{w}_\beta(\xi) (N \widehat{C}_\varphi^N)(t, \sigma, n - \xi; \sigma_0) \int_0^\sigma k_\lambda(\sigma, \tau) \widehat{f}(t, \tau, \xi) d\tau \\
& - \sum_{\xi \in \mathbb{Z}} n \xi \widehat{w}_\beta(\xi) \widehat{f}(t, \sigma, n - \xi) k_\lambda(\sigma, \sigma_0) \widehat{A}_\varphi(t, \sigma_0, \xi) \Big| \lesssim_\varphi \frac{1}{N^\zeta \sigma^2} e^{\sqrt{Ct \log(\sigma/\sigma_0)}} e^{Ct} \langle n \rangle^C.
\end{aligned}$$

Comparing with the solution  $C_\varphi$  of the limit equation (1.18), and using symmetry in  $n$  as in (4.5), we obtain for  $\alpha \gg 1$ ,

$$\begin{aligned}
& \partial_t \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |(N \widehat{C}_\varphi^N - \widehat{C}_\varphi)(t, \sigma, n; \sigma_0)|^2 \right)^{\frac{1}{2}} \\
& \lesssim \int_{\sigma_0}^\sigma k_\lambda(\sigma, \tau) \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |(N \widehat{C}_\varphi^N - \widehat{C}_\varphi)(t, \tau, n; \sigma_0)|^2 \right)^{\frac{1}{2}} d\tau \\
& + \left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |(N \widehat{C}_\varphi^N - \widehat{C}_\varphi)(t, \sigma, n; \sigma_0)|^2 \right)^{\frac{1}{2}} + \frac{C(\varphi)}{N^\zeta \sigma^2} e^{\sqrt{Ct \log(\sigma/\sigma_0)}} e^{Ct},
\end{aligned}$$

and thus, by Grönwall's inequality, with  $\widehat{C}_\varphi^N|_{t=0} = \widehat{C}_\varphi|_{t=0} = 0$ ,

$$\left( \sum_{n \in \mathbb{Z}} \langle n \rangle^{-\alpha} |(N \widehat{C}_\varphi^N - \widehat{C}_\varphi)(t, \sigma, n; \sigma_0)|^2 \right)^{\frac{1}{2}} \lesssim_\varphi \frac{1}{N^\zeta \sigma_0^2} e^{Ct},$$

which concludes the proof of (1.16).  $\square$

## 6. LOST IN THE MIDDLE

This section is devoted to the proof of Theorem 1.3. We work in the iid uniform case  $f \equiv 1$ ,  $A_\varphi \equiv \varphi - \int_{\mathbb{T}} \varphi$ , with fixed  $\lambda, \beta > 0$ . By Theorem 1.2, we recall that the soft accuracy admits the expansion (1.20), so that it remains to study the profile of the leading correction  $\mathcal{S}_t(\sigma_0)$  defined in (1.21). The argument relies on the explicit solvability of the Volterra–Hardy equation (1.19), as stated in Proposition 1.4. For this, we reduce the equation to a Goursat problem and solve it explicitly in terms of modified Bessel functions.

*Proof of Proposition 1.4.* In terms of

$$F(t, \sigma) := e^{\lambda \sigma_0} + \int_{\sigma_0}^\sigma e^{\lambda \sigma'} g_a(t, \sigma'; \sigma_0) d\sigma',$$

the Volterra–Hardy equation (1.19) for  $g_a$  becomes

$$\partial_t g_a(t, \sigma; \sigma_0) - a \frac{\lambda e^{-\lambda \sigma}}{1 - e^{-\lambda \sigma}} F(t, \sigma) = 0.$$

Since  $g_a = e^{-\lambda \sigma} \partial_\sigma F$ , we arrive at

$$\partial_t \partial_\sigma F(t, \sigma) - a \frac{\lambda}{1 - e^{-\lambda \sigma}} F(t, \sigma) = 0, \quad F(t, \sigma_0) = e^{\lambda \sigma_0}, \quad F(0, \sigma) = e^{\lambda \sigma_0}.$$

Now introduce

$$y := Y(\sigma; \sigma_0) = \int_{\sigma_0}^\sigma \frac{\lambda d\rho}{1 - e^{-\lambda \rho}} = \log \frac{e^{\lambda \sigma} - 1}{e^{\lambda \sigma_0} - 1},$$

and set  $U(t, y) := F(t, \sigma(y))$ . Since

$$\partial_\sigma y = \frac{\lambda}{1 - e^{-\lambda \sigma}}, \quad \partial_\sigma F(t, \sigma) = \frac{\lambda}{1 - e^{-\lambda \sigma}} \partial_y U(t, y),$$

we find

$$\partial_t \partial_\sigma F(t, \sigma) = \frac{\lambda}{1 - e^{-\lambda\sigma}} \partial_t \partial_y U(t, y).$$

Hence the equation for  $F$  reduces to the Goursat problem

$$\partial_t \partial_y U(t, y) - aU(t, y) = 0, \quad U(t, 0) = e^{\lambda\sigma_0}, \quad U(0, y) = e^{\lambda\sigma_0}.$$

We solve this by Laplace transform in  $t$ . Writing

$$\tilde{U}(\mu, y) := \int_0^\infty e^{-\mu t} U(t, y) dt,$$

the transformed equation reads

$$\mu \partial_y \tilde{U}(\mu, y) - a \tilde{U}(\mu, y) = 0, \quad \tilde{U}(\mu, 0) = \frac{e^{\lambda\sigma_0}}{\mu}.$$

Therefore

$$\tilde{U}(\mu, y) = \frac{e^{\lambda\sigma_0}}{\mu} e^{ay/\mu} = e^{\lambda\sigma_0} \sum_{k \geq 0} \frac{a^k y^k}{k!} \mu^{-(k+1)}.$$

Inverting termwise yields

$$U(t, y) = e^{\lambda\sigma_0} \sum_{k \geq 0} \frac{(aty)^k}{(k!)^2} = e^{\lambda\sigma_0} I_0(2\sqrt{aty}).$$

Differentiating and using  $I_0'(z) = I_1(z)$  gives

$$\partial_y U(t, y) = e^{\lambda\sigma_0} \sqrt{\frac{at}{y}} I_1(2\sqrt{aty}).$$

Returning to  $F$  and then to  $g_a$ ,

$$g_a(t, \sigma; \sigma_0) = e^{-\lambda\sigma} \partial_\sigma F(t, \sigma) = e^{-\lambda\sigma} \frac{\lambda}{1 - e^{-\lambda\sigma}} \partial_y U(t, Y(\sigma; \sigma_0)),$$

which is exactly (1.23). The formula (1.25) follows from the limit  $\lambda \downarrow 0$ , because  $Y(\sigma; \sigma_0) \rightarrow \log(\sigma/\sigma_0)$  and  $\frac{\lambda e^{-\lambda(\sigma-\sigma_0)}}{1-e^{-\lambda\sigma}} \rightarrow \sigma^{-1}$ .  $\square$

With Proposition 1.4 at hand, the expression for  $\mathcal{S}_t(\sigma_0)$  in (1.21) becomes explicit, and a direct analysis will prove its U-shape. Set

$$c_n := a_n t, \quad Y(\sigma_0) := Y(1; \sigma_0) = \log \frac{e^\lambda - 1}{e^{\lambda\sigma_0} - 1},$$

and define

$$\psi_c(y) := \sqrt{\frac{c}{y}} I_1(2\sqrt{cy}) = \sum_{k \geq 0} \frac{c^{k+1}}{k!(k+1)!} y^k.$$

At the output layer  $\sigma = 1$ , Proposition 1.4 yields

$$g_{a_n}(t, 1; \sigma_0) = \left( \frac{\lambda}{e^\lambda - 1} + \lambda e^{-Y(\sigma_0)} \right) \psi_{c_n}(Y(\sigma_0)).$$

*Proof of Theorem 1.3.* The map  $\sigma_0 \mapsto Y(\sigma_0)$  is a smooth decreasing bijection from  $(0, 1]$  onto  $[0, \infty)$ , so it suffices to consider

$$\mathcal{H}(y) := \sum_{n \geq 1} e^{-\frac{\pi^2}{2M^2} n^2} \left( \frac{\lambda}{e^\lambda - 1} + \lambda e^{-y} \right) \psi_{c_n}(y), \quad y \in [0, \infty).$$

For one mode, write

$$h_c(y) := \left( \frac{\lambda}{e^\lambda - 1} + \lambda e^{-y} \right) \psi_c(y).$$

A direct differentiation yields

$$h_c''(y) = \frac{\lambda}{e^\lambda - 1} \psi_c''(y) + \lambda e^{-y} (\psi_c''(y) - 2\psi_c'(y) + \psi_c(y)).$$

Since

$$\psi_c(y) = \sum_{k \geq 0} \frac{c^{k+1}}{k!(k+1)!} y^k,$$

we have

$$\psi_c''(y) - 2\psi_c'(y) + \psi_c(y) = \sum_{k \geq 0} \frac{c^{k+1}}{k!(k+3)!} (c^2 - 2(k+3)c + (k+2)(k+3)) y^k.$$

If  $0 < c \leq 3 - \sqrt{3}$ , then every coefficient in the last series is nonnegative, hence  $h_c''(y) > 0$  on  $[0, \infty)$ . Under (1.22), each summand is therefore strictly convex, so  $\mathcal{H}$  is strictly convex. At  $y = 0$  we have  $\psi_c(0) = c$  and  $\psi_c'(0) = c^2/2$ , hence

$$h_c'(0) = \lambda c \left( \frac{c}{2(1 - e^{-\lambda})} - 1 \right).$$

The second bound in (1.22) implies  $h_c'(0) < 0$  for every mode, hence  $\mathcal{H}'(0) < 0$ . Finally, because  $I_1(z) \sim e^z/\sqrt{2\pi z}$  as  $z \rightarrow \infty$ , each  $h_c(y)$  tends to  $\infty$  as  $y \rightarrow \infty$ , and so does  $\mathcal{H}(y)$ . A strictly convex function on  $[0, \infty)$  with negative initial slope and diverging right tail has a unique global minimizer. Pulling it back through the bijection  $\sigma_0 \mapsto Y(\sigma_0)$  proves the claim.  $\square$

**Acknowledgments.** The authors thank Thierry Paul for organizing several inspiring ‘‘Round Mean-field’’ workshops over the past years, as well as Emmanuel Trélat and Pierre Le Bris for motivating discussions, which resulted in the genesis of this work.

MD acknowledges financial support from the European Union (ERC, PASTIS, Grant Agreement n°101075879).<sup>2</sup> BG’s research was supported by a Sorbonne Emergences grant and a gift from Google.

**AI tool disclosure.** ChatGPT was used for proofreading, generating the figures in the text and gave the main clues behind the explicit computation of Proposition 1.4. Outside of these AI tool uses, the text of this paper was human generated.

## REFERENCES

- [1] Antonio Álvarez-López, Borjan Geshkovski, and Domènec Ruiz-Balet. Perceptrons and localization of attention’s mean-field landscape. *arXiv preprint arXiv:2601.21366*, 2026.
- [2] Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael M. Bronstein, Petar Veličković, and Razvan Pascanu. Why do LLMs attend to the first token? In *Conference on Language Modeling*, 2025.
- [3] Didier Bresch, Mitia Duerinckx, and Pierre-Emmanuel Jabin. A duality method for mean-field limits with singular interactions. Preprint, arXiv:2402.04695.
- [4] Didier Bresch, Pierre-Emmanuel Jabin, and Juan Soler. A new approach to the mean-field limit of Vlasov-Fokker-Planck equations. *Anal. PDE*, 18(4):1037–1064, 2025.
- [5] Didier Bresch, Pierre-Emmanuel Jabin, and Zhenfu Wang. Mean field limit and quantitative estimates with singular attractive kernels. *Duke Mathematical Journal*, 172(13):2591 – 2641, 2023.
- [6] Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. Emergence of meta-stable clustering in mean-field transformer models. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [7] Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. A multiscale analysis of mean-field transformers in the moderate interaction regime. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.
- [8] Valérie Castin, Pierre Ablin, José Antonio Carrillo, and Gabriel Peyré. A unified perspective on the dynamics of deep transformers. *arXiv preprint arXiv:2501.18322*, 2025.
- [9] Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Quantitative clustering in mean-field transformer models. *arXiv preprint arXiv:2504.14697*, 2025.

<sup>2</sup>Views and opinions expressed are however those of the authors 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.

- [10] Borun D Chowdhury. Lost in the middle at birth: An exact theory of transformer position bias. *arXiv preprint arXiv:2603.10123*, 2026.
- [11] Enrique Queipo de Llano, Alvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael M. Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in LLMs are two sides of the same coin. In *The Fourteenth International Conference on Learning Representations*, 2026.
- [12] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International conference on machine learning*, pages 2793–2803. PMLR, 2021.
- [13] Mitia Duerinckx. On the size of chaos via Glauber calculus in the classical mean-field dynamics. *Communications in Mathematical Physics*, 382(1):613–653, 2021.
- [14] Mitia Duerinckx and Pierre-Emmanuel Jabin. Correlation estimates for Brownian particles with singular interactions. Preprint, arXiv:2510.01507.
- [15] Lev Fedorov, Michaël E Sander, Romuald Elie, Pierre Marion, and Mathieu Laurière. Clustering in deep stochastic transformers. *arXiv preprint arXiv:2601.21942*, 2026.
- [16] Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model. *arXiv preprint arXiv:2410.06833*, 2024.
- [17] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36:57026–57037, 2023.
- [18] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bulletin of the American Mathematical Society*, 62(3):427–479, 2025.
- [19] Borjan Geshkovski, Philippe Rigollet, and Domènec Ruiz-Balet. Measure-to-measure interpolation using transformers. *arXiv preprint arXiv:2411.04551*, 2024.
- [20] François Golse. On the dynamics of large particle systems in the mean field limit. In Adrian Muntean, Jens Rademacher, and Antonios Zagaris, editors, *Macroscopic and Large Scale Phenomena: Coarse Graining, Mean Field Limits and Ergodicity*, pages 1–144. Springer International Publishing, Cham, 2016.
- [21] Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [22] Hanna Herasimchyk, Robin Labryga, Tomislav Prusina, and Sören Laue. A residual-aware theory of position bias in transformers. *arXiv preprint arXiv:2602.16837*, 2026.
- [23] Elias Hess-Childs and Keefer Rowan. Higher-order propagation of chaos in  $L^2$  for interacting diffusions. *Probab. Math. Phys.*, 6(2):581–646, 2025.
- [24] Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. Found in the middle: Calibrating positional attention bias improves long context utilization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14982–14995, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [25] Pierre-Emmanuel Jabin, David Poyato, and Juan Soler. Mean-field limit of non-exchangeable systems. *Communications on Pure and Applied Mathematics*, 78(4):651–741, 2025.
- [26] Pierre-Emmanuel Jabin and Zhenfu Wang. Mean field limit for stochastic particle systems. In Nicola Bellomo, Pierre Degond, and Eitan Tadmor, editors, *Active Particles, Volume 1 : Advances in Theory, Models, and Applications*, pages 379–402. Springer International Publishing, Cham, 2017.
- [27] Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1677, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [28] Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking. *Advances in neural information processing systems*, 37:115652–115681, 2024.
- [29] Hugo Koubbi, Borjan Geshkovski, and Philippe Rigollet. Homogenized transformers. *arXiv preprint arXiv:2604.01978*, 2026.
- [30] Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.
- [31] Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022.
- [32] T. Paul, M. Pulvirenti, and S. Simonella. On the Size of Chaos in the Mean Field Dynamics. *Arch. Ration. Mech. Anal.*, 231(1):285–317, 2019.
- [33] Yury Polyanskiy, Philippe Rigollet, and Andrew Yao. Synchronization of mean-field models on the circle. *arXiv preprint arXiv:2507.22857*, 2025.

- [34] Ofir Press, Noah Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*, 2022.
- [35] Matthew Rosenzweig and Sylvia Serfaty. Global-in-time mean-field convergence for singular Riesz-type diffusive flows. *The Annals of Applied Probability*, 33(2):954 – 998, 2023.
- [36] Sylvia Serfaty. Mean field limit for Coulomb-type flows. *Duke Mathematical Journal*, 169(15):2887–2935, 2020.
- [37] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2024.
- [38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.
- [39] Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. In *Forty-second International Conference on Machine Learning*, 2025.
- [40] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.
- [41] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021.

**Mitia Duerinckx**

Département de Mathématiques  
Université Libre de Bruxelles  
Boulevard du Triomphe  
B-1050 Bruxelles, Belgium  
e-mail: [mitia.duerinckx@ulb.be](mailto:mitia.duerinckx@ulb.be)

**Borjan Geshkovski**

Laboratoire Jacques-Louis Lions  
Inria & Sorbonne Université  
4 Place Jussieu  
75005 Paris, France  
e-mail: [borjan.geshkovski@inria.fr](mailto:borjan.geshkovski@inria.fr)

**Stefano Rossi**

Dipartimento di Matematica Guido Castelnuovo  
Sapienza Università di Roma  
Piazzale Aldo Moro 5  
00185 Rome, Italy  
e-mail: [stefano.rossi2@uniroma1.it](mailto:stefano.rossi2@uniroma1.it)