CRITICAL ATTENTION SCALING IN LONG-CONTEXT TRANSFORMERS

SHI CHEN, ZHENGJIANG LIN, YURY POLYANSKIY, AND PHILIPPE RIGOLLET

Abstract. As large language models scale to longer contexts, attention layers suffer from a fundamental pathology: attention scores collapse toward uniformity as context length n increases, causing tokens to cluster excessively, a phenomenon known as rank-collapse. While *attention scaling* effectively addresses this deficiency by rescaling attention scores with a polylogarithmic factor β_n, theoretical justification for this approach remains lacking.

We analyze a simplified yet tractable model that magnifies the effect of attention scaling. In this model, attention exhibits a phase transition governed by the scaling factor β_n: insufficient scaling collapses all tokens to a single direction, while excessive scaling reduces attention to identity, thereby eliminating meaningful interactions between tokens. Our main result identifies the critical scaling $\beta_n \asymp \log n$ and provides a rigorous justification for attention scaling in YaRN and Qwen, clarifying why logarithmic scaling maintains sparse, content-adaptive attention at large context lengths.

Contents

1. Introduction

The attention mechanism is a cornerstone of modern transformer architectures on which Large Language Models (LLMs) rely. Mathematically, an attention layer is a nonlinear operator ATT that maps a collection of *tokens* $\{x_1, \dots, x_n\}$ from $\mathbb{R}^d$ to $\mathbb{R}^d$. This operator is parametrized by three (possibly sparse) d by d matrices K, Q, and V and maps $\{x_1, \dots, x_n\}$ to $\{x'_1, \dots, x'_n\}$ using the following formula. Define the normalization operator $N(x) = x/\|x\|$ and for any $i = 1, \dots, n$ define $q_i = QN(x_i)$, $k_i = KN(x_i)$. Then $x'_i = \mathsf{ATT}(x_1, \dots, x_n)_i$ is defined as

$$x'_i = V \sum_{j=1}^{n} N(x_j) A_{ij}\,, \qquad A_{ij} = \frac{e^{a_{ij}}}{\sum_{k=1}^{n} e^{a_{ik}}}\,, \tag{1.1}$$

where the terms $a_{ij} = q_i^\top k_j$ are called *attention scores*.

A recent line of theoretical work has demonstrated that attention acts as a *contractive* operator that tends to cluster tokens together; see [DCL21, GLPR24, GLPR25, KPR24, GKPR24, BPA25a, PRY25, CLPR25, CNQG24, GG25]. This clustering effect is also known as "rank-collapse" or "token uniformity" and arises because the distribution of attention scores tends to flatten as the sequence length n grows, causing each token to disperse its attention across too many other tokens rather than focusing selectively.

Various practical solutions have been proposed to curb this clustering behavior. In this work, we focus on simple context-length-aware modifications of the attention mechanism following ideas practically implemented as YaRN [PQFS23], Qwen [BBC+23], SSMax [Nak25], and SWAN-GPT [PLS+25]. These methods employ a straightforward strategy that rescales attention scores a_{ij} by a single polylogarithmic factor β_n; see Table 1. Our goal in this paper is to answer the following fundamental question:

What is the optimal order of magnitude of the β_n scaling?

To address this question, we propose a highly simplified yet completely tractable model for attention. This model exhibits a phase transition governed by the parameter β_n as $n \to \infty$: when β_n is below a critical threshold, attention becomes overly contractive and collapses all tokens to a single direction, while when β_n is too large, attention acts as an identity operator and fails to process information effectively. More precisely, we establish that the critical parameter β_n scales as $\log n$, which corroborates the empirical guidelines underlying YaRN, Qwen, SSMax, and SWAN-GPT.

Method	β_n scaling
YaRN	$(\log n)^2$
Qwen	$\log n$
SSMax	$\log n$
SWAN-GPT	$\log n$

Table 1. Attention scaling factors for various methods. The standard attention score $\exp(k_i^\top q_j)$ is replaced with $\exp(C\beta_n k_i^\top q_j)$, $C > 0$.

Our work is intimately connected to the recent contributions of [CNQG24] and [GG25], who investigate the contractive effects of attention mechanisms with random key and query matrices K and Q to establish proper initialization schemes for these parameters. A crucial insight from [CNQG24] is that analyzing the evolution of symmetric token configurations provides a more mathematically tractable framework compared to the generic input distributions considered in [GLPR25].

This symmetric setting, while simplified, captures essential dynamics of the attention mechanism and enables rigorous theoretical analysis; see also [KGPR25].

The choice $\beta_n = \gamma \log n$ appears natural in retrospect. As noted in [Nak25], with such a scaling the attention weights A_{ij} in (1.1) become

$$A_{ij} = \frac{n^{\gamma a_{ij}}}{\sum_{k=1}^{n} n^{\gamma a_{ik}}} .$$

To illustrate the resulting dynamics, consider a simplified regime where all attention scores a_{ij} are of order one: specifically, let $a_{ii} = 1$ and $a_{ij} = \rho > 0$ for $i \neq j$. In this setting, the off-diagonal weights satisfy

$$A_{ij} = \frac{n^{\gamma\rho}}{n^{\gamma} + (n-1)n^{\gamma\rho}} \sim \begin{cases} 1/n & \text{if } \gamma < \frac{1}{1-\rho} \\ 1/n^{\gamma(1-\rho)} & \text{if } \gamma > \frac{1}{1-\rho} \end{cases}$$

This analysis reveals two distinct regimes. When γ is small (subcritical regime), attention weights are asymptotically uniform, resulting in diffuse attention that, as we demonstrate below, leads to severe token contraction. Conversely, when γ is large (supercritical regime), off-diagonal weights become negligible with respect to the diagonal ones so that the attention mechanism is effectively suppressed.

The critical regime emerges at the phase boundary $\gamma = \frac{1}{1-\rho}$ where attention can concentrate on a sublinear yet nontrivial number of tokens so as to maintain sufficient connections to facilitate information flow from a small set of important tokens. This sparse attention is related to structured attention mechanisms employed in long-context architectures such as Longformer [BPC20] and SWIN [LLC+21] which implement a sliding window over $k \ll n$-nearest neighbors but where proximity is measured in terms of token position rather than embedding. Unlike these structurally constrained approaches that rely on fixed positional neighborhoods, the logarithmic scaling enables the attention pattern to be entirely *content-adaptive*, allowing each token to dynamically select its most relevant context based on semantic similarity rather than positional proximity.

Following similar motivations, [GG25] establish a compelling analogy between attention dynamics and the random energy model from statistical physics [Der81]. Using the replica method—an analytical heuristic from statistical physics—they identify a phase transition occurring at $\beta_n \sim \sqrt{\log n}$, which differs from the scalings presented in Table 1. This result represents a significant discrepancy from our findings and highlights fundamental differences in modeling assumptions. More specifically, their approach assumes that the attention scores a_{ij} are correlated Gaussian random variables. This assumption effectively induces a random geometry on the token space, where similarity between tokens is treated as fundamentally random. In this sense, their model bears closer resemblance to recent Kuramoto models on random graphs studied in [ABK+22, JMS25], where the authors investigate the synchronization of oscillators interacting across the edges of a (sparse) Erdős–Rényi random graph with unit edge weights. However, in the case of [GG25], the random graph is both directed and dense, with the edge pointing from token j to token i having weight given by

$$A_{ij} = \frac{e^{\beta_n a_{ij}}}{\sum_{k=1}^{n} e^{\beta_n a_{ik}}} \tag{1.2}$$

where a_{ij} are Gaussian random variables. While [GG25] assumes a specific correlation structure between the Gaussian random variables, the phase transition they

uncover is expected to be universal within a large class of random matrices including Wigner ones. Crucially though, in such models, the interaction strength A_{ij} is independent of the positional relationship between tokens i and j, making this model qualitatively different from standard attention mechanisms where attention is focused on few (or all) of the preceding tokens.

[BPA25b] adopt a different approach to studying the regime where $n \to \infty$ and $\beta_n \to \infty$, in a more general setting than ours. By considering various levels of generality for the matrices K, Q, V, this work identifies distinct regimes of token dynamics and relates them to the hardmax ($\beta = \infty$) limit. Importantly, the analysis is conducted in the subcritical regime and differs from the present work in focusing on a broader class of models, for which the critical regime has yet to be precisely characterized. We believe that combining the analytical tools developed in both papers could yield a deeper understanding of this critical regime and represents a promising direction for future research.

The remainder of the paper is organized as follows. Section 2 provides a precise mathematical formulation of the phase transition phenomena for the rescaled attention layer. We begin by analyzing token angles and the contractive behavior of tokens under two settings: an idealized but intuitive simplex model (Section 2.1) and a more realistic model with the simplex constraint relaxed (Section 2.2). In both cases, we identify three distinct regimes of the scaling parameter, each leading to qualitatively different contrastive behaviors of the self-attention layer. Section 2.3 turns to the gradient norm of the rescaled attention operator. Because rank collapse is often accompanied by vanishing gradients, we characterize the gradient dynamics across scaling regimes and show when gradients vanish, or stabilize to non-trivial limits. Section 3 presents our numerical experiments, which validate these theoretical predictions.

Throughout this paper, when we denote a quantity as $o_n(1)$, where n is the number of tokens, we mean there are positive constants C_1, C_2 independent of the dimension d, such that $|o_n(1)| \leqslant C_1 n^{-C_2}$. The constants C_1, C_2 depend on the assumptions in theorems.

2. A phase transition for attention

In this section, we establish the main theorem of this paper, namely a phase transition for the contractive properties of the attention layer when $\beta_n = \gamma \log n$ for some $\gamma > 0$.

Following [GLPR25], we study a simplified version of the attention layer with pre-layer norm that is described in the introduction by assuming that $K = Q = V = I_d$. More specifically, the model we study is given as follows.

For any two points $x, y \in \mathbb{R}^d$, let $\langle x, y\rangle = x^\top y$ denote the standard Euclidean inner product in $\mathbb{R}^d$, and $\|x\| = \sqrt{\langle x, x\rangle}$. Finally, recall that $N(x) := x/\|x\|$.

For any collection of tokens $\{x_1, \dots, x_n\}$ in $\mathbb{R}^d$, define $y_i = N(x_i) \in \mathbb{S}^{d-1}$ for $i = 1, \dots, n$ and

$$Z_i := \sum_{k=1}^{n} e^{a_{ik}}\,, \qquad A_{ij} := \frac{e^{a_{ij}}}{Z_i}\,, \qquad a_{ij} := \beta \langle y_i, y_j\rangle\,, \tag{2.1}$$

for $i, j = 1, \dots, n$. We then define

$$\mathsf{ATT}(y_i) \coloneqq \sum_{j=1}^{n} A_{ij} y_j. \tag{2.2}$$

Since the seminal work of [HZRS16], residual connections are added to modern architectures and naturally act as a regularization scheme of the attention map towards the identity; see [CLC+25]. With said residual connections, each token x_i is mapped to x_i' using the following update rule

$$x_i' \coloneqq \mathsf{ATT}(y_i) + \alpha x_i\,, \qquad \alpha \geqslant 0\,. \tag{2.3}$$

Our first goal is to understand where the angle $\measuredangle(x_i', x_j')$ compares to $\measuredangle(x_i, x_j)$. If $\measuredangle(x_i', x_j') < \measuredangle(x_i, x_j)$—or equivalently $\langle y_i', y_j' \rangle > \langle y_i, y_j \rangle$, with $y_i' = N(x_i')$—we say that attention is *contractive*.

The nonlinear update rule (2.3) can produce complex dynamics, in which some pairs of tokens move closer together while others drift apart. This diversity of motion is in fact the most desirable outcome in practice, and it emerges precisely at the phase transition identified in this study. Beyond this critical regime, the tokens exhibit an unexpectedly cohesive behavior. To delineate the boundaries of the critical regime, we assume that the size and relative positions of the initial tokens are governed by constants independent of the number n of tokens. As an analytically tractable extreme of this assumption, we first consider the case in which the tokens form a regular simplex in $\mathbb{R}^d$ as in [CNQG24]. Despite its symmetry, this configuration is sufficient to capture and predict the onset of the phase transition. We subsequently relax this constraint in Section 2.2 to show that the same phase transition occurs in more realistic configurations.

2.1. **The simplex case.** The following assumption was made in [CNQG24] and subsequently in [GG25]. While rather stringent—in particular, it requires $d \geqslant n$—it turns out to provide a tractable yet predictive setup to study the contractive properties of attention.

Assumption 1. *There exists nonnegative constants $q \geqslant 0$ and $\rho \in (0, 1)$ such that $\|x_i\|^2 = q$ and $\langle y_i, y_j \rangle = \rho$, for any $i, j = 1, \dots, n$ and $i \neq j$.*

Under Assumption 1, it is easy to see that there are positive constants ρ' and q' such that $\langle y_i', y_j' \rangle = \rho'$ for all $i \neq j$ and $\|x_i'\|^2 = q'$ for all i. This simplification gives rise to a tractable phase transition.

Theorem 2.1. *Under Assumption 1, there is a $\rho' \in (0, 1)$ such that $\langle y_i', y_j' \rangle = \rho'$ for all $i \neq j$. Moreover, if $\beta = \gamma \log n$ where γ is a positive constant, then for any $i \neq j$, it holds*

$$\lim_{n \to +\infty} \langle y_i', y_j' \rangle = \begin{cases} \frac{\rho(\alpha\sqrt{q}+1)^2}{\alpha^2 q + 2\alpha\sqrt{q}\rho + \rho} & \text{if } \gamma < \frac{1}{1-\rho}, \\ \frac{\rho(\alpha\sqrt{q}+1)^2}{\alpha^2 q + \alpha\sqrt{q}(1+\rho) + \frac{1+3\rho}{4}} & \text{if } \gamma = \frac{1}{1-\rho}, \\ \rho & \text{if } \gamma > \frac{1}{1-\rho}. \end{cases} \tag{2.4}$$

Note that when $\gamma \leqslant \frac{1}{1-\rho}$, the right hand sides of (2.4) are strictly larger than ρ for any $\alpha \geqslant 0$. In other words, in the critical and subcritical regimes attention is contractive even in the presence of a residual connection. Of course, when $\alpha \to \infty$, the effects of attention dissipates and the limit tends to ρ for all phases. This is

expected as the update from y_i to y_i' tends to the identity map, an effect known to "mitigate oversmoothing" in residual neural networks; see [CLC+25].

Note also that for $\alpha = 0$, that is in absence of residual connections, the limit in (2.4) reduces to

$$\lim_{n\to+\infty} \langle y_i', y_j' \rangle = \begin{cases} 1 & \text{if } \gamma < \frac{1}{1-\rho}, \\ \frac{4\rho}{1+3\rho} & \text{if } \gamma = \frac{1}{1-\rho}, \\ \rho & \text{if } \gamma > \frac{1}{1-\rho}. \end{cases} \tag{2.5}$$

In the subcritical case, the tokens contract in one step towards a single cluster when $n \to \infty$ while in the supercritical case, their inner product does not change. In fact, a careful inspection of the proof reveals that in this supercritical regime the attention operator converges to the identity as $n \to \infty$. When $\alpha > 0$, the subcritical case is mitigated by the residual connection which prevents token to collapse to a single point in one step. Nevertheless, this singular behavior reveals a major limitation in the simplex case: since the tokens are equidistant the phase transition reveals an all-or-nothing phenomenon where attention transitions from $A_{ij} \sim 1/n$ so that $\mathsf{ATT}(y_i) = \bar{y} = \frac{1}{n}\sum_{j=1}^n y_j$ for all i to $A_{ij} = \delta_{ij}$ so that $\mathsf{ATT}(y_i) = y_i$ for all i. In the next section, we present a similar result Theorem 2.2, where the simplex assumption is relaxed.

Before we end this section, we present the proof for (2.5) as a special case of Theorem 2.1. The detailed proof for Theorem 2.1 and the later Theorem 2.2 in Section 2.2 is included in Appendix A.

Proof of (2.5). In (2.3), when $\alpha = 0$, we have that $x_i' = \mathsf{ATT}(y_i)$ for each $i = 1, 2, \ldots, n$. In (2.1), under Assumption 1, we notice that the quantity $\sum_{k=1}^n e^{a_{ik}}$ in the denominator of A_{ij} is independent of the choice of i, and equals to $e^\beta + (n-1)e^{\rho\beta}$. Denote this as $Z := e^\beta + (n-1)e^{\rho\beta}$. Then (2.2) and (2.3) become

$$x_i' = \mathsf{ATT}(y_i) = \frac{1}{Z}\left(e^\beta y_i + \sum_{m\neq i} e^{\rho\beta} y_m\right).$$

Under Assumption 1, a direct computation shows that for any $i = 1, 2, \ldots, n$,

$$\langle x_i', x_i' \rangle = \frac{1}{Z^2}\left(e^{2\beta} + 2(n-1)\rho e^{(1+\rho)\beta} + (n-1)(1+(n-2)\rho)e^{2\rho\beta}\right),$$

and for any two different $i, j = 1, 2, \ldots, n$,

$$\langle x_i', x_j' \rangle = \frac{1}{Z^2}\left(\rho e^{2\beta} + 2(1+(n-2)\rho)e^{\beta(1+\rho)} + \left((n-2) + (n^2-3n+3)\rho\right)e^{2\beta\rho}\right).$$

See also Lemma A.3 and Lemma A.4 for more detailed computations for $\langle x_i', x_i' \rangle$ and $\langle x_i', x_j' \rangle$.

For $Z = e^\beta + (n-1)e^{\rho\beta}$, when we let $\beta = \gamma \log n$, we see that $e^\beta = n^\gamma$ and $ne^{\rho\beta} = n^{1+\rho\gamma}$ in Z. The largest term in Z then depends on the relation between γ and $1 + \rho\gamma$: when $\gamma < \frac{1}{1-\rho}$, $n^{1+\rho\gamma}$ is the largest term; when $\gamma > \frac{1}{1-\rho}$, n^γ is the largest term. We then directly get the following three phases for Z from the above arguments:

$$Z = \begin{cases} (1+o_n(1))\cdot ne^{\rho\beta} & \text{if } \gamma < \frac{1}{1-\rho}, \\ (2+o_n(1))\cdot e^\beta & \text{if } \gamma = \frac{1}{1-\rho}, \\ (1+o_n(1))\cdot e^\beta & \text{if } \gamma > \frac{1}{1-\rho}, \end{cases} \tag{2.6}$$

where the terms $o_n(1)$ go to 0 as $n \to +\infty$. Similarly, we can get the the following three phases for $\langle x_i', x_i' \rangle$:

$$\lim_{n\to+\infty} \langle x_i', x_i' \rangle = \begin{cases} \rho & \text{if } \gamma < \frac{1}{1-\rho}, \\ \frac{1+3\rho}{4} & \text{if } \gamma = \frac{1}{1-\rho}, \\ 1 & \text{if } \gamma > \frac{1}{1-\rho}. \end{cases} \tag{2.7}$$

For $\langle x_i', x_j' \rangle$, we always have that $\lim_{n\to+\infty} \langle x_i', x_j' \rangle = \rho$ for γ in these three different regimes. Then (2.5) follows from these two limits since $\langle y_i', y_j' \rangle = \langle x_i'/\|x_i'\|, x_j'/\|x_j'\| \rangle$. □

2.2. **The almost-simplex case.** In this section, we relax Assumption 1 to allow pairwise angles and lengths to vary slightly. This relaxation makes it possible for tokens to lie in a dimension $d \ll n$. Although the resulting bounds are not as sharp as those obtained under Assumption 1, they demonstrate that the critical scaling $\beta_n = \Theta(\log n)$ is intrinsic and not merely an artifact of a particular geometric construction.

Assumption 2. *There exist constants $q_1, q_2 \in (0, \infty), \rho_1, \rho_2 \in (0, 1)$ such that $q_1 \leqslant \|x_i\|^2 \leqslant q_2$ and $\rho_1 \leqslant \langle y_i, y_j \rangle \leqslant \rho_2$, for any $i, j = 1, \ldots, n$ and $i \neq j$. Moreover, $\rho_1 = \langle y_i, y_j \rangle$ for some i, j.*

It is easy to see using standard probabilistic tools that Assumption 2 holds with high probability when the y_i's are independent random vectors uniformly distributed on a half-sphere for example.

Theorem 2.2. *Under Assumption 2, we have the following phase transition when $\beta = \gamma \log n$ for some fixed $\gamma > 0$.*

If $\gamma < \frac{1}{1-\rho_1}$, then there is a constant $\varepsilon > 0$ depending on α, ρ_2, q_1, q_2, such that

$$\varliminf_{n\to+\infty} \min_{i\neq j} \langle y_i', y_j' \rangle \geqslant \rho_1 + \varepsilon > \rho_1, \tag{2.8}$$

which implies that the angle between tokens becomes strictly smaller after an attention layer (2.3).

If $\gamma > \frac{1}{1-\rho_2}$, then for any $i \in [\![1, n]\!]$,

$$\mathsf{ATT}(y_i) = y_i + o_n(1), \text{ and hence } x_i' = y_i + \alpha x_i + o_n(1), \tag{2.9}$$

where the term $o_n(1)$ goes to 0 as $n \to +\infty$ with a speed uniform in i. Hence, when $\gamma > \frac{1}{1-\rho_2}$, for any two different $i, j \in [\![1, n]\!]$,

$$\lim_{n\to+\infty} \langle y_i', y_j' \rangle = \langle y_i, y_j \rangle. \tag{2.10}$$

which implies that the angle between tokens does not change after an attention layer (2.3).

The proof for Theorem 2.2 is included in Appendix A, but the general intuition is similar to the proof for (2.5) in Section 2.1. As we have seen in that proof, the first step to build up phase transition regimes for $\langle y_i', y_j' \rangle$ is to study the phase transition regimes for Z_i in (2.1). Adjusting the logarithmic scaling factor γ causes different phase transition regimes for Z_i first. When γ is small enough, the weights $e^{a_{ik}}$ consisting of Z_i are asymptotically uniform, and each token almost equally interacts with the other tokens. When γ is large enough, each token mostly focuses on itself.

Building on this observation, Theorem 2.1 and Theorem 2.2 together demonstrate that γ controls the effective interaction range of each token. In particular, we have seen in Theorem 2.1 the existence of the critical regime when $\gamma = \frac{1}{1-\rho}$. In this case, although the tokens continue to contract, their rate of shrinkage is evidently slower than in the subcritical regime, as shown in (2.4) and (2.5).

It is hence natural to ask whether further regimes emerge when γ is varied between the supercritical and subcritical threshold. Indeed, in Appendix C, we prove the existence of a nontrivial middle phase when γ is between the two extrema $\frac{1}{1-\rho_1}$ and $\frac{1}{1-\rho_2}$, under a refined assumption on the distribution of tokens, which allows for a sharper characterization of the transition. Under this refined assumption, Theorem C.2 show the existence of γ_1, γ_2 such that (2.3) presents three different phases: $\gamma < \gamma_1$, $\gamma_1 < \gamma < \gamma_2$, and $\gamma > \gamma_2$. In the extreme regimes, when $\gamma < \gamma_1$, each token interacts with almost all the remaining tokens, while when $\gamma > \gamma_2$, each token only focuses on itself, consistent with Theorem 2.2. In the intermediate regime $\gamma_1 < \gamma < \gamma_2$, however, the weights $e^{a_{ik}}$ concentrate on only a small subset of tokens, so that each Z_i and hence the update in (2.3) is dominated by a few highly relevant interactions. This shows that the logarithmic scaling enables each token to dynamically select its most relevant context.

We conclude by noting that those $o_n(1)$ terms in our theorems satisfy the bound $|o_n(1)| \leqslant C_1 n^{-C_2}$ for some positive constants C_1, C_2 that are independent of d (though varying across theorems). As a result, the simplex configuration (Assumption 1) and the almost simplex configuration (Assumption 2) remains valid under repeated application of the ATT operator up to $\mathrm{poly}(n)$ iterations. In particular, the accumulated error remains negligible at this scale, so our theorems and arguments extend to transformers with many layers.

2.3. **Propagation of Gradients under Attention Layer.** In the previous section, we established how attention scaling influences the propagation of token representations, corresponding to running the Transformer in the forward (inference) direction. During training, however, the Transformer is also executed in the *backward* direction to compute gradients via backpropagation [RHW86]. In this section, we show that a similar phase transition arises in the backward pass: in the subcritical regime—where token representations rapidly collapse in the forward pass—the gradients also collapse, whereas in the supercritical regime they retain their scale. The stability of gradients is a crucial computational consideration that strongly affects a model's ability to be trained effectively. For this reason, several theoretical analyses of gradient dynamics in Transformers have been conducted, albeit without attention scaling; see, for example, [CNQG24, DCL21, NAB+22].

Let the input token configuration be denoted by $X(0)$, and let $X(t)$ represent the positions of all tokens at the output of Transformer layer t. To compute gradients, one needs to evaluate the end-to-end input–output Jacobian across L layers of the Transformer. By the chain rule, this Jacobian can be expressed as

$$\frac{\partial X(L)}{\partial X(0)} = \frac{\partial X(L)}{\partial X(L-1)} \frac{\partial X(L-1)}{\partial X(L-2)} \cdots \frac{\partial X(1)}{\partial X(0)}.$$

Thus, the end-to-end Jacobian can be obtained by recursively computing and multiplying the layer-wise Jacobians. This procedure is known as the *adjoint method* in dynamical systems theory [Lio71], and as *backpropagation* in the machine learning community.

Our main result shows that when $\beta_n = \gamma \log n$ with subcritical γ, the typical singular values of $\frac{\partial X(t+1)}{\partial X(t)}$ are close to zero (apart from the contribution of the residual connection). In contrast, for supercritical values of γ, the contribution of the attention component to the Jacobian is non-trivial and behaves as a normalization map. ZL In this case, the attention component does not amplify the total noise cross layers.

We now proceed with formal definitions. For $x \in \mathbb{R}^d$, let $(x)_u$ denote its u-th coordinate for $u = 1, 2, \ldots, d$. The concatenation $X = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{nd}$ represents the configuration of all tokens. The normalization map is defined by

$$\mathcal{N}(X) = \mathcal{N}(x_1, x_2, \ldots, x_n) := \big(N(x_1), N(x_2), \ldots, N(x_n)\big), \tag{2.11}$$

and the attention map by

$$\mathcal{ATT}(Y) = \mathcal{ATT}(y_1, y_2, \ldots, y_n) := \big(\mathsf{ATT}(y_1), \mathsf{ATT}(y_2), \ldots, \mathsf{ATT}(y_n)\big), \tag{2.12}$$

where $\mathsf{ATT}(y_i)$ is defined in (2.2) and $Y = (y_1, \ldots, y_n)$. Under these definitions, the update (2.3) can be written compactly as

$$X' = \mathcal{ATT}(\mathcal{N}(X)) + \alpha X, \tag{2.13}$$

where $X' = (x'_1, x'_2, \ldots, x'_n)$.

We define the $nd \times nd$ Jacobian matrix as

$$\nabla_X X' := \left(\frac{\partial (x'_j)_v}{\partial (x_i)_u}\right)_{(j,v),(i,u)}, \tag{2.14}$$

for $i, j = 1, \ldots, n$ and $u, v = 1, \ldots, d$. The matrix norm of $\nabla_X X'$ is given by

$$\|\nabla_X X'\|^2 := \mathrm{tr}\big[(\nabla_X X')^\top \nabla_X X'\big] = \sum_{i,j=1}^{n} \sum_{u,v=1}^{d} \left(\frac{\partial (x'_j)_v}{\partial (x_i)_u}\right)^2. \tag{2.15}$$

Let $\sigma_1, \sigma_2, \ldots, \sigma_{nd}$ denote the singular values of $\nabla_X X'$. Then the normalized Jacobian norm satisfies

$$\frac{1}{nd}\|\nabla_X X'\|^2 = \frac{1}{nd} \sum_{i=1}^{nd} \sigma_i^2, \tag{2.16}$$

which represents the mean squared singular value of the Jacobian.

Before stating our results on $\frac{1}{nd}\|\nabla_X X'\|^2$, we note that the Jacobian $\nabla_X X'$ can be decomposed into the residual part αI_{nd} and the attention part $\nabla_X\big(\mathcal{ATT}(\mathcal{N}(X))\big)$. As shown in Theorems 2.1 and 2.2, the residual component αI_{nd} does not affect the phase transition behavior. Therefore, to streamline the analysis, we focus exclusively on the attention term $\nabla_X\big(\mathcal{ATT}(\mathcal{N}(X))\big)$ by setting $\alpha = 0$ in (2.13). The following theorems characterize $\frac{1}{nd}\|\nabla_X X'\|^2$ under this setting.

Theorem 2.3. *Adopt Assumption 1 and* (2.13) *with* $\alpha = 0$. *Then, we have the following phase transition phenomenon: let* $\beta = \gamma \log n$ *where* γ *is a positive constant.*

If $\gamma < \frac{1}{1-\rho}$,

$$\frac{1}{nd}\|\nabla_X X'\|^2 = 0 + o_n(1). \tag{2.17}$$

If $\gamma = \frac{1}{1-\rho}$

$$\frac{1}{nd}\|\nabla_X X'\|^2 = \frac{1}{4q}\left(1 - \frac{1}{d}\right) + o_n(1). \tag{2.18}$$

If $\gamma > \frac{1}{1-\rho}$

$$\frac{1}{nd}\|\nabla_X X'\|^2 = \frac{1}{q}\left(1 - \frac{1}{d}\right) + o_n(1). \tag{2.19}$$

In both cases, the terms $o_n(1)$ *go to* 0 *as* $n \to +\infty$*, with speeds depending on* γ, ρ, q*.*

The results of the previous theorem show that under the simplex assumption, the phase transition in the backward dynamics (for gradients) is as sharp as for the forward pass: for small γ, gradients do not flow through the attention block.

We can also extend the analysis for Theorem 2.3 to the relaxed Assumption 2.

Theorem 2.4. *Adopt Assumption 2 and* (2.13) *with* $\alpha = 0$*. Then, we have the following phase transition phenomenon: let* $\beta = \gamma \log n$ *where* γ *is a positive constant.*

If $\gamma < \frac{1}{1-\rho_1}$*,*

$$\frac{1}{nd}\|\nabla_X X'\|^2 \leqslant 4\frac{\gamma^2(\log(n))^2}{q_1 d} + o_n(1), \tag{2.20}$$

If $\gamma > \frac{1}{1-\rho_2}$*,*

$$\frac{1}{nd}\|\nabla_X X'\|^2 \geqslant \frac{1}{q_2}\left(1 - \frac{1}{d}\right) + o_n(1), \tag{2.21}$$

which is away from 0 *even when* d, n *is very large. Indeed, when* $\gamma > \frac{1}{1-\rho_2}$*, for any fixed* $i, j \in [\![1, n]\!]$*,*

$$\left(\frac{\partial(\mathsf{ATT}(N(x_j)))_v}{\partial(x_i)_u}\right)_{d\times d} = \frac{\delta_{ij}}{\|x_i\|}\left(I_d - y_i y_i^T\right) + \mathbf{o}_n(1) + o_n(1)\cdot I_d, \tag{2.22}$$

where the leading order term is exactly $\frac{\partial(N(x_j))_v}{\partial(x_i)_u}$ *as shown in Proposition B.1. Here,* I_d *is the* $d \times d$ *identity matrix, the term* $\mathbf{o}_n(1)$ *(*$o_n(1)$*, respectively) is a* $d \times d$ *matrix (constant, respectively) with matrix norm as defined in* (2.15) *(value, respectively) going to* 0 *as* $n \to +\infty$*, with a speed independent of* i, j *but only depending on* γ, ρ_2, q_1*.*

We present the proofs for Theorem 2.3 and Theorem 2.4 in Appendix B. Note that the $\frac{\log^2 n}{d}$ term in (2.20) is small for typical values of n and d used in Transformers. Theorem 2.3 and Theorem 2.4 also corroborate the fact that tokens collapse fast when γ is in the subcritical regime, while each token only focuses on itself when γ is in the supercritical regime.

3. Numerical Experiments

This section reports numerical experiments designed to corroborate our theoretical predictions. In the following numerical experiments, we test the phase transition in the almost-simplex case as Section 2.2. We generate samples $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$

such that the expectations $\mathbb{E}\|x_i\|^2 = 1$ and $\mathbb{E}\langle x_i, x_j\rangle = \rho \in [0,1]$ for $i \neq j$. More precisely, we generate x_i according to

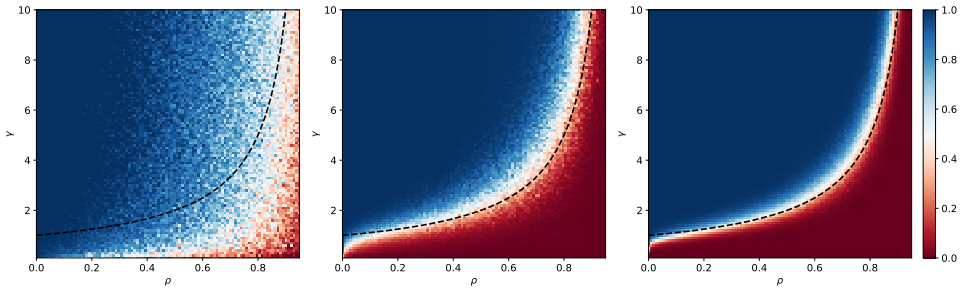
$$x_i = \sqrt{\rho}\, z_0 + \sqrt{1-\rho}\, z_i\,, \tag{3.1}$$

where $z_0, z_1, \ldots, z_n$ are i.i.d. standard Gaussian vectors in $\mathbb{R}^d$. The generated samples satisfy the Assumption 2 with high probability.

In Figure 1, we plot the input-to-output angle ratio λ, defined as

$$\lambda = \frac{2}{n(n-1)} \sum_{1 \leqslant i < j \leqslant n} \frac{1 - \langle y_i', y_j'\rangle}{1 - \langle y_i, y_j\rangle}\,, \tag{3.2}$$

for samples processed through a single self-attention layer with different γ and of different dimensions d. Consistent with our theoretical predictions, the layer acts as a contraction mapping when γ is small, reducing pairwise output angles, whereas for large γ the output angles remain nearly unchanged from the input. Moreover, in the large d regime the angle between input tokens $\langle y_i, y_j\rangle$ $(i \neq j)$ concentrate near ρ, so that the simplex Assumption 1 is effectively satisfied. In this setting, we observe a sharp phase transition in agreement with Theorem 2.1. In the small d regime, however, the input tokens $\langle y_i, y_j\rangle$ randomly distributed in an interval (ρ_1, ρ_2), and an intermediate phase emerges in which the contraction is only partial: some angles shrink significantly while others remain close to their original values, which smooths out the transition.



$d = 2$ $d = 32$ $d = 512$

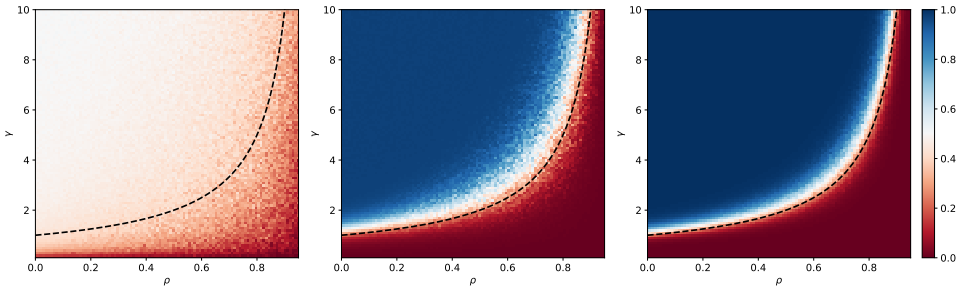
Figure 1. Plots of the input-to-output angle ratio λ, defined in (3.2), as a function of ρ and γ. The tokens are first normalized by a pre-layer normalization and then passed through a single self-attention layer (2.2), with residual connections and MLP layers omitted. The dashed curve corresponds to $\gamma = \frac{1}{1-\rho}$, which approximates the actual phase transition with increasing accuracy as d grows, as implied by Theorem 2.1.

In Figure 2, we plot the normalized matrix norm for the $nd \times nd$ matrix $\nabla_X X'$, defined as

$$\eta = \frac{1}{nd}\|\nabla_X X'\|^2\,, \tag{3.3}$$

for samples passed through a single self-attention layer with varying γ and dimension d. Across all three plots, the normalized gradient norm remains close to 0 when γ is small, while for large γ it approaches $1 - 1/d$, consistent with Theorem 2.4. Similar to the token-angle behavior, a sharp phase transition emerges

near $\gamma = \frac{1}{1-\rho}$ in the large-d regime, in agreement with the predictions under the simplex assumption. In lower dimensions, fluctuations in the pairwise angle prevent perfect concentration, and the transition is smoothed into an intermediate regime where the gradient norm only partially stabilizes.



$d = 2$ $d = 32$ $d = 512$

Figure 2. Plots of the normalized norm η of the gradient, defined by (3.3), as a function of ρ and γ. The tokens are first normalized by a pre-layer normalization and then passed through a single self-attention layer (2.2), with residual connections and MLP layers omitted. The dash curve shows $\frac{1}{1-\rho}$, which approximate the actual phase transition with increasing accuracy as d grows, as implied by Theorem 2.3. The matrix norm η is computed by the Hutchinson trace estimator [Hut89], based on the definition in (2.15).

Appendix A. Proof of Theorem 2.1 and Theorem 2.2

In this section, we adopt Assumption 2 and prove Theorem 2.2 first. Then, we prove Theorem 2.1. To simplify notations, we define $[\![1, n]\!] := \{1, 2, \dots, n\}$ for any $n \in \mathbb{Z}_+$.

We study the asymptotics of the quantity $\langle x_i', x_j' \rangle$ as $n \to +\infty$. We use the notation

$$Z_i := \sum_{k=1}^{n} e^{a_{ik}} = e^{\beta} + \sum_{k \neq i} e^{a_{ik}}. \tag{A.1}$$

Lemma A.1. *Let $\beta = \gamma \log n$ where γ is a positive constant. Under Assumption 2 and (2.3), for any $i \in [\![1, n]\!]$,*

$$Z_i = \begin{cases} (1 + o_n(1)) \cdot \left(\sum_{k \neq i} e^{a_{ik}}\right) & \text{if } \gamma < \frac{1}{1-\rho_1}, \\ (1 + o_n(1)) \cdot e^{\beta} & \text{if } \gamma > \frac{1}{1-\rho_2}, \end{cases} \tag{A.2}$$

where the terms $o_n(1)$ go to 0 as $n \to +\infty$ with speeds independent of i but only depending on γ, ρ_1, ρ_2.

Proof of Lemma A.1. We notice that

$$Z_i = e^{\beta} + \sum_{k \neq i} e^{a_{ik}}. \tag{A.3}$$

We also notice that $e^{\beta t} = n^{\gamma t}$ for any t. It then holds that $e^{\beta} = n^{\gamma}$ and

$$n^{\gamma\rho_1}(n-1) \leqslant \sum_{k\neq i} e^{a_{ik}} \leqslant n^{\gamma\rho_2}(n-1)\,. \tag{A.4}$$

Hence, when $\gamma < \frac{1}{1-\rho_1}$, $n^{\gamma} < n^{1+\gamma\rho_1}$, the leading order term in Z_i is $\sum_{k\neq i} e^{a_{ik}}$. We also see that

$$Z_i = \left(\frac{e^{\beta}}{\left(\sum_{k\neq i} e^{a_{ik}}\right)} + 1\right) \cdot \left(\sum_{k\neq i} e^{a_{ik}}\right), \tag{A.5}$$

with

$$\frac{e^{\beta}}{\left(\sum_{k\neq i} e^{a_{ik}}\right)} \leqslant \frac{n^{\gamma}}{n^{\gamma\rho_1}(n-1)}, \tag{A.6}$$

which goes to 0 as $n \to +\infty$, and is independent of i but only depending on γ, ρ_1. Similarly, when $\gamma > \frac{1}{1-\rho_2}$, $n^{\gamma} > n^{1+\gamma\rho_2}$, the leading order term in Z_i is e^{β}, and similar arguments hold true. □

Lemma A.2. *Let $\beta = \gamma \log n$ where γ is a positive constant. Under Assumption 2 and (2.3), if $\gamma > \frac{1}{1-\rho_2}$, then for any $i \in [\![1, n]\!]$,*

$$\mathsf{ATT}(y_i) = y_i + o_n(1), \text{ and hence } x_i' = y_i + \alpha x_i + o_n(1), \tag{A.7}$$

where the term $o_n(1)$ goes to 0 as $n \to +\infty$ with a speed independent of i but only depending on γ, ρ_2.

Proof of Lemma A.2. According to Lemma A.1, we see that when $\gamma > \frac{1}{1-\rho_2}$, $n^{\gamma} > n^{1+\gamma\rho_2}$, and hence,

$$\mathsf{ATT}(y_i) = Z_i^{-1}\left(e^{\beta} y_i + \sum_{j\neq i} e^{a_{ij}} y_j\right) = (1 + o_n(1))\left(y_i + e^{-\beta}\sum_{j\neq i} e^{a_{ij}} y_j\right). \tag{A.8}$$

Because $\|y_j\| = 1$,

$$\left\| e^{-\beta}\sum_{j\neq i} e^{a_{ij}} y_j \right\| \leqslant e^{-\beta}\sum_{j\neq i} e^{a_{ij}} \leqslant n^{-\gamma} \cdot n^{\gamma\rho_2}(n-1), \tag{A.9}$$

which goes to 0 as $n \to +\infty$, and is independent of i but only depending on γ, ρ_2. This shows that when $\gamma > \frac{1}{1-\rho_2}$,

$$\mathsf{ATT}(y_i) = (1 + o_n(1))(y_i + o_n(1)) = y_i + o_n(1). \tag{A.10}$$

□

Lemma A.3. *Under Assumption 2 and (2.3), for any $i \in [\![1, n]\!]$,*

$$\begin{aligned}\langle x_i', x_i'\rangle = \alpha^2\|x_i\|^2 &+ \frac{2\alpha\|x_i\|}{Z_i}\left(e^{\beta} + \sum_{j\neq i} e^{a_{ij}}\langle y_i, y_j\rangle\right) \\ &+ \frac{1}{Z_i^2}\left(e^{2\beta} + 2e^{\beta}\sum_{j\neq i} e^{a_{ij}}\langle y_i, y_j\rangle + \sum_{j\neq i}\sum_{k\neq i} e^{a_{ij}+a_{ik}}\langle y_k, y_j\rangle\right).\end{aligned} \tag{A.11}$$

Let $\beta = \gamma \log n$ where γ is a positive constant. When $\gamma < \frac{1}{1-\rho_1}$,

(A.12)
$$\langle x_i', x_i' \rangle = \alpha^2 \|x_i\|^2 + 2\alpha \|x_i\| \frac{\sum_{k\neq i} e^{a_{ik}} \langle y_i, y_k \rangle}{\sum_{k\neq i} e^{a_{ik}}} + \frac{\sum_{k\neq i} \sum_{l\neq i} e^{a_{ik}+a_{il}} \langle y_k, y_l \rangle}{\left(\sum_{k\neq i} e^{a_{ik}}\right)^2} + o_n(1).$$

When $\gamma > \frac{1}{1-\rho_2}$,

$$\langle x_i', x_i' \rangle = (\alpha \|x_i\| + 1)^2 + o_n(1). \tag{A.13}$$

In both cases, the terms $o_n(1)$ go to 0 as $n \to +\infty$ with speeds independent of i but only depending on $\gamma, \rho_1, \rho_2, \alpha$.

Proof of Lemma A.3. According to (2.3), we see that

$$\langle x_i', x_i' \rangle = \alpha^2 \|x_i\|^2 + 2\alpha \langle x_i, \mathsf{ATT}(y_i) \rangle + \langle \mathsf{ATT}(y_i), \mathsf{ATT}(y_i) \rangle. \tag{A.14}$$

(A.11) follows from direct computations. Two phase transitions (A.12) and (A.13) follow from similar arguments as in Lemma A.1. □

Lemma A.4. *Under Assumption 2 and (2.3), for any two different $i, j \in [\![1, n]\!]$,*

(A.15)
$$\begin{aligned}
\langle x_i', x_j' \rangle &= \alpha^2 \langle x_i, x_j \rangle \\
&+ \frac{\alpha \|x_j\|}{Z_i} \left(e^{\beta} \langle y_j, y_i \rangle + \sum_{k\neq i} e^{a_{ik}} \langle y_j, y_k \rangle \right) + \frac{\alpha \|x_i\|}{Z_j} \left(e^{\beta} \langle y_i, y_j \rangle + \sum_{l\neq j} e^{a_{jl}} \langle y_i, y_l \rangle \right) \\
&+ \frac{1}{Z_i Z_j} \left(e^{2\beta} \langle y_i, y_j \rangle + e^{\beta} \sum_{k\neq i} e^{a_{ik}} \langle y_j, y_k \rangle + e^{\beta} \sum_{l\neq j} e^{a_{jl}} \langle y_i, y_l \rangle + \sum_{k\neq i} \sum_{l\neq j} e^{a_{ik}+a_{jl}} \langle y_k, y_l \rangle \right).
\end{aligned}$$

Let $\beta = \gamma \log n$ where γ is a positive constant. When $\gamma < \frac{1}{1-\rho_1}$,

$$\begin{aligned}
\langle x_i', x_j' \rangle = \alpha^2 \langle x_i, x_j \rangle &+ \alpha \|x_j\| \frac{\sum_{k\neq i} e^{a_{ik}} \langle y_j, y_k \rangle}{\sum_{k\neq i} e^{a_{ik}}} + \alpha \|x_i\| \frac{\sum_{l\neq j} e^{a_{jl}} \langle y_i, y_l \rangle}{\sum_{l\neq j} e^{a_{jl}}} \\
&+ \frac{\sum_{k\neq i} \sum_{l\neq j} e^{a_{ik}+a_{jl}} \langle y_k, y_l \rangle}{\left(\sum_{k\neq i} e^{a_{ik}}\right) \cdot \left(\sum_{l\neq j} e^{a_{jl}}\right)} + o_n(1).
\end{aligned} \tag{A.16}$$

When $\gamma > \frac{1}{1-\rho_2}$,

$$\langle x_i', x_j' \rangle = (\alpha \|x_i\| + 1)(\alpha \|x_j\| + 1) \langle y_i, y_j \rangle + o_n(1). \tag{A.17}$$

Proof of Lemma A.4. According to (2.3), we see that for two different $i, j \in [\![1, n]\!]$,

$$\langle x_i', x_j' \rangle = \alpha^2 p + \alpha \langle x_i, \mathsf{ATT}(y_j) \rangle + \alpha \langle x_j, \mathsf{ATT}(y_i) \rangle + \langle \mathsf{ATT}(y_i), \mathsf{ATT}(y_j) \rangle. \tag{A.18}$$

(A.15) follows from direct computations. Two phase transitions (A.16) and (A.17) follow from similar arguments as in Lemma A.1. □

Next, we prove Theorem 2.2.

Proof of Theorem 2.2. We first discuss the case when $\gamma < \frac{1}{1-\rho_1}$. According to (A.16) and Assumption 2, we see that

$$\begin{aligned}
\langle x_i', x_j' \rangle &\geqslant \alpha^2 \|x_i\| \|x_j\| \rho_1 + \alpha \|x_j\| \rho_1 + \alpha \|x_i\| \rho_1 + \rho_1 + o_n(1) \\
&= \rho_1 (\alpha \|x_i\| + 1)(\alpha \|x_j\| + 1) + o_n(1).
\end{aligned} \tag{A.19}$$

By (A.12), we see that

$$\begin{aligned}\langle x_i', x_i'\rangle &\leqslant \alpha^2\|x_i\|^2 + 2\alpha\|x_i\|\rho_2 + \rho_2 + o_n(1)\\ &= \alpha^2\|x_i\|^2 + 2\alpha\|x_i\| + 1 - (1-\rho_2)(1+2\alpha\|x_i\|) + o_n(1)\\ &\leqslant (\alpha\|x_i\|+1)^2 - (1-\rho_2)(1+2\alpha q_1) + o_n(1).\end{aligned} \tag{A.20}$$

We have a similar inequality for $\langle x_j', x_j'\rangle$. So, there is a constant $\delta > 0$ depending on ρ_2, α, q_1, q_2 and independent of n, such that

$$\frac{1}{\|x_i'\|} \geqslant \frac{1+\delta}{\alpha\|x_i\|+1} + o_n(1), \text{ and } \frac{1}{\|x_j'\|} \geqslant \frac{1+\delta}{\alpha\|x_j\|+1} + o_n(1). \tag{A.21}$$

Hence,

$$\langle y_i', y_j'\rangle \geqslant \rho_1(1+\delta)^2 + o_n(1) \geqslant \rho_1 + \varepsilon + o_n(1), \tag{A.22}$$

for $\varepsilon = \rho_1(1+2\delta)\delta > 0$ independent of n.

For the case when $\gamma < \frac{1}{1-\rho_2}$, (2.9) and (2.10) follow directly from Lemma A.2, Lemma A.3, and Lemma A.4. □

Proof of Theorem 2.1. We notice that Assumption 1 corresponds to the special case when $q_1 = q_2 = q$ and $\rho_1 = \rho_2 = \rho$ in Assumption 2. Clearly, Z_i is independent of the choice of $i \in [\![1, n]\!]$ by its definition (A.1). According to the explicit forms (A.11) in Lemma A.3 and (A.15) in Lemma A.4, one directly sees that both $\langle x_i, x_i\rangle$ and $\langle x_i, x_j\rangle$ are independent of the choices of $i, j \in [\![1, n]\!]$. We can further compute that for any $i \in [\![1, n]\!]$,

$$\lim_{n\to+\infty}\langle x_i', x_i'\rangle = \begin{cases}\alpha^2 q + 2\alpha\sqrt{q}\rho + \rho & \text{if } \gamma < \frac{1}{1-\rho},\\ \alpha^2 q + \alpha\sqrt{q}(1+\rho) + \frac{1+3\rho}{4} & \text{if } \gamma = \frac{1}{1-\rho},\\ (\alpha\sqrt{q}+1)^2 & \text{if } \gamma > \frac{1}{1-\rho},\end{cases} \tag{A.23}$$

and for any two different $i, j \in [\![1, n]\!]$,

$$\lim_{n\to+\infty}\langle x_i', x_j'\rangle = \rho(\alpha\sqrt{q}+1)^2. \tag{A.24}$$

(2.4) follows from (A.23) and (A.24).

When $\gamma < \frac{1}{1-\rho}$, we see that

$$\lim_{n\to+\infty}\langle y_i', y_j'\rangle = \frac{\rho(\alpha\sqrt{q}+1)^2}{\alpha^2 q + 2\alpha\sqrt{q}\rho + \rho} > \frac{\rho(\alpha\sqrt{q}+1)^2}{\alpha^2 q + 2\alpha\sqrt{q} + 1} = \rho, \tag{A.25}$$

where the strict inequality is because $\rho < 1$. When $\gamma = \frac{1}{1-\rho}$, we can similarly show that $\lim_{n\to+\infty}\langle y_i', y_j'\rangle > \rho$. This completes the proof for Theorem 2.1. □

Appendix B. Proof of Theorem 2.3 and Theorem 2.4

We prove Theorem 2.4 first. We need to explicitly compute terms in $\frac{\partial(\mathsf{ATT}(N(x_j)))_v}{\partial(x_i)_u}$, for which we need the following lemmas.

B.1. **Proof of Theorem 2.4.**

Lemma B.1. *For any $i, k \in [\![1, n]\!]$ and $u, w \in [\![1, d]\!]$,*

$$\frac{\partial(N(x_k))_w}{\partial(x_i)_u} = \delta_{ik}\frac{\delta_{wu}\|x_k\|^2 - (x_k)_w(x_k)_u}{\|x_k\|^3}. \tag{B.1}$$

Proof of Lemma B.1.

$$\frac{\partial(N(x_k))_w}{\partial(x_i)_u} = \frac{\partial((x_k)_w \cdot \|x_k\|^{-1})}{\partial(x_i)_u} = \delta_{ik}\frac{\delta_{wu}\|x_k\| - (x_k)_w \cdot \frac{(x_k)_u}{\|x_k\|}}{\|x_k\|^2}. \tag{B.2}$$

□

Lemma B.2. *For any $k, j \in [\![1, n]\!]$ and $w, v \in [\![1, d]\!]$,*

$$\begin{aligned}
&\frac{\partial(\mathsf{ATT}(y_j))_v}{\partial(y_k)_w} \\
&= \Bigg[\left(\delta_{kj}\beta\left(\sum_{m=1}^n e^{\beta\langle y_j, y_m\rangle}(y_m)_w(y_m)_v\right) + e^{\beta\langle y_j, y_k\rangle}(\beta(y_j)_w(y_k)_v + \delta_{wv})\right) \cdot \left(\sum_{l=1}^n e^{\beta\langle y_j, y_l\rangle}\right) \\
&\quad - \left(\delta_{kj}\beta\left(\sum_{l=1}^n e^{\beta\langle y_j, y_l\rangle}(y_l)_w\right) + \beta e^{\beta\langle y_j, y_k\rangle}(y_j)_w\right) \cdot \left(\sum_{m=1}^n e^{\beta\langle y_j, y_m\rangle}(y_m)_v\right) \Bigg] \\
&\quad \cdot \left(\sum_{l=1}^n e^{\beta\langle y_j, y_l\rangle}\right)^{-2}.
\end{aligned} \tag{B.3}$$

Proof of Lemma B.2. By (2.2),

$$(\mathsf{ATT}(y_j))_v = \frac{\sum_{m=1}^n e^{\beta\langle y_j, y_m\rangle}(y_m)_v}{\sum_{l=1}^n e^{\beta\langle y_j, y_l\rangle}}. \tag{B.4}$$

Lemma B.2 then follows from a direct computation. □

For $x, y \in \mathbb{R}^d$, we use $x \otimes y$ to denote the $d \times d$ matrix with (u, v)-th element $(x \otimes y)_{uv} = (x)_u(y)_v$, i.e., $x \otimes y \coloneqq xy^T$. We then have the following proposition.

Lemma B.3. *Adopt Assumption 2 and (2.3). For any $i, j \in [\![1, n]\!]$, consider the $d \times d$ matrix formed by $\frac{\partial(\mathsf{ATT}(N(x_j)))_v}{\partial(x_i)_u}$, for $u, v \in [\![1, d]\!]$. Denote $y_k = N(x_k)$ for each $k \in [\![1, n]\!]$. Then, this matrix has the following form:*

$$\left(\frac{\partial(\mathsf{ATT}(N(x_j)))_v}{\partial(x_i)_u}\right)_{d\times d} = \|x_i\|^{-\frac{1}{2}}\left[(\mathbf{R}_1 + \mathbf{R}_2)Z_j - (\mathbf{U}_1 + \mathbf{U}_2) \otimes \mathbf{V}_j\right] \cdot Z_j^{-2}, \tag{B.5}$$

where $Z_j = \sum_{l=1}^n e^{\beta\langle y_j, y_l\rangle}$ as in (A.1),

$$\mathbf{R}_1 \coloneqq \delta_{ij}\beta\left(\mathbf{W}_j - y_i \otimes (\mathbf{W}_j y_i)\right), \quad \mathbf{R}_2 \coloneqq e^{\beta\langle y_j, y_i\rangle}\left((-y_i + \beta\mathbf{P}_{y_i}y_j) \otimes y_i + I_d\right), \tag{B.6}$$

and

$$\mathbf{U}_1 \coloneqq \delta_{ij}\beta\left(\mathbf{P}_{y_i}\mathbf{V}_j\right), \quad \mathbf{U}_2 \coloneqq \beta e^{\beta\langle y_j, y_i\rangle}\left(\mathbf{P}_{y_i}y_j\right). \tag{B.7}$$

In (B.6) *and* (B.7),

(B.8)
$$\mathbf{V}_j \coloneqq \sum_{m=1}^{n} e^{\beta\langle y_j, y_m\rangle} y_m, \quad \mathbf{W}_j \coloneqq \sum_{m=1}^{n} e^{\beta\langle y_j, y_m\rangle} y_m \otimes y_m, \quad \mathbf{P}_x y \coloneqq y - \langle y, x\rangle x.$$

Proof of Lemma B.3. By chain rule and Proposition B.1, we have that

(B.9)
$$\begin{aligned}
&\frac{\partial(\mathsf{ATT}(N(x_j)))_v}{\partial(x_i)_u} \\
&= \sum_{k=1}^{n}\sum_{w=1}^{d} \frac{\partial(\mathsf{ATT}(y_j))_v}{\partial(y_k)_w}\bigg|_{Y=\mathcal{N}(X)} \cdot \frac{\partial(N(x_k))_w}{\partial(x_i)_u} \\
&= \|x_i\|^{-\frac{3}{2}} \left(\|x_i\| \cdot \frac{\partial(\mathsf{ATT}(y_j))_v}{\partial(y_i)_u} - \sum_{w=1}^{d} (x_i)_u (x_i)_w \frac{\partial(\mathsf{ATT}(y_j))_v}{\partial(y_i)_w} \right)\bigg|_{Y=\mathcal{N}(X)} \\
&= \|x_i\|^{-\frac{1}{2}} \left(\frac{\partial(\mathsf{ATT}(y_j))_v}{\partial(y_i)_u} - (y_i)_u \sum_{w=1}^{d} (y_i)_w \frac{\partial(\mathsf{ATT}(y_j))_v}{\partial(y_i)_w} \right)\bigg|_{Y=\mathcal{N}(X)}.
\end{aligned}$$

According to Proposition B.2 and the notation $Z_j = \sum_{l=1}^{n} e^{a_{jl}}$, we see that

(B.10)
$$\begin{aligned}
&\sum_{w=1}^{d} (y_i)_w \frac{\partial(\mathsf{ATT}(y_j))_v}{\partial(y_i)_w} \\
&= \Bigg[\left(\delta_{ij}\beta \left(\sum_{m=1}^{n} e^{\beta\langle y_j, y_m\rangle} \langle y_m, y_i\rangle (y_m)_v \right) + e^{\beta\langle y_j, y_i\rangle} (\beta\langle y_j, y_i\rangle + 1)(y_i)_v \right) \cdot Z_j \\
&\quad - \left(\delta_{ij}\beta \left(\sum_{l=1}^{n} e^{\beta\langle y_j, y_l\rangle} \langle y_l, y_i\rangle \right) + \beta e^{\beta\langle y_j, y_i\rangle} \langle y_j, y_i\rangle \right) \cdot \left(\sum_{m=1}^{n} e^{\beta\langle y_j, y_m\rangle} (y_m)_v \right) \Bigg] \cdot Z_j^{-2}.
\end{aligned}$$

Hence,

(B.11)
$$\begin{aligned}
&\|x_i\|^{\frac{1}{2}} \cdot \frac{\partial(\mathsf{ATT}(N(x_j)))_v}{\partial(x_i)_u} \\
&= \left(\frac{\partial(\mathsf{ATT}(y_j))_v}{\partial(y_i)_u} - (y_i)_u \sum_{w=1}^{d} (y_i)_w \frac{\partial(\mathsf{ATT}(y_j))_v}{\partial(y_i)_w} \right)\bigg|_{Y=\mathcal{N}(X)} \\
&= \Bigg[\Bigg[\delta_{ij}\beta \left(\sum_{m=1}^{n} e^{\beta\langle y_j, y_m\rangle} \left((y_m)_u (y_m)_v - \langle y_m, y_i\rangle (y_m)_v (y_i)_u \right) \right) \\
&\qquad + e^{\beta\langle y_j, y_i\rangle} \left(\beta (y_j)_u (y_i)_v + \delta_{uv} - (\beta\langle y_j, y_i\rangle + 1)\, (y_i)_v (y_i)_u \right) \Bigg] \cdot Z_j \\
&\qquad - \Bigg[\delta_{ij}\beta \left(\sum_{l=1}^{n} e^{\beta\langle y_j, y_l\rangle} \left((y_l)_u - \langle y_l, y_i\rangle (y_i)_u \right) \right) \\
&\qquad + \beta e^{\beta\langle y_j, y_i\rangle} \left((y_j)_u - \langle y_j, y_i\rangle (y_i)_u \right) \Bigg] \cdot \left(\sum_{m=1}^{n} e^{\beta\langle y_j, y_m\rangle} (y_m)_v \right) \Bigg] \cdot Z_j^{-2}.
\end{aligned}$$

We then adopt the notation (B.8), i.e.,

(B.12)
$$\mathbf{V}_j = \sum_{m=1}^{n} e^{\beta\langle y_j, y_m\rangle} y_m, \quad \mathbf{W}_j = \sum_{m=1}^{n} e^{\beta\langle y_j, y_m\rangle} y_m \otimes y_m, \quad \mathbf{P}_x y \coloneqq y - \langle y, x\rangle x.$$

So, the matrix form of (B.11) becomes

(B.13)
$$\begin{aligned}
&\left[\left[\delta_{ij}\beta\left(\mathbf{W}_j - y_i \otimes (\mathbf{W}_j y_i)\right) + e^{\beta\langle y_j, y_i\rangle}(\beta y_j \otimes y_i + I_d - (\beta\langle y_j, y_i\rangle + 1)\, y_i \otimes y_i)\right] \cdot Z_j \right.\\
&\left. - \left[\delta_{ij}\beta\left(\mathbf{V}_j - \langle \mathbf{V}_j, y_i\rangle y_i\right) + \beta e^{\beta\langle y_j, y_i\rangle}\left(y_j - \langle y_j, y_i\rangle y_i\right)\right] \otimes \mathbf{V}_j\right] \cdot Z_j^{-2}\\
&= \left[\left[\delta_{ij}\beta\left(\mathbf{W}_j - y_i \otimes (\mathbf{W}_j y_i)\right) + e^{\beta\langle y_j, y_i\rangle}((-y_i + \beta \mathbf{P}_{y_i} y_j) \otimes y_i + I_d)\right] \cdot Z_j \right.\\
&\left. - \left[\delta_{ij}\beta\left(\mathbf{P}_{y_i}\mathbf{V}_j\right) + \beta e^{\beta\langle y_j, y_i\rangle}\left(\mathbf{P}_{y_i} y_j\right)\right] \otimes \mathbf{V}_j\right] \cdot Z_j^{-2}.
\end{aligned}$$

We further use the notations in (B.6) and (B.7), i.e.,

(B.14)
$$\mathbf{R}_1 = \delta_{ij}\beta e^{\beta\rho}\left(\mathbf{W}_j - y_i \otimes (\mathbf{W}_j y_i)\right), \quad \mathbf{R}_2 = e^{\beta\langle y_j, y_i\rangle}\left((-y_i + \beta \mathbf{P}_{y_i} y_j) \otimes y_i + I_d\right),$$

and

(B.15)
$$\mathbf{U}_1 = \delta_{ij}\beta\left(\mathbf{P}_{y_i}\mathbf{V}_j\right), \quad \mathbf{U}_2 = \beta e^{\beta\langle y_j, y_i\rangle}\left(\mathbf{P}_{y_i} y_j\right).$$

Finally, the matrix form of (B.11) becomes

(B.16)
$$\left[(\mathbf{R}_1 + \mathbf{R}_2)Z - (\mathbf{U}_1 + \mathbf{U}_2) \otimes \mathbf{V}_j\right] \cdot Z_j^{-2}.$$

□

Lemma B.4. *Let $\beta = \gamma \log n$ where γ is a positive constant. Under Assumption 2 and (2.3), if $\gamma > \frac{1}{1-\rho_2}$, then for any fixed $i, j \in [\![1, n]\!]$, the $d \times d$ matrix satisfies*

(B.17)
$$\left(\frac{\partial(\mathsf{ATT}(N(x_j)))_v}{\partial(x_i)_u}\right)_{d\times d} = \frac{\delta_{ij}}{\|x_i\|}\left(I_d - y_i \otimes y_i\right) + \mathbf{o}_n(1) + o_n(1) \cdot I_d,$$

where the leading order term is exactly $\frac{\partial(N(x_j))_v}{\partial(x_i)_u}$. The term $\mathbf{o}_n(1)$ ($o_n(1)$, respectively) is a $d \times d$ matrix (constant, respectively) with matrix norm as defined in (2.15) (value, respectively) going to 0 as $n \to +\infty$, with a speed independent of i, j but only depending on γ, ρ_2, q_1.

Proof of Lemma B.4. We frequently use this formula: for two vectors V_1, V_2, the matrix norm of $V_1 \otimes V_2$ as defined in (2.15) is $\|V_1\|\|V_2\|$. When $\gamma > \frac{1}{1-\rho_2}$, $n^\gamma > n^{1+\gamma\rho_2}$, and we know from Lemma A.1 that $Z_j = (1 + o_n(1)) \cdot e^\beta$ for any $j \in [\![1, n]\!]$. Adopt the notations in Proposition B.3, we then show the following facts when $\gamma > \frac{1}{1-\rho_2}$:

(B.18)
$$\mathbf{R}_1 Z_j^{-1} = \mathbf{o}_n(1), \quad \mathbf{R}_2 Z_j^{-1} = \delta_{ij}\left(-y_i \otimes y_i + I_d\right) + \mathbf{o}_n(1) + o_n(1) \cdot I_d,$$

and

(B.19)
$$\left[(\mathbf{U}_1 + \mathbf{U}_2) \otimes \mathbf{V}_j\right] \cdot Z_j^{-2} = \mathbf{o}_n(1).$$

First, for $\mathbf{R}_1 Z_j^{-1}$, when $i \neq j$, we have that $\mathbf{R}_1 = 0$ by its definition. When $i = j$, $\mathbf{R}_1 = \beta \sum_{m=1}^n e^{\beta\langle y_i, y_m\rangle} (y_m \otimes y_m - \langle y_m, y_i\rangle y_i \otimes y_m)$ and we notice that the term when $m = i$ is 0. So, because $\|y_m \otimes y_m - \langle y_m, y_i\rangle y_i \otimes y_m\| \leqslant \|y_m\|^2 + \|y_m\|^2\|y_i\|^2 = 2$, $e^\beta = n^\gamma$,

$$\|\mathbf{R}_1\| Z_j^{-1} \leqslant \beta(n-1)e^{\beta\rho_2} \cdot 2Z_j^{-1} \leqslant 2\gamma \log(n) \cdot n^{\gamma\rho_2 + 1 - \gamma}(1 + o_n(1)), \tag{B.20}$$

which goes to 0 with a speed independent of i, j, because $\gamma\rho_2 + 1 - \gamma < 0$.

For $\mathbf{R}_2 Z_j^{-1}$, we notice that when $i \neq j$, $e^{\beta\langle y_i, y_j\rangle} Z_j^{-1} \leqslant e^{\beta(\rho_2 - 1)}(1 + o_n(1)) = n^{\gamma(\rho_2 - 1)}(1 + o_n(1))$, which goes to 0 with a speed independent of i, j. So, $\mathbf{R}_2 Z_j^{-1} = \mathbf{o}_n(1) + o_n(1) \cdot I_d$ when $i \neq j$. When $i = j$, $\mathbf{R}_2 = e^\beta (-y_i \otimes y_i + I_d)$, and so $\mathbf{R}_2 Z_j^{-1} = (-y_i \otimes y_i + I_d) + \mathbf{o}_n(1) + o_n(1) \cdot I_d$.

For $[(\mathbf{U}_1 + \mathbf{U}_2) \otimes \mathbf{V}_j] \cdot Z_j^{-2}$, we see that when $i \neq j$, $\mathbf{U}_1 = 0$, and so

$$\begin{aligned} &\|(\mathbf{U}_1 + \mathbf{U}_2) \otimes \mathbf{V}_j\| \cdot Z_j^{-2} \leqslant Z_j^{-2}\beta \sum_{m=1}^n e^{\beta\langle y_j, y_m + y_i\rangle} \|y_m\| \|\mathbf{P}_{y_i} y_j\| \\ &\leqslant Z_j^{-2}\beta e^{\beta(1+\rho_2)} n = \gamma \log(n) n^{\gamma(\rho_2 - 1) + 1}(1 + o_n(1)), \end{aligned} \tag{B.21}$$

which goes to 0 with a speed independent of i, j because $\gamma > \frac{1}{1-\rho_2}$. When $i = j$, $\mathbf{U}_2 = 0$, and so

$$\begin{aligned} &\|(\mathbf{U}_1 + \mathbf{U}_2) \otimes \mathbf{V}_j\| \cdot Z_j^{-2} \leqslant Z_j^{-2}\beta \|\mathbf{P}_{y_i} \mathbf{V}_i\| \|\mathbf{V}_i\| \\ &\leqslant Z_j^{-2}\beta \left(\sum_{m \neq i} e^{\beta\langle y_i, y_m\rangle} \|\mathbf{P}_{y_i} y_m\| \right) \cdot \left(e^\beta + \sum_{m \neq i} e^{\beta\langle y_i, y_m\rangle} \|\mathbf{P}_{y_i} y_m\| \right) \\ &\leqslant Z_j^{-2}\beta \left(e^{\beta\rho_2} n \right) \cdot \left(e^\beta + e^{\beta\rho_2} n \right) \\ &= \gamma \log(n) n^{\gamma(\rho_2 - 1) + 1} (1 + n^{\gamma(\rho_2 - 1) + 1})(1 + o_n(1)), \end{aligned} \tag{B.22}$$

which goes to 0 with a speed independent of i, j because $\gamma > \frac{1}{1-\rho_2}$. Hence, $[(\mathbf{U}_1 + \mathbf{U}_2) \otimes \mathbf{V}_j] \cdot Z_j^{-2} = \mathbf{o}_n(1)$. □

Lemma B.5. *Let $\beta = \gamma \log n$ where γ is a positive constant. Under Assumption 2 and (2.3), if $\gamma < \frac{1}{1-\rho_1}$, then for fixed $i, j \in [\![1, n]\!]$, the $d \times d$ matrix satisfies*

$$\left\| \left(\frac{\partial(\mathsf{ATT}(N(x_j)))_v}{\partial(x_i)_u} \right)_{d \times d} \right\| \leqslant \|x_i\|^{-\frac{1}{2}} \cdot \left(2\beta\delta_{ij} + (2\beta + \sqrt{d}) e^{a_{ij}} Z_j^{-1} \right). \tag{B.23}$$

Proof of Lemma B.5. According to Lemma A.1, when $\gamma < \frac{1}{1-\rho_1}$, $Z_j = (1 + o_n(1)) \cdot \left(\sum_{k \neq j} e^{a_{jk}} \right)$ for any $j \in [\![1, n]\!]$, and $Z_j \geqslant n^{\gamma\rho_1 + 1}(1 + o_n(1)) > n^\gamma(1 + o_n(1))$, because $\gamma\rho_1 + 1 > \gamma$. Adopt the notations in Proposition B.3, we then show the following facts when $\gamma < \frac{1}{1-\rho_1}$:

$$\|\mathbf{R}_1\| Z_j^{-1} \leqslant \delta_{ij}\beta, \quad \|\mathbf{R}_2\| Z_j^{-1} \leqslant Z_j^{-1} e^{a_{ij}} \left(\beta + \sqrt{d - 1} \right), \tag{B.24}$$

and

$$\|(\mathbf{U}_1 + \mathbf{U}_2) \otimes \mathbf{V}_j\| \cdot Z_j^{-2} \leqslant \beta \left(\delta_{ij} + e^{a_{ij}} Z_j^{-1} \right). \tag{B.25}$$

First, for $\mathbf{R}_1 Z_j^{-1}$, when $i \neq j$, we have that $\mathbf{R}_1 = 0$ by its definition. When $i = j$, $\mathbf{R}_1 = \beta \sum_{m=1}^n e^{\beta\langle y_i, y_m\rangle} (y_m \otimes y_m - \langle y_m, y_i\rangle y_i \otimes y_m)$. So, because we have

that $\|y_m \otimes y_m - \langle y_m, y_i\rangle y_i \otimes y_m\| = \|\mathbf{P}_{y_i} y_n \otimes y_m\| = \|\mathbf{P}_{y_i} y_n\| \|y_m\| \leqslant 1$,

$$\|\mathbf{R}_1\| Z_j^{-1} \leqslant \beta Z_j \cdot \cdot Z_j^{-1} = \beta. \tag{B.26}$$

For $\mathbf{R}_2 Z_j^{-1}$, because $\| - y_i \otimes y_i + I_d\| = \sqrt{d-1}$, we have that

$$\|\mathbf{R}_2\| Z_j^{-1} \leqslant Z_j^{-1} e^{a_{ij}} \left(\beta + \sqrt{d-1}\right). \tag{B.27}$$

For $[(\mathbf{U}_1 + \mathbf{U}_2) \otimes \mathbf{V}_j] \cdot Z_j^{-2}$, we see that $\|\mathbf{V}_j\| \leqslant \sum_{m=1}^n e^{\beta\langle y_j, y_m\rangle} = Z_j$. Also, $\|\mathbf{U}_1\| Z_j^{-1} \leqslant \delta_{ij}\beta \|\mathbf{V}_j\| Z_j^{-1} \leqslant \delta_{ij}\beta$, $\|\mathbf{U}_2\| Z_j^{-1} \leqslant \beta e^{a_{ij}} Z_j^{-1}$. Hence, we have that

$$\|(\mathbf{U}_1 + \mathbf{U}_2) \otimes \mathbf{V}_j\| \cdot Z_j^{-2} \leqslant \beta \left(\delta_{ij} + e^{a_{ij}} Z_j^{-1}\right). \tag{B.28}$$

□

Proof of Theorem 2.4. Theorem 2.4 follows directly from Lemma B.4 and Lemma B.5. □

B.2. **Proof of Theorem 2.3.** The proof for Theorem 2.3 requires more delicate arguments. The part when $\gamma > \frac{1}{1-\rho}$ in Theorem 2.3 directly follows from Lemma B.4, so we only focus on the part when $\gamma \leqslant \frac{1}{1-\rho}$. We remark that when $\gamma < \frac{1}{1-\rho}$, our result is that $\frac{1}{nd}\|\nabla_X X'\|^2 = 0 + o_n(1)$, which is a better estimate than (2.20) in Theorem 2.4.

We first have the following lemma which replaces Lemma B.3 when we adopt Assumption 1.

Lemma B.6. *Adopt Assumption 1 and* (2.3)*. For any $i, j \in [\![1, n]\!]$, consider the $d \times d$ matrix formed by $\frac{\partial(\mathsf{ATT}(N(x_j)))_v}{\partial(x_i)_u}$, for $u, v \in [\![1, d]\!]$. Denote $y_k = N(x_k)$ for each $k \in [\![1, n]\!]$. Then, this matrix has the following form:*

$$\left(\frac{\partial(\mathsf{ATT}(N(x_j)))_v}{\partial(x_i)_u}\right)_{d\times d} = q^{-\frac{1}{2}} \left[(\mathbf{R}_1 + \mathbf{R}_2)Z - (\mathbf{U}_1 + \mathbf{U}_2) \otimes (\mathbf{U}_3 + \mathbf{U}_4)\right] \cdot Z^{-2}, \tag{B.29}$$

where $Z = e^\beta + (n-1)e^{\beta\rho}$,

$$\mathbf{R}_1 \coloneqq \delta_{ij}\beta e^{\beta\rho}\left(\mathbf{W} - y_i \otimes (\mathbf{W} y_i)\right), \quad \mathbf{R}_2 \coloneqq e^{\beta\langle y_j, y_i\rangle}\left((-y_i + \beta \mathbf{P}_{y_i} y_j) \otimes y_i + I_d\right), \tag{B.30}$$

and

$$\begin{aligned} &\mathbf{U}_1 \coloneqq \delta_{ij}\beta e^{\beta\rho}\left(\mathbf{P}_{y_i}\mathbf{V}\right), \quad \mathbf{U}_2 \coloneqq \beta e^{\beta\langle y_j, y_i\rangle}\left(\mathbf{P}_{y_i} y_j\right), \\ &\mathbf{U}_3 \coloneqq (e^\beta - e^{\beta\rho}) y_j, \quad \mathbf{U}_4 \coloneqq e^{\beta\rho}\mathbf{V}. \end{aligned} \tag{B.31}$$

In (B.30) *and* (B.31)*,*

$$\mathbf{V} \coloneqq \sum_{m=1}^n y_m, \quad \mathbf{W} \coloneqq \sum_{m=1}^n y_m \otimes y_m, \quad \mathbf{P}_x y \coloneqq y - \langle y, x\rangle x. \tag{B.32}$$

Proof of Lemma B.6. We first apply Lemma B.3 to get (B.5). After replacing $\langle y_j, y_m\rangle = \rho$ for $m \neq j$, we can obtain (B.29). The only remark is that the term $\delta_{ij}(\mathbf{W}_j - y_i \otimes (\mathbf{W}_j y_i))$ in $\mathbf{R}_1$ of (B.6) is nonzero when $i = j$. Then, when $i = j$, $\mathbf{W}_i - y_i \otimes (\mathbf{W}_i y_i) = \sum_{m=1}^n e^{\beta\langle y_i, y_m\rangle}(y_m \otimes y_m - y_i \otimes y_m \langle y_m, y_i\rangle)$. If $m = i$, the summand $(y_m \otimes y_m - y_i \otimes y_m \langle y_m, y_i\rangle)$ becomes 0. Hence, $\mathbf{W}_i - y_i \otimes (\mathbf{W}_i y_i) = \sum_{m\neq i} e^{\beta\langle y_i, y_m\rangle}(y_m \otimes y_m - y_i \otimes y_m \langle y_m, y_i\rangle) = e^{\beta\rho}\sum_{m\neq i}(y_m \otimes y_m - y_i \otimes y_m \langle y_m, y_i\rangle) = e^{\beta\rho}\left(\mathbf{W} - y_i \otimes (\mathbf{W} y_i)\right)$. □

Next, to compute the matrix norm of (B.29), we see that for any matrix K, its matrix norm square equals to $\mathrm{Tr}(K^TK)$. Hence, the matrix norm square of (B.29) equals to

(B.33)
$$q^{-1}Z^{-4}\cdot\Big(\mathrm{Tr}\left[Z^2(\mathbf{R}_1+\mathbf{R}_2)^T(\mathbf{R}_1+\mathbf{R}_2)\right]-2Z(\mathbf{U}_1+\mathbf{U}_2)^T(\mathbf{R}_1+\mathbf{R}_2)(\mathbf{U}_3+\mathbf{U}_4)$$
$$+\|\mathbf{U}_1+\mathbf{U}_2\|^2\|\mathbf{U}_3+\mathbf{U}_4\|^2\Big).$$

We then compute these terms separately, and sum them in i,j. We first have the following basic equalities for the notations $\mathbf{V},\mathbf{W}$ in (B.32).

Lemma B.7. *For the notations in* (B.32)*, i.e.,*

$$\mathbf{V}\coloneqq\sum_{m=1}^n y_m,\quad \mathbf{W}\coloneqq\sum_{m=1}^n y_m\otimes y_m,\quad \mathbf{P}_x y\coloneqq y-\langle y,x\rangle x, \tag{B.34}$$

we have that

$$\begin{aligned}&\mathrm{Tr}(\mathbf{W}^2)=\sum_{m,l}\langle y_m,y_l\rangle^2=n(n\rho^2+(1-\rho^2)),\\ &\mathrm{Tr}(\mathbf{W})=n,\quad \mathrm{Tr}(\mathbf{W}y_iy_i^T)=n\rho^2+(1-\rho^2),\quad \|\mathbf{P}_{y_i}y_j\|^2=1-\rho^2.\end{aligned} \tag{B.35}$$

Also,

$$\begin{aligned}&\mathbf{W}y_i=\sum_{m=1}^n\langle y_m,y_i\rangle y_m=(1-\rho)y_i+\rho\mathbf{V},\\ &\langle\mathbf{V},y_i\rangle=\sum_{m=1}^n\langle y_m,y_i\rangle=n\rho+(1-\rho),\\ &\|\mathbf{V}\|^2=\sum_{m,l}\langle y_m,y_l\rangle=n+\rho n(n-1)=n(n\rho+(1-\rho)),\\ &\|\mathbf{P}_{y_i}\mathbf{V}\|^2=\|\mathbf{V}\|^2-\langle\mathbf{V},y_i\rangle^2=(n-1)(n\rho+(1-\rho))(1-\rho),\\ &\|\mathbf{W}y_i\|^2=n^2\rho^3+3n\rho^2(1-\rho)+(1+2\rho)(1-\rho)^2.\end{aligned} \tag{B.36}$$

Proof of Lemma B.7. Direct Computations. □

Lemma B.8. *For terms $\mathbf{R}_1,\mathbf{R}_2$ in Lemma B.6, we have that*

$$\begin{aligned}&\sum_{i,j}\mathrm{Tr}\left[(\mathbf{R}_1+\mathbf{R}_2)^T(\mathbf{R}_1+\mathbf{R}_2)\right]\\ &=\beta^2e^{2\beta\rho}n\left[n^2\rho^2(1-\rho)+n(1-\rho)(1+\rho-3\rho^2)-(1+2\rho)(1-\rho)^2\right]\\ &+\beta e^{\beta(\rho+1)}n(n-1)(1-\rho^2)\\ &+e^{2\beta}(d-1)n+e^{2\beta\rho}\left[\beta^2(1-\rho^2)+d-1\right]n(n-1).\end{aligned} \tag{B.37}$$

As a corollary, when we pick $\beta=\gamma\log n$, we have the following phase transition limits as $n\to+\infty$:

(B.38)
$$\frac{1}{nZ^2}\sum_{i,j}\mathrm{Tr}\left[(\mathbf{R}_1+\mathbf{R}_2)^T(\mathbf{R}_1+\mathbf{R}_2)\right]=\begin{cases}\beta^2\rho^2(1-\rho)+o_n(1) & \text{if } \gamma<\frac{1}{1-\rho},\\ \frac{d-1+\beta^2\rho^2(1-\rho)}{4}+o_n(1) & \text{if } \gamma=\frac{1}{1-\rho},\\ d-1+o_n(1) & \text{if } \gamma>\frac{1}{1-\rho}.\end{cases}$$

Proof of Lemma B.8. We first notice that $\mathbf{W}$ is a symmetric matrix and $\|y_i\| = 1$. We then expand each term in Lemma B.8 and use Lemma B.7.

$$\begin{aligned}
&\sum_{i,j} \mathrm{Tr}\left[(\mathbf{R}_1)^T \mathbf{R}_1\right] \\
&= \beta^2 e^{2\beta\rho} \sum_i \left(\mathrm{Tr}\left(\mathbf{W}^2 - 2\mathbf{W} y_i (\mathbf{W} y_i)^T\right) + \|y_i\|^2 \|\mathbf{W} y_i\|^2\right) \\
&= \beta^2 e^{2\beta\rho} \sum_i \left(\mathrm{Tr}\mathbf{W}^2 - 2\|\mathbf{W} y_i\|^2 + \|\mathbf{W} y_i\|^2\right) \\
&= \beta^2 e^{2\beta\rho} n \left[n^2\rho^2(1-\rho) + n(1-\rho)(1+\rho-3\rho^2) - (1+2\rho)(1-\rho)^2\right].
\end{aligned} \tag{B.39}$$

Then,

$$\begin{aligned}
&\sum_{i,j} \mathrm{Tr}\left[(\mathbf{R}_1)^T \mathbf{R}_2\right] \\
&= \mathrm{Tr} \sum_{i,j} \delta_{ij} \beta e^{\beta\rho} \left(\mathbf{W} - \mathbf{W} y_i y_i^T\right) e^{\beta\langle y_j, y_i\rangle} \left((-y_i + \beta \mathbf{P}_{y_i} y_j) \otimes y_i + I_d\right) \\
&= \beta e^{\beta(\rho+1)} \mathrm{Tr} \sum_i \left(\mathbf{W} - \mathbf{W} y_i y_i^T\right)(-y_i y_i^T + I_d) = \beta e^{\beta(\rho+1)} \mathrm{Tr} \sum_i \left(\mathbf{W} - \mathbf{W} y_i y_i^T\right) \\
&= \beta e^{\beta(\rho+1)} n(n-1)(1-\rho^2),
\end{aligned} \tag{B.40}$$

where the second equality is because $\mathbf{P}_{y_i} y_i = 0$.

$$\begin{aligned}
&\sum_{i,j} \mathrm{Tr}\left[(\mathbf{R}_2)^T \mathbf{R}_2\right] \\
&= \sum_{i,j} e^{2\beta\langle y_j, y_i\rangle} \mathrm{Tr}\left[(-y_i + \beta \mathbf{P}_{y_i} y_j) y_i^T + I_d\right)\left(y_i (-y_i + \beta \mathbf{P}_{y_i} y_j)^T + I_d\right] \\
&= \sum_{i \neq j} e^{2\beta\rho} \left[\left(1 + \beta^2(1-\rho^2)\right) - 2 + d\right] + \sum_i e^{2\beta}(d-1) \\
&= e^{2\beta}(d-1)n + e^{2\beta\rho}\left[\beta^2(1-\rho^2) + d - 1\right] n(n-1).
\end{aligned} \tag{B.41}$$

Next, we show the asymptotics (B.38) as $n \to +\infty$. According to Lemma A.1, we have that

$$Z = \begin{cases} (1 + o_n(1)) \cdot n e^{\beta\rho} & \text{if } \gamma < \frac{1}{1-\rho}, \\ (1 + o_n(1)) \cdot e^{\beta} & \text{if } \gamma > \frac{1}{1-\rho}. \end{cases} \tag{B.42}$$

That is, when $\gamma < \frac{1}{1-\rho}$, the leading order terms are those terms involving $ne^{\beta\rho}$, and all the remaining terms go to 0 after dividing $ne^{\beta\rho}$; when $\gamma > \frac{1}{1-\rho}$, the leading order terms are those terms involving e^{β}, and all the remaining terms go to 0 after dividing e^{β}. Hence, when $\gamma < \frac{1}{1-\rho}$, the leading order term in (B.37) is the term $\beta^2 e^{2\beta\rho} n^3 \rho^2 (1-\rho)$; when $\gamma > \frac{1}{1-\rho}$, the leading order term is $e^{2\beta}(d-1)n$. This proves (B.38). □

Lemma B.9. *For terms* $\mathbf{R}_1, \mathbf{R}_2, \mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{U}_4$ *in Lemma B.6, we have that*

$$\begin{aligned}
&\sum_{i,j}(\mathbf{U}_1+\mathbf{U}_2)^T(\mathbf{R}_1+\mathbf{R}_2)(\mathbf{U}_3+\mathbf{U}_4)\\
&=\rho\beta^2e^{2\beta\rho}(e^\beta-e^{\beta\rho})n(n-1)(n\rho+(1-\rho))(1-\rho)\\
&\quad+\beta^2e^{3\beta\rho}n(n-1)(n\rho+(1-\rho))^2(1-\rho)\\
&\quad+\beta e^{\beta(2\rho+1)}n(n-1)(n\rho+(1-\rho))(1-\rho)\\
&\quad+\beta e^{2\beta\rho}(e^\beta-e^{\beta\rho})(\beta\rho+1)n(n-1)(1-\rho^2)\\
&\quad+\beta e^{3\beta\rho}n(n-1)(n\rho+(1-\rho))(\beta(1-\rho^2)+(1-\rho)).
\end{aligned}\tag{B.43}$$

As a corollary, when we pick $\beta=\gamma\log n$*, we have the following phase transition limits as* $n\to+\infty$*:*

$$\frac{1}{nZ^3}\sum_{i,j}(\mathbf{U}_1+\mathbf{U}_2)^T(\mathbf{R}_1+\mathbf{R}_2)(\mathbf{U}_3+\mathbf{U}_4)=\begin{cases}\beta^2\rho^2(1-\rho)+o_n(1) & \text{if } \gamma<\frac{1}{1-\rho},\\ \frac{\beta^2\rho^2(1-\rho)}{4}+o_n(1) & \text{if } \gamma=\frac{1}{1-\rho}.\\ 0+o_n(1) & \text{if } \gamma>\frac{1}{1-\rho}.\end{cases}\tag{B.44}$$

Proof of Lemma B.9. We expand each term in Lemma B.9 and also apply Lemma B.7 to each term. We first estimate terms involving $\mathbf{U}_1$.

$$\begin{aligned}
&\sum_{i,j}\mathbf{U}_1^T\mathbf{R}_1\mathbf{U}_3=\sum_i\beta^2e^{2\beta\rho}(e^\beta-e^{\beta\rho})\left(\mathbf{P}_{y_i}\mathbf{V}\right)^T\left(\mathbf{W}-y_i\otimes(\mathbf{W}y_i)\right)y_i\\
&=\beta^2e^{2\beta\rho}(e^\beta-e^{\beta\rho})\sum_i\left(\mathbf{P}_{y_i}\mathbf{V}\right)^T\mathbf{W}y_i=\rho\beta^2e^{2\beta\rho}(e^\beta-e^{\beta\rho})\sum_i\left(\mathbf{P}_{y_i}\mathbf{V}\right)^T\mathbf{V}\\
&=\rho\beta^2e^{2\beta\rho}(e^\beta-e^{\beta\rho})n(n-1)(n\rho+(1-\rho))(1-\rho),
\end{aligned}\tag{B.45}$$

where the second and the third equality is because $\langle\mathbf{P}_{y_i}\mathbf{V},y_i\rangle=0$.

$$\begin{aligned}
&\sum_{i,j}\mathbf{U}_1^T\mathbf{R}_1\mathbf{U}_4=\sum_i\beta^2e^{3\beta\rho}\left(\mathbf{P}_{y_i}\mathbf{V}\right)^T\left(\mathbf{W}-y_i\otimes(\mathbf{W}y_i)\right)\mathbf{V}\\
&=\beta^2e^{3\beta\rho}\sum_i\left(\mathbf{P}_{y_i}\mathbf{V}\right)^T\mathbf{W}\mathbf{V}=\beta^2e^{3\beta\rho}(n\rho+(1-\rho))\sum_i\|\mathbf{P}_{y_i}\mathbf{V}\|^2\\
&=\beta^2e^{3\beta\rho}n(n-1)(n\rho+(1-\rho))^2(1-\rho),
\end{aligned}\tag{B.46}$$

where the second equality is because $\langle\mathbf{P}_{y_i}\mathbf{V},y_i\rangle=0$.

$$\sum_{i,j}\mathbf{U}_1^T\mathbf{R}_2\mathbf{U}_3=\sum_i\beta e^{\beta(\rho+1)}(e^\beta-e^{\beta\rho})\left(\mathbf{P}_{y_i}\mathbf{V}\right)^T\left((-y_i+\beta\mathbf{P}_{y_i}y_i)\otimes y_i+I_d\right)y_i=0,\tag{B.47}$$

where the second equality is because $\langle\mathbf{P}_{y_i}\mathbf{V},y_i\rangle=0$ and $\mathbf{P}_{y_i}y_i=0$.

$$\begin{aligned}
&\sum_{i,j}\mathbf{U}_1^T\mathbf{R}_2\mathbf{U}_4=\sum_i\beta e^{\beta(2\rho+1)}\left(\mathbf{P}_{y_i}\mathbf{V}\right)^T\left((y_i+\beta\mathbf{P}_{y_i}y_i)\otimes y_i+I_d\right)\mathbf{V}\\
&=\beta e^{\beta(2\rho+1)}\sum_i\|\mathbf{P}_{y_i}\mathbf{V}\|^2=\beta e^{\beta(2\rho+1)}n(n-1)(n\rho+(1-\rho))(1-\rho),
\end{aligned}\tag{B.48}$$

where the second equality is because $\langle\mathbf{P}_{y_i}\mathbf{V},y_i\rangle=0$ and $\mathbf{P}_{y_i}y_i=0$.

Next, we estimate the terms involving $\mathbf{U}_2$. We recall that $\mathbf{U}_2=\beta e^{\beta\langle y_j,y_i\rangle}\left(\mathbf{P}_{y_i}y_j\right)$. Because $\mathbf{P}_{y_i}y_j=0$ when $i=j$, we can just replace $e^{\beta\langle y_j,y_i\rangle}$ with $e^{\beta\rho}$ in $\mathbf{U}_2$, i,e,

$\mathbf{U}_2 = \beta e^{\beta\rho}\,(\mathbf{P}_{y_i} y_j)$. Hence,

$$\sum_{i,j} \mathbf{U}_2^T \mathbf{R}_1 \mathbf{U}_3 = 0, \quad \sum_{i,j} \mathbf{U}_2^T \mathbf{R}_1 \mathbf{U}_4 = 0, \tag{B.49}$$

because $\delta_{ij}(\mathbf{P}_{y_i} y_j) = 0$ for any i, j in $\mathbf{U}_2^T \mathbf{R}_1$.

$$\begin{aligned}
&\sum_{i,j} \mathbf{U}_2^T \mathbf{R}_2 \mathbf{U}_3 = \sum_{i,j} \beta e^{\beta(\rho + \langle y_j, y_i\rangle)} (e^\beta - e^{\beta\rho}) \left(\mathbf{P}_{y_i} y_j\right)^T \left((-y_i + \beta \mathbf{P}_{y_i} y_j) \otimes y_i + I_d\right) y_j \\
&= \beta e^{2\beta\rho}(e^\beta - e^{\beta\rho}) \sum_{i \neq j} \left(\mathbf{P}_{y_i} y_j\right)^T \left((-y_i + \beta \mathbf{P}_{y_i} y_j)\rho + y_j\right) \\
&= \beta e^{2\beta\rho}(e^\beta - e^{\beta\rho})(\beta\rho + 1) \sum_{i \neq j} \|\mathbf{P}_{y_i} y_j\|^2 \\
&= \beta e^{2\beta\rho}(e^\beta - e^{\beta\rho})(\beta\rho + 1) n(n-1)(1-\rho^2).
\end{aligned} \tag{B.50}$$

where the second equality is because $\mathbf{P}_{y_i} y_j \neq 0$ only when $i \neq j$, on which $\langle y_j, y_i\rangle = \rho$, and the third equality is because $\langle \mathbf{P}_{y_i} y_j, y_i\rangle = 0$.

$$\begin{aligned}
&\sum_{i,j} \mathbf{U}_2^T \mathbf{R}_2 \mathbf{U}_4 = \sum_{i,j} \beta e^{\beta(2\rho + \langle y_j, y_i\rangle)} \left(\mathbf{P}_{y_i} y_j\right)^T \left((-y_i + \beta \mathbf{P}_{y_i} y_j) \otimes y_i + I_d\right) \mathbf{V} \\
&= \beta e^{3\beta\rho} \sum_{i \neq j} \left(\beta \|\mathbf{P}_{y_i} y_j\|^2 (n\rho + (1-\rho)) + \left(\mathbf{P}_{y_i} y_j\right)^T \mathbf{V}\right) \\
&= \beta e^{3\beta\rho} \sum_{i \neq j} \left(\beta(1-\rho^2)(n\rho + (1-\rho)) + (1-\rho)(n\rho + (1-\rho))\right) \\
&= \beta e^{3\beta\rho} n(n-1)(n\rho + (1-\rho))(\beta(1-\rho^2) + (1-\rho)).
\end{aligned} \tag{B.51}$$

where the second equality is because $\mathbf{P}_{y_i} y_j \neq 0$ only when $i \neq j$, on which $\langle y_j, y_i\rangle = \rho$, and the third equality is because $\langle \mathbf{P}_{y_i} y_j, y_i\rangle = 0$.

The proof for (B.44) is similar to the proof for (B.38) in Lemma B.8. Notice that when $\gamma < \frac{1}{1-\rho}$, we need to pick up terms involving $ne^{\beta\rho}$, and the leading order term in (B.43) is the one in the second line of (B.43), which is $\beta^2 n^4 e^{3\beta\rho}\rho^2(1-\rho)$; when $\gamma > \frac{1}{1-\rho}$, after diving nZ^3, all terms in (B.43) are $o_n(1)$ terms. □

Lemma B.10. *For terms $\mathbf{U}_1, \mathbf{U}_2, \mathbf{U}_3, \mathbf{U}_4$ in Lemma B.6, we have that*

$$\begin{aligned}
&\sum_{i,j} \|\mathbf{U}_1 + \mathbf{U}_2\|^2 \|\mathbf{U}_3 + \mathbf{U}_4\|^2 \\
&= \beta^2 e^{2\beta\rho} n(n-1)(n\rho + 2)(1-\rho) \\
&\quad \cdot \left[(e^\beta - e^{\beta\rho})^2 + 2e^{\beta\rho}(e^\beta - e^{\beta\rho})(n\rho + (1-\rho)) + e^{2\beta\rho} n(n\rho + (1-\rho))\right].
\end{aligned} \tag{B.52}$$

As a corollary, when we pick $\beta = \gamma \log n$, we have the following phase transition limits as $n \to +\infty$:

$$\frac{1}{nZ^4} \sum_{i,j} \|\mathbf{U}_1 + \mathbf{U}_2\|^2 \|\mathbf{U}_3 + \mathbf{U}_4\|^2 = \begin{cases} \beta^2\rho^2(1-\rho) + o_n(1) & \text{if } \gamma < \frac{1}{1-\rho}, \\ \frac{\beta^2\rho(1-\rho)(1+3\rho)}{16} + o_n(1) & \text{if } \gamma = \frac{1}{1-\rho}, \\ 0 + o_n(1) & \text{if } \gamma > \frac{1}{1-\rho}. \end{cases} \tag{B.53}$$

Proof of Lemma B.10. We notice that $\langle \mathbf{U}_1, \mathbf{U}_2\rangle = 0$ because $\delta_{ij}\mathbf{P}_{y_i}y_j = 0$ for any i, j. So,

$$\begin{aligned}\|\mathbf{U}_1+\mathbf{U}_2\|^2 &= \delta_{ij}\beta^2 e^{2\beta\rho}\|\mathbf{P}_{y_i}\mathbf{V}\|^2 + \beta^2 e^{2\beta\langle y_j,y_i\rangle}\|\mathbf{P}_{y_i}y_j\|^2 \\ &= \delta_{ij}\beta^2 e^{2\beta\rho}(n-1)(n\rho+(1-\rho))(1-\rho) + (1-\delta_{ij})\beta^2 e^{2\beta\rho}(1-\rho^2),\end{aligned} \tag{B.54}$$

where the second equality is because $e^{2\beta\langle y_j,y_i\rangle}\|\mathbf{P}_{y_i}y_j\|^2 \neq 0$ only if $i \neq j$, on which $e^{2\beta\langle y_j,y_i\rangle}\|\mathbf{P}_{y_i}y_j\|^2 = e^{2\beta\rho}(1-\rho^2)$.

$$\begin{aligned}\|\mathbf{U}_3+\mathbf{U}_4\|^2 &= (e^\beta - e^{\beta\rho})^2 + 2e^{\beta\rho}(e^\beta - e^{\beta\rho})\langle \mathbf{V}, y_j\rangle + e^{2\beta\rho}\|\mathbf{V}\|^2 \\ &= (e^\beta - e^{\beta\rho})^2 + 2e^{\beta\rho}(e^\beta - e^{\beta\rho})(n\rho+(1-\rho)) + e^{2\beta\rho}n(n\rho+(1-\rho)),\end{aligned} \tag{B.55}$$

which is independent of i, j. Hence,

$$\begin{aligned}&\sum_{i,j}\|\mathbf{U}_1+\mathbf{U}_2\|^2\|\mathbf{U}_3+\mathbf{U}_4\|^2 \\ &= \left[\beta^2 e^{2\beta\rho}n(n-1)(n\rho+(1-\rho))(1-\rho) + n(n-1)\beta^2 e^{2\beta\rho}(1-\rho^2)\right]\|\mathbf{U}_3+\mathbf{U}_4\|^2 \\ &= \beta^2 e^{2\beta\rho}n(n-1)(n\rho+2)(1-\rho)\|\mathbf{U}_3+\mathbf{U}_4\|^2 \\ &= \beta^2 e^{2\beta\rho}n(n-1)(n\rho+2)(1-\rho) \\ &\quad\cdot\left[(e^\beta - e^{\beta\rho})^2 + 2e^{\beta\rho}(e^\beta - e^{\beta\rho})(n\rho+(1-\rho)) + e^{2\beta\rho}n(n\rho+(1-\rho))\right].\end{aligned} \tag{B.56}$$

The proof for (B.53) is similar to the proof for (B.38) in Lemma B.8. Notice that when $\gamma < \frac{1}{1-\rho}$, we need to pick up terms involving $ne^{\beta\rho}$, and the leading order term in(B.52) is is $\beta^2 n^5 e^{4\beta\rho}\rho^2(1-\rho)$; when $\gamma > \frac{1}{1-\rho}$, after diving nZ^4, all terms in (B.43) are $o_n(1)$ terms. □

Proof of Theorem 2.3. As we have mentioned at the beginning of Appendix B.2, we only need to focus the case when $\gamma \leqslant \frac{1}{1-\rho}$, which follows directly from Lemma B.8, Lemma B.9, and Lemma B.10. We notice that, in these three lemmas, the leading order terms are the same, $\beta^2\rho^2(1-\rho)$, which cancels in (B.33). Hence, when $\gamma < \frac{1}{1-\rho}$, $\frac{1}{nd}\|\nabla_X X'\|^2 = 0 + o_n(1)$. When $\gamma = \frac{1}{1-\rho}$, we also only need to use the corresponding cases in these three lemmas and combine them in (B.33) to get the conclusion in Theorem 2.3. One remark is that under Assumption 1, we have that $n \leqslant d$ implicitly. So, when $\gamma = \frac{1}{1-\rho}$, terms in (B.33) involving $\frac{\beta^2}{d} = \frac{\gamma^2(\log n)^2}{d}$ also become $o_n(1)$. □

Appendix C. Modified Assumptions with More Median Phases

In this section, we modify Assumption 2, so that we can prove the existence of three different phases like Lemma A.1, Theorem 2.2, Theorem 2.4. We remark that we only showed the existence of two phases (two extrema) in Lemma A.1, Theorem 2.2, Theorem 2.4, but it doesn't mean under Assumption 2, there is no other transition phase between these two phases (two extrema). Under the following Assumption 3, we can show there are indeed at least three phases. Recall that for any $i \in [\![1, n]\!]$, we defined $y_i = N(x_i)$.

Assumption 3.

- *For any $i \in [\![1, n]\!]$, $\|x_i\|^2 \in [q_1, q_2]$ for some positive constants $q_1 \leqslant q_2$.*

- *There is a $\tau \in (0,1]$, four positive constants $\rho_3, \rho_4, \kappa_3, \kappa_4$ with $\rho_3 \leqslant \rho_4$, $\kappa_3 \leqslant \kappa_4$, and $\rho_4 < 1$, such that for any $i \in [\![1,n]\!]$, if we define*

$$\mathcal{K}_i = \{m \neq i \mid \langle y_m, y_i \rangle \in [\rho_3, \rho_4]\}, \tag{C.1}$$

then we have that

$$\kappa_3 \leqslant \frac{|\mathcal{K}_i|}{n^\tau} \leqslant \kappa_4. \tag{C.2}$$

- *For any $i \in [\![1,n]\!]$ and any $j \notin \mathcal{K}_i \cup \{i\}$, $\langle y_i, y_j \rangle \in [\rho_1, \rho_2]$ for some nonnegative constants ρ_1, ρ_2 satisfying $\rho_1 \leqslant \rho_2 < \rho_3 \leqslant \rho_4$.*
- *For technical reason, we further assume that $(1-\tau)(1-\rho_2) + \rho_2 < \rho_3$.*

Lemma C.1. *Let $\beta = \gamma \log n$ where γ is a positive constant. Under Assumption 2 and (2.3), for any $i \in [\![1,n]\!]$,*

$$Z_i = \begin{cases} (1+o_n(1)) \cdot \left(\sum_{m \notin \mathcal{K}_i \cup \{i\}} e^{a_{im}}\right) & \text{if } \gamma < \min\left\{\frac{1}{1-\rho_1}, \frac{1-\tau}{\rho_4-\rho_1}\right\}, \\ (1+o_n(1)) \cdot \left(\sum_{m \in \mathcal{K}_i} e^{a_{im}}\right) & \text{if } \frac{1-\tau}{\rho_3-\rho_2} < \gamma < \frac{\tau}{1-\rho_3}, \\ (1+o_n(1)) \cdot e^{\beta} & \text{if } \gamma > \max\left\{\frac{1}{1-\rho_2}, \frac{\tau}{1-\rho_4}\right\}, \end{cases} \tag{C.3}$$

where the terms $o_n(1)$ go to 0 as $n \to +\infty$ with speeds independent of i but only depending on $\gamma, \rho_1, \rho_2, \rho_3, \rho_4, \tau, \kappa_3, \kappa_4$.

Proof. The proof is similar to Lemma A.1. We notice that

$$\begin{aligned} Z_i &= e^{\beta} + \sum_{m \in \mathcal{K}_i} e^{a_{im}} + \sum_{m \notin \mathcal{K}_i \cup \{i\}} e^{a_{im}} \\ &= n^{\gamma} + \sum_{m \in \mathcal{K}_i} n^{\gamma \langle y_i, y_m \rangle} + \sum_{m \notin \mathcal{K}_i \cup \{i\}} n^{\gamma \langle y_i, y_m \rangle}. \end{aligned} \tag{C.4}$$

We also notice that $\kappa_3 n^\tau \leqslant |\mathcal{K}_i| \leqslant \kappa_4 n^\tau$ according to Assumption 3. Hence,

$$\kappa_3 n^{\tau+\gamma\rho_3} \leqslant |\mathcal{K}_i| \cdot n^{\gamma\rho_3} \leqslant \sum_{m \in \mathcal{K}_i} n^{\gamma \langle y_i, y_m \rangle} \leqslant |\mathcal{K}_i| \cdot n^{\gamma\rho_4} \leqslant \kappa_4 n^{\tau+\gamma\rho_4}, \tag{C.5}$$

and

$$(n - \kappa_4 n^\tau - 1) \cdot n^{\gamma\rho_1} \leqslant \sum_{m \notin \mathcal{K}_i \cup \{i\}} n^{\gamma \langle y_i, y_m \rangle} \leqslant (n - |\mathcal{K}_i| - 1) \cdot n^{\gamma\rho_2} \leqslant n^{1+\gamma\rho_2}. \tag{C.6}$$

When $\gamma < \min\left\{\frac{1}{1-\rho_1}, \frac{1-\tau}{\rho_4-\rho_1}\right\}$, the leading order term in Z_i is $\sum_{m \notin \mathcal{K}_i \cup \{i\}} n^{\gamma \langle y_i, y_m \rangle}$; when $\frac{1-\tau}{\rho_3-\rho_2} < \gamma < \frac{\tau}{1-\rho_3}$, the leading order term in Z_i is $\sum_{m \in \mathcal{K}_i} n^{\gamma \langle y_i, y_m \rangle}$; when $\gamma > \max\left\{\frac{1}{1-\rho_2}, \frac{\tau}{1-\rho_4}\right\}$, the leading order term in Z_i is n^γ. We also remark that the last assumption in Assumption 3 is to ensure the existence of the middle phase, i.e., $\frac{1-\tau}{\rho_3-\rho_2} < \gamma < \frac{\tau}{1-\rho_3}$. This finishes the proof for Lemma C.1 by similar arguments as in Lemma A.1. □

A direct corollary of Lemma C.1 is the following theorem.

Theorem C.2. *Under Assumption 2 and (2.3) we have the following phase transition phenomena: let $\beta = \gamma \log n$ where γ is a positive constant. For any $i \in [\![1,n]\!]$*

the updating dynamics (2.3) *can be written as*

$$x_i' = \alpha x_i + \begin{cases} \frac{\sum_{m\notin\mathcal{K}_i\cup\{i\}} e^{a_{im}} y_m}{\sum_{m\notin\mathcal{K}_i\cup\{i\}} e^{a_{im}}} + \mathbf{o}_n(1) & \text{if } \gamma < \min\left\{\frac{1}{1-\rho_1}, \frac{1-\tau}{\rho_4-\rho_1}\right\}, \\ \frac{\sum_{m\in\mathcal{K}_i} e^{a_{im}} y_m}{\sum_{m\in\mathcal{K}_i} e^{a_{im}}} + \mathbf{o}_n(1) & \text{if } \frac{1-\tau}{\rho_3-\rho_2} < \gamma < \frac{\tau}{1-\rho_3}, \\ y_i + \mathbf{o}_n(1) & \text{if } \gamma > \max\left\{\frac{1}{1-\rho_2}, \frac{\tau}{1-\rho_4}\right\}, \end{cases} \tag{C.7}$$

The terms $\mathbf{o}_n(1)$ *represent vectors in* $\mathbb{R}^d$ *with norms going to* 0 *as* $n \to +\infty$*, with a speed independent of* i *but only depending on* $\gamma, \rho_1, \rho_2, \rho_3, \rho_4, \tau, \kappa_3, \kappa_4$*.*

The proof of Theorem C.2 is similar to Lemma C.1 so we omit its proof.

References

[ABK+22] Pedro Abdalla, Afonso S Bandeira, Martin Kassabov, Victor Souza, Steven H Strogatz, and Alex Townsend. Expander graphs are globally synchronising, 2022. arXiv:2210.12788.

[BBC+23] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report, 2023. arXiv:2309.16609.

[BPA25a] Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. Emergence of meta-stable clustering in mean-field transformer models. In *International Conference on Learning Representations*, 2025.

[BPA25b] Giuseppe Bruno, Federico Pasqualotto, and Andrea Agazzi. A multi-scale analysis of mean-field transformers in the moderate interaction regime. *NeurIPS*, 2025.

[BPC20] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[CLC+25] Ziang Chen, Zhengjiang Lin, Shi Chen, Yury Polyanskiy, and Philippe Rigollet. Residual connections provably mitigate oversmoothing in graph neural networks, 2025. arXiv:2501.00762.

[CLPR25] Shi Chen, Zhengjiang Lin, Yury Polyanskiy, and Philippe Rigollet. Quantitative clustering in mean-field transformer models, 2025. arXiv:2504.14697.

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

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

[Der81] B. Derrida. Random-energy model: An exactly solvable model of disordered systems. *Phys. Rev. B*, 24:2613–2626, 1981.

[GG25] Alessio Giorlandino and Sebastian Goldt. Two failure modes of deep transformers and how to avoid them: a unified theory of signal propagation at initialisation, 2025. arXiv:2505.24333.

[GKPR24] Borjan Geshkovski, Hugo Koubbi, Yury Polyanskiy, and Philippe Rigollet. Dynamic metastability in the self-attention model, 2024. arXiv:2410.06833.

[GLPR24] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. The emergence of clusters in self-attention dynamics. *Advances in Neural Information Processing Systems*, 36, 2024.

[GLPR25] Borjan Geshkovski, Cyril Letrouit, Yury Polyanskiy, and Philippe Rigollet. A mathematical perspective on transformers. *Bull. Amer. Math. Soc.*, 2025.

[Hut89] Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

[HZRS16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[JMS25] Vishesh Jain, Clayton Mizgerd, and Mehtaab Sawhney. The random graph process is globally synchronizing, 2025. arXiv:2501.12205.

[KGPR25] Nikita Karagodin, Shu Ge, Yury Polyanskiy, and Philippe Rigollet. Normalization in attention dynamics, 2025. arXiv.

[KPR24] Nikita Karagodin, Yury Polyanskiy, and Philippe Rigollet. Clustering in causal attention masking, 2024. arXiv:2411.04990.

[Lio71] Jacques Louis Lions. *Optimal control of systems governed by partial differential equations*, volume 170. Springer, 1971.

[LLC+21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

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

[Nak25] Ken M Nakanishi. Scalable-softmax is superior for attention, 2025. arXiv:2501.19399.

[PLS+25] Krishna C Puvvada, Faisal Ladhak, Santiago Akle Serrano, Cheng-Ping Hsieh, Shantanu Acharya, Somshubra Majumdar, Fei Jia, Samuel Kriman, Simeng Sun, Dima Rekesh, et al. Swan-gpt: An efficient and scalable approach for long-context language modeling, 2025. arXiv:2504.08719.

[PQFS23] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. Yarn: Efficient context window extension of large language models, 2023. arXiv:2309.00071.

[PRY25] Yury Polyanskiy, Philippe Rigollet, and Andrew Yao. Synchronization of mean-field models on the circle, 2025. arXiv:2507.22857.

[RHW86] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

(SC) Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA
Email address: schen636@mit.edu

(ZL) Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA
Email address: linzj@mit.edu

(YP) Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA
Email address: yp@mit.edu

(PR) Department of Mathematics, Massachusetts Institute of Technology, 77 Massachusetts Ave, 02139 Cambridge MA, USA
Email address: rigollet@math.mit.edu