

## A Simulation Details and Supplementary Analysis

### A.1 General Experimental Setup

Throughout all simulations presented in this paper, we consistently set the spatial dimension to  $s = 4$ . The core experimental setup follows that described in Example 3.1. Specifically, the vectors  $x^1, \dots, x^s$  are outputs of MatMul, defined by

$$x^i = W^i h \quad (i \in [s]).$$

Each element of the initial vector  $h \in \mathbb{R}^n$  is sampled independently from a standard normal distribution.

For all weight matrices involved in the attention mechanism  $W^{Q,a}, W^{K,a}, W^{V,a}, W^{O,a}$  and the matrices  $W^i$  generating  $x^i$ , we set

$$\sigma_{W^{Q,a}}^2 = \sigma_{W^{K,a}}^2 = \sigma_{W^{V,a}}^2 = \sigma_{W^{O,a}}^2 = \sigma_{W^i}^2 = 1.$$

The elements of these weight matrices are independently sampled from  $\mathcal{N}(0, \sigma_W^2/n)$ , where  $\sigma_W^2$  is the respective variance (here, 1).

Under this setup, the input vectors  $x^j$  to the attention layer are designed such that their infinite-width limits  $Z^{x^j}$  are uncorrelated for  $j \neq j'$ , i.e.,

$$\mathbb{E}[Z^{x^j} Z^{x^{j'}}] = 0 \quad (j, j' \in [s], j \neq j').$$

Furthermore, the infinite-width limit  $Z^{x^i}$  has the variance

$$\mathbb{E}[(Z^{x^i})^2] = 1 \quad (i \in [s]).$$

Consequently, the covariances of the limiting variables for the vectors  $\tilde{v}^{a,j}$  and dot-product scores  $\tilde{p}_{i,j}^{(a)}$  simplify as described in Example 3.1.

$$\text{Cov}(Z^{\tilde{v}^{a,j}}, Z^{\tilde{v}^{a',j'}}) = \begin{cases} \mathbb{E}[(Z^{x^j})^2] & (a = a', j = j'), \\ 0 & (\text{otherwise}). \end{cases}$$

and

$$\text{Cov}(\tilde{p}_{i,j}^{(a)}, \tilde{p}_{i',j'}^{(a')}) = \begin{cases} \left(\mathbb{E}[(Z^{x^i})^2]\right)^2 & (a = a', i = i', j = j'), \\ 0 & (\text{otherwise}). \end{cases}$$

To estimate the empirical distributions of finite-width attention outputs and their corresponding infinite-width limits, we employ Monte Carlo sampling. For each such estimation, 300,000 samples are drawn. Kernel density estimation (KDE) is used to visualize these empirical distributions.

### A.2 Analysis of Low-Rank Attention

#### A.2.1 Specific Setup for Low-Rank Attention

In practice, large-scale Transformers typically assume a specific embedding dimensionality for multi-head self-attention layers. For head counts  $H$ , the embedding dimension  $n$  is set linearly as  $n = H n_H$ , where  $n_H$  denotes the head dimension and determines the sizes of weight matrices as  $W^{Q,a}, W^{K,a}, W^{V,a} \in \mathbb{R}^{n_H \times n}$ . Thus, the QK product becomes low-rank relative to the embedding dimension  $n$ , and the scaling factor is given by  $1/\sqrt{n_H}$  as follows:

$$p_{i,j}^{(a)} = \frac{1}{\sqrt{n_H}} (W^{Q,a} x^i)^\top (W^{K,a} x^j) \quad (i, j \in [s], a \in [H]).$$

For example, the original Transformer architecture sets  $H = 8$  and  $n_H = 64$ . Large-scale models often increase the number of heads to be on the order of the hidden embedding dimension [EXW+24], as seen in GPT-3 (175B parameters), which sets  $H = 96$  and  $n_H = 128$ .

Additionally, since the output from each head is also of dimension  $n_H$  through the value matrix, an output weight  $W^{O,a} \in \mathbb{R}^{n \times n_H}$  is applied to map it back to the  $n$ -dimensional input for the subsequent layer:

$$y^i = \sum_{a=1}^H \sum_{j=1}^s \text{SoftMax}_j(p_{i,1}^{(a)}, \dots, p_{i,s}^{(a)}) W^{O,a} W^{V,a} x^j \quad (i \in [s]).$$

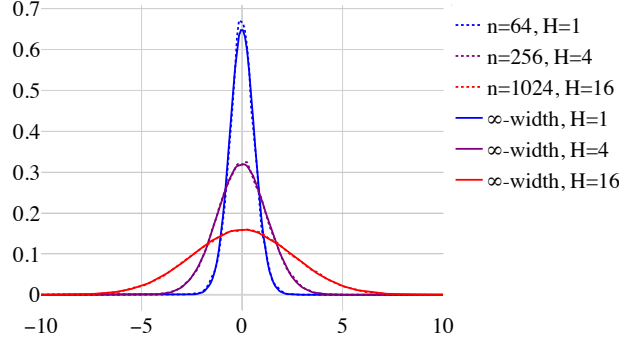


Figure 3: Comparison of the distribution of the attention output  $y_1^1$  under the low-rank setting and its infinite-width limit  $Z^y$ . Kernel density estimates of the empirical distribution (via Monte Carlo sampling) of  $y_1^1$  for various widths  $n$  and head counts  $H$  (with  $n_H = n/H = 64$  is fixed, dashed lines) alongside that of  $Z^y$  (solid lines). The plot shows convergence to the infinite-width limit as  $n$  and  $H$  increase proportionally.

Note that, to ensure the dot-product scores and the attention outputs are of order 1, the weight matrices are randomly initialized with the following scales:

$$W_{\alpha\beta}^{Q,a}, W_{\alpha\beta}^{K,a}, W_{\alpha\beta}^{V,a} \sim N(0, \sigma_W^2/n), \quad W_{\alpha\beta}^{O,a} \sim N(0, \sigma_W^2/n_H) \quad (\alpha, \beta \in [n]).$$

### A.2.2 Results and Discussion

Finally, we investigate the behavior of the attention output  $y_1^1$  in the low-rank regime described above, where the number of heads  $H$  increases proportionally with  $n$ . Fixing the head dimension  $n_H = n/H = 64$ , Figure 3 presents kernel density estimates of the distribution of  $y_1^1$  for  $(n, H) \in \{(64, 1), (256, 4), (1024, 16)\}$ , alongside the distributions of their corresponding infinite-width limit  $Z^y$ , approximated via Monte Carlo sampling. In all cases, the finite-width estimates (dashed lines) closely track the infinite-width limits (solid lines).

Notably, even in these practically relevant settings employing low-rank attention (where head-specific projections are  $n_H \times n$  or  $n \times n_H$  rather than the  $n \times n$  matrices primarily assumed in our formal derivations in Theorem 3.1), our infinite-width framework continues to provide an excellent approximation. This agreement suggests that the core principles of convergence captured by our theory extend robustly to common architectural variants like low-rank attention (with appropriate scaling considerations), underscoring the practical utility of our theoretical predictions under these structural assumptions common in modern attention models."

## B Mathematical Tools

### B.1 Basics

**Lemma B.1.** For  $1 \leq m \leq \infty$  and  $a_1, \dots, a_k \in \mathbb{R}$ , we have

$$\left| \sum_{i=1}^k a_i \right|^m \leq \left( \sum_{i=1}^k |a_i| \right)^m \leq k^{m-1} \sum_{i=1}^k |a_i|^m.$$

*Proof.* The first inequality is an application of the triangle inequality. The second inequality follows from Jensen's inequality. Since  $m \geq 1$ , the function  $x \mapsto x^m$  on  $[0, \infty)$  is convex, and thus Jensen's inequality implies

$$\left( \sum_{i=1}^k |a_i| \right)^m = k^m \left( \frac{1}{k} \sum_{i=1}^k |a_i| \right)^m \leq k^m \frac{1}{k} \sum_{i=1}^k |a_i|^m = k^{m-1} \sum_{i=1}^k |a_i|^m$$

as desired.  $\square$

817 **Lemma B.2** (Portmanteau lemma (Lemma 2.2 in [Vaa98])). *The following conditions are equivalent.*

818 (i)  $X_n \xrightarrow{d} X$ .

819 (ii)  $\mathbb{E}[f(X_n)] \rightarrow \mathbb{E}[f(X)]$  for all bounded and continuous function  $f$ .

820 (iii)  $\mathbb{E}[f(X_n)] \rightarrow \mathbb{E}[f(X)]$  for all bounded and Lipschitz function  $f$ .

821 (iv)  $\mathbb{P}(X_n \in B) \rightarrow \mathbb{P}(X \in B)$  for all Borel sets  $B$  with  $\mathbb{P}(X \in \delta B) = 0$ , where  $\delta B$  denotes the  
822 boundary of  $B$ .

823 **Fact B.1.** *Suppose  $\{X_n\}_{n \in \mathbb{N}}$  is a sequence of integrable random variables that converges in probability  
824 to  $X$ . Then the following statements are equivalent.*

825 (i) *The sequence  $\{X_n\}_{n \in \mathbb{N}}$  is uniformly integrable.*

826 (ii)  $\mathbb{E}(|X_n|) \rightarrow \mathbb{E}(|X|) < \infty$ .

827 **Remark B.1.** If  $X_n$  converges to 0 in probability, then by Fact [B.1] we have

$$\mathbb{E}(|X_n|) = o(1) \iff \{X_n\}_{n \in \mathbb{N}} \text{ is uniformly integrable.}$$

828 Moreover, since  $|\mathbb{E}(X_n)| \leq \mathbb{E}(|X_n|)$ , it follows that  $\mathbb{E}(X_n) = o(1)$ .

829 **Fact B.2.** *Suppose there exists  $\delta > 1$  such that  $\sup_n \mathbb{E}(|X_n|^\delta) < \infty$ . Then the sequence  $\{X_n\}_{n \in \mathbb{N}}$  is  
830 uniformly integrable.*

## 831 B.2 Pseudo-Lipschitz Functions

832 **Definition B.1** (Pseudo-Lipschitz functions [BM11]). Let  $d > 1$  be a constant. A function  $f : \mathbb{R}^k \rightarrow \mathbb{R}$   
833 is pseudo-Lipschitz of order  $d$  if there exists a constant  $C > 0$  such that, for all  $x, y \in \mathbb{R}^k$ ,

$$|f(x) - f(y)| \leq C\|x - y\|(1 + \|x\|^{d-1} + \|y\|^{d-1})$$

834 holds.

835 **Fact B.3.** *The following statements hold.*

836 (i) *A Lipschitz function is pseudo-Lipschitz of order  $d$  for all  $d > 1$ .*

837 (ii) *A pseudo-Lipschitz function (of any given order) is continuous.*

838 In this paper, we refer to a function as pseudo-Lipschitz if it is pseudo-Lipschitz of order  $d$  for some  
839  $d \in [2, \infty)$ .

840 **Proposition B.3.** *Suppose  $f : \mathbb{R}^k \rightarrow \mathbb{R}$  and  $g_i : \mathbb{R}^\ell \rightarrow \mathbb{R}$  ( $i \in [k]$ ) are pseudo-Lipschitz. Then the  
841 function  $h : \mathbb{R}^\ell \rightarrow \mathbb{R}$  defined by  $h(x) = f(g_1(x), \dots, g_k(x))$  is also pseudo-Lipschitz.*

842 *Proof.* Suppose  $f$  is pseudo-Lipschitz of order  $d_0 + 1$  and each  $g_i$  is pseudo-Lipschitz of order  $d_i + 1$ .

843 Define a function  $g : \mathbb{R}^\ell \rightarrow \mathbb{R}^k$  and a constant  $d \geq 1$  by

$$g(x) = (g_1(x), \dots, g_k(x)), \quad d = \max\{d_1, \dots, d_k\}.$$

844 Applying the pseudo-Lipschitz bounds for  $f$  and each  $g_i$  gives

$$|h(x) - h(x')| \lesssim \|g(x) - g(x')\| \left(1 + \|g(x)\|^{d_0} + \|g(x')\|^{d_0}\right)$$

845 and

$$|g_i(x) - g_i(x')| \lesssim \|x - x'\| \left(1 + \|x\|^{d_i} + \|x'\|^{d_i}\right) \lesssim \|x - x'\| \left(1 + \|x\|^d + \|x'\|^d\right).$$

846 The last inequality implies

$$\|g(x) - g(x')\| = \left( \sum_{i=1}^k |g_i(x) - g_i(x')|^2 \right)^{1/2} \lesssim \|x - x'\| \left(1 + \|x\|^d + \|x'\|^d\right).$$

847 On the other hand, since

$$\|g(x)\|^2 = \sum_{i=1}^k |g_i(x)|^2 \leq \left( \sum_{i=1}^k |g_i(x)| \right)^2$$

848 holds in general, Lemma B.1 implies

$$\|g(x)\|^{d_0} \leq \left( \sum_{i=1}^k |g_i(x)| \right)^{d_0} \lesssim \sum_{i=1}^k |g_i(x)|^{d_0}.$$

849 The pseudo-Lipschitz property of each  $g_i$  yields

$$|g_i(x)| \leq |g_i(0)| + |g_i(x) - g_i(0)| \lesssim 1 + \|x\|(1 + \|x\|^{d_i}) \lesssim 1 + \|x\|^{d_i+1} \lesssim 1 + \|x\|^{d+1}.$$

850 Hence, by Lemma B.1, we have

$$\|g(x)\|^{d_0} \lesssim (1 + \|x\|^{d+1})^{d_0} \lesssim 1 + \|x\|^{d_0(d+1)}.$$

851 Combining these elements gives

$$|h(x) - h(x')| \lesssim \|x - x'\| \left( 1 + \|x\|^d + \|x'\|^d \right) \left( 1 + \|x\|^{d_0(d+1)} + \|x'\|^{d_0(d+1)} \right).$$

852 We expand the product as

$$\begin{aligned} & (1 + \|x\|^d + \|x'\|^d)(1 + \|x\|^{d_0(d+1)} + \|x'\|^{d_0(d+1)}) \\ &= 1 + \|x\|^d + \|x'\|^d + \|x\|^{d_0(d+1)} + \|x'\|^{d_0(d+1)} + \|x\|^{d+d_0(d+1)} + \|x'\|^{d+d_0(d+1)} \\ & \quad + \|x\|^d \|x'\|^{d_0(d+1)} + \|x'\|^d \|x\|^{d_0(d+1)}. \end{aligned}$$

853 Observe that each of the first seven terms are bounded by  $1 + \|x\|^a + \|x'\|^a$ , where  $a$  is given by

$$a = d + d_0(d + 1).$$

854 For the remaining two terms, we apply the weighted AM-GM inequality to obtain

$$\|x\|^d \|x'\|^{d_0(d+1)} = (\|x\|^a)^{\frac{d}{a}} (\|x'\|^a)^{\frac{d_0(d+1)}{a}} \leq \frac{d}{a} \|x\|^a + \frac{d_0(d+1)}{a} \|x'\|^a \leq \|x\|^a + \|x'\|^a.$$

855 The same bound applies to  $\|x'\|^d \|x\|^{d_0(d+1)}$ . Therefore the entire product satisfies

$$(1 + \|x\|^d + \|x'\|^d)(1 + \|x\|^{d_0(d+1)} + \|x'\|^{d_0(d+1)}) \lesssim 1 + \|x\|^a + \|x'\|^a,$$

856 which implies that  $h$  is pseudo-Lipschitz of order  $a$ .  $\square$

857 **Lemma B.4.** Define  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  by  $f(x, y) = xy$ . Then  $f$  is pseudo-Lipschitz of order  $d$  for every  
858  $d \in [2, \infty)$ .

859 *Proof.* For  $(x, y), (x', y') \in \mathbb{R}^2$ , we have

$$\begin{aligned} |f(x, y) - f(x', y')| &= |xy - x'y'| = |x(y - y') + (x - x')y'| \\ &\leq |x||y - y'| + |x - x'||y'| \leq \|(x, y) - (x', y')\|(|x| + |y'|). \end{aligned}$$

860 Observe that for any  $d \geq 2$ , we have

$$|x| \leq \|(x, y)\| \leq 1 + \|(x, y)\|^{d-1}, \quad |y'| \leq \|(x', y')\| \leq 1 + \|(x', y')\|^{d-1},$$

861 and consequently, we have

$$|x| + |y'| \lesssim 1 + \|(x, y)\|^{d-1} + \|(x', y')\|^{d-1}.$$

862 This gives us

$$|f(x, y) - f(x', y')| \lesssim \|(x, y) - (x', y')\| (1 + \|(x, y)\|^{d-1} + \|(x', y')\|^{d-1}),$$

863 which shows that  $f$  is pseudo-Lipschitz of order  $d$ .  $\square$

## 864 C Remaining proofs

### 865 C.1 Proof of Theorem 4.1

866 In this section we provide detailed proofs for the results sketched in Section 4, thereby completing the  
867 proof of Theorem 4.1

868 Throughout,  $\mathbb{E}[\cdot | X]$  denotes the conditional expectation with respect to the  $\sigma$ -algebra  $\sigma(X)$ . Since  
869 conditional expectations are only defined up to almost-sure equality, we omit “a.s.” when writing  
870 “ $\stackrel{\text{a.s.}}{=}$ ” in this context.

### 871 C.1.1 Weak Convergence of the Dot Products

872 In this section we prove the following proposition.

873 **Proposition C.1.** *Under the assumptions of Theorem 4.1 the vector  $(p_1, \dots, p_r)$  converges in*  
874 *distribution to  $(\hat{p}_1, \dots, \hat{p}_r)$ , which is the Gaussian vector defined in Definition 3.1*

875 The proof of Proposition C.1 relies on the next lemma, which is an application of Theorem 2 in  
876 [BCRT58].

877 **Lemma C.2.** *For each  $n \in \mathbb{N}$ , let  $\{X_\alpha\}_{\alpha \in [n]}$  be an exchangeable sequence of random variables*  
878 *satisfying*

$$\mathbb{E}[X_\alpha] = 0, \quad \mathbb{E}[X_\alpha^2] = \sigma_n^2, \quad \sigma_n^2 \rightarrow \sigma_*^2 \geq 0 \quad (n \rightarrow \infty).$$

879 Set  $S = \sum_{\alpha=1}^n X_\alpha / \sqrt{n}$ . Assume the following conditions:

880 (a)  $\mathbb{E}(X_1 X_2) = o(1/n)$ .

881 (b)  $\lim_{n \rightarrow \infty} \mathbb{E}(X_1^2 X_2^2) = \sigma_*^4$ .

882 (c)  $\mathbb{E}(|X_1|^3) = o(\sqrt{n})$ .

883 Then,  $S$  converges in distribution to  $Z$ , where the random variable  $Z$  satisfies

$$Z \stackrel{\text{a.s.}}{=} 0 \quad \text{if } \sigma_*^2 = 0, \quad Z \sim N(0, \sigma_*^2) \quad \text{otherwise.}$$

884 We introduce the following notation. For any two matrices  $W^{i,j}$  and  $W^{i',j'}$  appearing in the program,  
885 define  $d_{(i,j)}^{(i',j')}$  by

$$d_{(i,j)}^{(i',j')} = \begin{cases} 1 & \text{(if } W^{i,j} \text{ and } W^{i',j'} \text{ are the same matrices),} \\ 0 & \text{(otherwise).} \end{cases}$$

886 It is important to note that  $d_{(i,j)}^{(i',j')}$  is always a deterministic value that is independ of  $n$ , and is fixed by  
887 the program architecture. According to the sampling rule explained in Section 2.1 the matrices  $W^{i,j}$   
888 and  $W^{i',j'}$  are sampled independently whenever  $d_{(i,j)}^{(i',j')}$  is zero. In particular, Assumption 3.1 gives

$$d_{(i,1)}^{(i,2)} = 0 \quad (i \in [r]). \quad (4)$$

889 Let  $t_1, \dots, t_r$  be arbitrary constants. We define

$$S = \sum_{i=1}^r t_i p_i = \sum_{i=1}^r t_i \left( \frac{1}{\sqrt{n}} \sum_{\alpha=1}^n \sum_{\gamma_1=1}^n \sum_{\gamma_2=1}^n W_{\alpha\gamma_1}^{i,1} W_{\alpha\gamma_2}^{i,2} x_{\gamma_1}^{i,1} x_{\gamma_2}^{i,2} \right) = \frac{1}{\sqrt{n}} \sum_{\alpha=1}^n X_\alpha,$$

890 where  $X_\alpha$  is given by

$$X_\alpha = \sum_{i=1}^r \sum_{\gamma_1, \gamma_2=1}^n t_i W_{\alpha\gamma_1}^{i,1} W_{\alpha\gamma_2}^{i,2} x_{\gamma_1}^{i,1} x_{\gamma_2}^{i,2}.$$

891 Lemmas C.3–C.8 show that the sequence  $\{X_\alpha\}_{\alpha \in [n]}$  satisfies the conditions of Lemma C.2. The  
892 proof of Proposition C.1 is completed by applying Lemma C.2 to  $S = \sum_{\alpha=1}^n X_\alpha / \sqrt{n}$  and then invoking  
893 the Cramér–Wold device.

894 **Lemma C.3.** *The sequence  $\{X_\alpha\}_{\alpha \in [n]}$  is exchangeable.*

895 *Proof.* By the sampling rule, an element  $W_{\alpha\beta}^{i,j}$  of the random matrix  $W^{i,j}$  independently and identically  
896 follows  $\mathcal{N}(0, \sigma_{W^{i,j}}^2/n)$  for  $\alpha, \beta \in [n]$ . Hence, conditional on  $\{x^{i,j} : i \in [r], j \in [2]\}$ , the random  
897 variables  $X_1, \dots, X_n$  are i.i.d. Hence, by de Finetti's theorem,  $\{X_\alpha\}_{\alpha \in [n]}$  is exchangeable.  $\square$

898 **Lemma C.4.**  $\mathbb{E}(X_\alpha) = 0$  holds for every  $\alpha \in [n]$ .

899 *Proof.* We compute  $\mathbb{E}(X_\alpha) = \sum_{i=1}^r \sum_{\gamma_1, \gamma_2=1}^n t_i \mathbb{E}(W_{\alpha\gamma_1}^{i,1}) \mathbb{E}(W_{\alpha\gamma_2}^{i,2}) \mathbb{E}(x_{\gamma_1}^{i,1} x_{\gamma_2}^{i,2}) = 0$ .  $\square$

900 **Lemma C.5.** We have  $\lim_{n \rightarrow \infty} \mathbb{E}(X_\alpha^2) = \sigma_*^2$  with

$$\begin{aligned} \sigma_*^2 &= \sum_{i_1, i_2=1}^r t_{i_1} t_{i_2} \mathbb{E} \left[ Z^{g^{i_1,1}} Z^{g^{i_1,2}} Z^{g^{i_2,1}} Z^{g^{i_2,2}} \right] \\ &= \sum_{i_1, i_2=1}^r \sum_{(j, j') \in J} t_{i_1} t_{i_2} \sigma_{W^{i_1,1}}^2 \sigma_{W^{i_1,2}}^2 \mathbb{E} \left[ Z^{x^{i_1,1}} Z^{x^{i_2,j}} \right] \mathbb{E} \left[ Z^{x^{i_1,2}} Z^{x^{i_2,j'}} \right] d_{(i_1,1)}^{(i_2,j)} d_{(i_1,2)}^{(i_2,j')}, \end{aligned}$$

901 where we defined  $J = \{(1, 2), (2, 1)\}$ .

902 *Proof.* For any  $\alpha \in [n]$ , we have

$$\begin{aligned} \mathbb{E}(X_\alpha^2) &= \mathbb{E} \left[ \left( \sum_{i_1=1}^r \sum_{\gamma_1, \gamma_2=1}^n t_{i_1} W_{\alpha\gamma_1}^{i_1,1} W_{\alpha\gamma_2}^{i_1,2} x_{\gamma_1}^{i_1,1} x_{\gamma_2}^{i_1,2} \right) \left( \sum_{i_2=1}^r \sum_{\gamma_3, \gamma_4=1}^n t_{i_2} W_{\alpha\gamma_3}^{i_2,1} W_{\alpha\gamma_4}^{i_2,2} x_{\gamma_3}^{i_2,1} x_{\gamma_4}^{i_2,2} \right) \right] \\ &= \sum_{i_1, i_2=1}^r \sum_{\gamma_1, \dots, \gamma_4=1}^n t_{i_1} t_{i_2} \mathbb{E} \left( W_{\alpha\gamma_1}^{i_1,1} W_{\alpha\gamma_2}^{i_1,2} W_{\alpha\gamma_3}^{i_2,1} W_{\alpha\gamma_4}^{i_2,2} \right) \mathbb{E} \left( x_{\gamma_1}^{i_1,1} x_{\gamma_2}^{i_1,2} x_{\gamma_3}^{i_2,1} x_{\gamma_4}^{i_2,2} \right) \\ &= \sum_{i_1, i_2=1}^r \sum_{\gamma_1, \gamma_2=1}^n \sum_{(j, j') \in J} t_{i_1} t_{i_2} \mathbb{E} \left[ (W_{\alpha\gamma_1}^{i_1,1})^2 \right] \mathbb{E} \left[ (W_{\alpha\gamma_2}^{i_1,2})^2 \right] \mathbb{E} \left( x_{\gamma_1}^{i_1,1} x_{\gamma_1}^{i_2,j} x_{\gamma_2}^{i_1,2} x_{\gamma_2}^{i_2,j'} \right) d_{(i_1,1)}^{(i_2,j)} d_{(i_1,2)}^{(i_2,j')} \\ &= \sum_{i_1, i_2=1}^r \sum_{(j, j') \in J} t_{i_1} t_{i_2} \sigma_{W^{i_1,1}}^2 \sigma_{W^{i_1,2}}^2 \mathbb{E} \left[ \left( \frac{1}{n} \sum_{\gamma_1=1}^n x_{\gamma_1}^{i_1,1} x_{\gamma_1}^{i_2,j} \right) \left( \frac{1}{n} \sum_{\gamma_2=1}^n x_{\gamma_2}^{i_1,2} x_{\gamma_2}^{i_2,j'} \right) \right] d_{(i_1,1)}^{(i_2,j)} d_{(i_1,2)}^{(i_2,j')} \\ &\xrightarrow{n \rightarrow \infty} \sum_{i_1, i_2=1}^r \sum_{(j, j') \in J} t_{i_1} t_{i_2} \sigma_{W^{i_1,1}}^2 \sigma_{W^{i_1,2}}^2 \mathbb{E} \left[ Z^{x^{i_1,1}} Z^{x^{i_2,j}} \right] \mathbb{E} \left[ Z^{x^{i_1,2}} Z^{x^{i_2,j'}} \right] d_{(i_1,1)}^{(i_2,j)} d_{(i_1,2)}^{(i_2,j')}, \end{aligned}$$

903 where the convergence follows from Lemma C.9. Finally, Eq. (4) and Definition 3.1 imply that

$$\sigma_{W^{i_1,1}}^2 \sigma_{W^{i_1,2}}^2 \mathbb{E} \left[ Z^{x^{i_1,1}} Z^{x^{i_2,j}} \right] \mathbb{E} \left[ Z^{x^{i_1,2}} Z^{x^{i_2,j'}} \right] d_{(i_1,1)}^{(i_2,j)} d_{(i_1,2)}^{(i_2,j')} = \mathbb{E} \left[ Z^{g^{i_1,1}} Z^{g^{i_1,2}} Z^{g^{i_2,1}} Z^{g^{i_2,2}} \right]$$

904 holds for any  $i_1, i_2 \in [r]$  and  $(j, j') \in J$ . □

905 **Lemma C.6.**  $\mathbb{E}(X_\alpha X_\beta) = 0$  holds for every  $\alpha, \beta \in [n], \alpha \neq \beta$ .

906 *Proof.* For  $\alpha \neq \beta$ , we have

$$\begin{aligned} \mathbb{E}(X_\alpha X_\beta) &= \mathbb{E} \left[ \left( \sum_{i_1=1}^r \sum_{\gamma_1, \gamma_2=1}^n t_{i_1} W_{\alpha\gamma_1}^{i_1,1} W_{\alpha\gamma_2}^{i_1,2} x_{\gamma_1}^{i_1,1} x_{\gamma_2}^{i_1,2} \right) \left( \sum_{i_2=1}^r \sum_{\gamma_3, \gamma_4=1}^n t_{i_2} W_{\beta\gamma_3}^{i_2,1} W_{\beta\gamma_4}^{i_2,2} x_{\gamma_3}^{i_2,1} x_{\gamma_4}^{i_2,2} \right) \right] \\ &= \sum_{i_1, i_2=1}^r \sum_{\gamma_1, \dots, \gamma_4=1}^n t_{i_1} t_{i_2} \mathbb{E} \left( W_{\alpha\gamma_1}^{i_1,1} \right) \mathbb{E} \left( W_{\alpha\gamma_2}^{i_1,2} \right) \mathbb{E} \left( W_{\beta\gamma_3}^{i_2,1} \right) \mathbb{E} \left( W_{\beta\gamma_4}^{i_2,2} \right) \mathbb{E} \left( x_{\gamma_1}^{i_1,1} x_{\gamma_2}^{i_1,2} x_{\gamma_3}^{i_2,1} x_{\gamma_4}^{i_2,2} \right) \\ &= 0 \end{aligned}$$

907 as desired. □

908 **Lemma C.7.**  $\lim_{n \rightarrow \infty} \mathbb{E}(X_\alpha^2 X_\beta^2) = \sigma_*^4$  holds for every  $\alpha, \beta \in [n], \alpha \neq \beta$ .

909 *Proof.* By a calculation similar to Lemma C.5, we have

$$\begin{aligned}
& \mathbb{E}(X_\alpha^2 X_\beta^2) \\
&= \sum_{i_1, \dots, i_4=1}^r \sum_{(j_1, j'_1), (j_2, j'_2) \in J^2} t_{i_1} t_{i_2} t_{i_3} t_{i_4} \sigma_{W^{i_1,1}}^2 \sigma_{W^{i_2,2}}^2 \sigma_{W^{i_3,1}}^2 \sigma_{W^{i_4,2}}^2 d_{(i_1,1)}^{(i_2, j_1)} d_{(i_1,2)}^{(i_2, j'_1)} d_{(i_3,1)}^{(i_4, j_2)} d_{(i_3,2)}^{(i_4, j'_2)} \\
&\quad \times \mathbb{E} \left[ \left( \frac{1}{n} \sum_{\gamma_1=1}^n x_{\gamma_1}^{i_1,1} x_{\gamma_1}^{i_2, j_1} \right) \left( \frac{1}{n} \sum_{\gamma_2=1}^n x_{\gamma_2}^{i_1,2} x_{\gamma_2}^{i_2, j'_1} \right) \left( \frac{1}{n} \sum_{\gamma_3=1}^n x_{\gamma_3}^{i_3,1} x_{\gamma_3}^{i_4, j_2} \right) \left( \frac{1}{n} \sum_{\gamma_4=1}^n x_{\gamma_4}^{i_3,2} x_{\gamma_4}^{i_4, j'_2} \right) \right] \\
&\xrightarrow{n \rightarrow \infty} \sum_{i_1, \dots, i_4=1}^r \sum_{(j_1, j'_1), (j_2, j'_2) \in J^2} t_{i_1} t_{i_2} t_{i_3} t_{i_4} \sigma_{W^{i_1,1}}^2 \sigma_{W^{i_2,2}}^2 \sigma_{W^{i_3,1}}^2 \sigma_{W^{i_4,2}}^2 d_{(i_1,1)}^{(i_2, j_1)} d_{(i_1,2)}^{(i_2, j'_1)} d_{(i_3,1)}^{(i_4, j_2)} d_{(i_3,2)}^{(i_4, j'_2)} \\
&\quad \times \mathbb{E} \left[ Z^{x^{i_1,1}} Z^{x^{i_2, j_1}} \right] \mathbb{E} \left[ Z^{x^{i_1,2}} Z^{x^{i_2, j'_1}} \right] \mathbb{E} \left[ Z^{x^{i_3,1}} Z^{x^{i_4, j_2}} \right] \mathbb{E} \left[ Z^{x^{i_3,2}} Z^{x^{i_4, j'_2}} \right],
\end{aligned}$$

910 where the convergence follows from Lemma C.9. Observe that this limit is equivalent to  $\sigma_*^4$ .  $\square$

911 **Lemma C.8.**  $\mathbb{E}(|X_\alpha|^3) = o(\sqrt{n})$  holds as  $n \rightarrow \infty$  for every  $\alpha \in [n]$ .

912 *Proof.* By the Lyapunov inequality, we have

$$\frac{1}{\sqrt{n}} \mathbb{E}(|X_\alpha|^3) \leq \frac{1}{\sqrt{n}} \left( \mathbb{E}(X_\alpha^4) \right)^{\frac{3}{4}} \leq \frac{1}{\sqrt{n}} \left( \sup_n \mathbb{E}(X_\alpha^4) \right)^{\frac{3}{4}}.$$

913 Thus, it suffices to show that  $\sup_n \mathbb{E}(X_\alpha^4) < \infty$  holds. We can express  $\mathbb{E}(X_\alpha^4)$  as

$$\begin{aligned}
\mathbb{E}(X_\alpha^4) &= \sum_{i_1, \dots, i_4=1}^r \sum_{\gamma_1, \dots, \gamma_8=1}^n t_{i_1} t_{i_2} t_{i_3} t_{i_4} \mathbb{E} \left[ W_{\alpha \gamma_1}^{i_1,1} W_{\alpha \gamma_2}^{i_2,1} W_{\alpha \gamma_3}^{i_3,1} W_{\alpha \gamma_4}^{i_4,1} W_{\alpha \gamma_5}^{i_1,2} W_{\alpha \gamma_6}^{i_2,2} W_{\alpha \gamma_7}^{i_3,2} W_{\alpha \gamma_8}^{i_4,2} \right] \\
&\quad \times \mathbb{E} \left[ x_{\gamma_1}^{i_1,1} x_{\gamma_2}^{i_2,1} x_{\gamma_3}^{i_3,1} x_{\gamma_4}^{i_4,1} x_{\gamma_5}^{i_1,2} x_{\gamma_6}^{i_2,2} x_{\gamma_7}^{i_3,2} x_{\gamma_8}^{i_4,2} \right].
\end{aligned}$$

914 Applying the Cauchy–Schwarz inequality yields

$$\begin{aligned}
& \mathbb{E} \left[ x_{\gamma_1}^{i_1,1} x_{\gamma_2}^{i_2,1} x_{\gamma_3}^{i_3,1} x_{\gamma_4}^{i_4,1} x_{\gamma_5}^{i_1,2} x_{\gamma_6}^{i_2,2} x_{\gamma_7}^{i_3,2} x_{\gamma_8}^{i_4,2} \right] \\
&\leq \left( \prod_{k=1}^4 \mathbb{E} \left[ (x_{\gamma_j}^{i_j,1})^8 \right] \right)^{\frac{1}{8}} \left( \prod_{k=1}^4 \mathbb{E} \left[ (x_{\gamma_{4+j}}^{i_j,2})^8 \right] \right)^{\frac{1}{8}} = \left( \prod_{k=1}^4 \mathbb{E} \left[ (x_1^{i_j,1})^8 \right] \right)^{\frac{1}{8}} \left( \prod_{k=1}^4 \mathbb{E} \left[ (x_1^{i_j,2})^8 \right] \right)^{\frac{1}{8}} \\
&\leq \sup_{i \in [r], j \in [2]} \mathbb{E} \left[ (x_1^{i,j})^8 \right].
\end{aligned}$$

915 By the boundedness of  $x^{i,j}$  ( $i \in [r]$ ,  $j \in [2]$ ), the last term is bounded uniformly in  $n$ , and  
916 consequently, it holds that

$$\sup_n \mathbb{E} \left[ x_{\gamma_1}^{i_1,1} x_{\gamma_2}^{i_2,1} x_{\gamma_3}^{i_3,1} x_{\gamma_4}^{i_4,1} x_{\gamma_5}^{i_1,2} x_{\gamma_6}^{i_2,2} x_{\gamma_7}^{i_3,2} x_{\gamma_8}^{i_4,2} \right] < \infty.$$

917 Hence, we compute

$$\begin{aligned}
\mathbb{E}(X_\alpha^4) &\lesssim \sum_{i_1, \dots, i_4=1}^r \sum_{\gamma_1, \dots, \gamma_8=1}^n \mathbb{E} \left[ W_{\alpha\gamma_1}^{i_1,1} W_{\alpha\gamma_2}^{i_2,1} W_{\alpha\gamma_3}^{i_3,1} W_{\alpha\gamma_4}^{i_4,1} W_{\alpha\gamma_5}^{i_1,2} W_{\alpha\gamma_6}^{i_2,2} W_{\alpha\gamma_7}^{i_3,2} W_{\alpha\gamma_8}^{i_4,2} \right] \\
&= \sum_{i=1}^r \sum_{\gamma_1, \dots, \gamma_8=1}^n \mathbb{E} \left[ W_{\alpha\gamma_1}^{i,1} W_{\alpha\gamma_2}^{i,1} W_{\alpha\gamma_3}^{i,1} W_{\alpha\gamma_4}^{i,1} W_{\alpha\gamma_5}^{i,2} W_{\alpha\gamma_6}^{i,2} W_{\alpha\gamma_7}^{i,2} W_{\alpha\gamma_8}^{i,2} \right] \\
&\quad + 4 \sum_{i_1=1}^r \sum_{i_2 \neq i_1} \sum_{\gamma_1, \dots, \gamma_8=1}^n \mathbb{E} \left[ W_{\alpha\gamma_1}^{i_1,1} W_{\alpha\gamma_2}^{i_1,1} W_{\alpha\gamma_3}^{i_1,1} W_{\alpha\gamma_4}^{i_2,1} W_{\alpha\gamma_5}^{i_1,2} W_{\alpha\gamma_6}^{i_1,2} W_{\alpha\gamma_7}^{i_1,2} W_{\alpha\gamma_8}^{i_2,2} \right] \\
&\quad + 3 \sum_{i_1=1}^r \sum_{i_2 \neq i_1} \sum_{\gamma_1, \dots, \gamma_8=1}^n \mathbb{E} \left[ W_{\alpha\gamma_1}^{i_1,1} W_{\alpha\gamma_2}^{i_1,1} W_{\alpha\gamma_3}^{i_2,1} W_{\alpha\gamma_4}^{i_2,1} W_{\alpha\gamma_5}^{i_1,2} W_{\alpha\gamma_6}^{i_1,2} W_{\alpha\gamma_7}^{i_2,2} W_{\alpha\gamma_8}^{i_2,2} \right] \\
&\quad + 6 \sum_{i_1=1}^r \sum_{i_2 \neq i_1} \sum_{i_3 \notin \{i_1, i_2\}} \sum_{\gamma_1, \dots, \gamma_8=1}^n \mathbb{E} \left[ W_{\alpha\gamma_1}^{i_1,1} W_{\alpha\gamma_2}^{i_1,1} W_{\alpha\gamma_3}^{i_2,1} W_{\alpha\gamma_4}^{i_3,1} W_{\alpha\gamma_5}^{i_1,2} W_{\alpha\gamma_6}^{i_1,2} W_{\alpha\gamma_7}^{i_2,2} W_{\alpha\gamma_8}^{i_3,2} \right] \\
&\quad + \sum_{\substack{i_1, i_2, i_3, i_4=1 \\ \text{all distinct}}}^r \sum_{\gamma_1, \dots, \gamma_8=1}^n \mathbb{E} \left[ W_{\alpha\gamma_1}^{i_1,1} W_{\alpha\gamma_2}^{i_2,1} W_{\alpha\gamma_3}^{i_3,1} W_{\alpha\gamma_4}^{i_4,1} W_{\alpha\gamma_5}^{i_1,2} W_{\alpha\gamma_6}^{i_2,2} W_{\alpha\gamma_7}^{i_3,2} W_{\alpha\gamma_8}^{i_4,2} \right] \\
&=: A_1 + 4A_2 + 3A_3 + 6A_4 + A_5.
\end{aligned}$$

918 Define  $\sigma$  by  $\sigma = \max\{\sigma_{W^{i,j}} : i \in [r], j \in [2]\}$ . Then, we compute  $A_1$  as

$$\begin{aligned}
A_1 &= \sum_{i=1}^r \sum_{\gamma_1, \dots, \gamma_8=1}^n \mathbb{E} \left[ W_{\alpha\gamma_1}^{i,1} W_{\alpha\gamma_2}^{i,1} W_{\alpha\gamma_3}^{i,1} W_{\alpha\gamma_4}^{i,1} \right] \mathbb{E} \left[ W_{\alpha\gamma_5}^{i,2} W_{\alpha\gamma_6}^{i,2} W_{\alpha\gamma_7}^{i,2} W_{\alpha\gamma_8}^{i,2} \right] \\
&= \sum_{i=1}^r \sum_{\gamma_1, \gamma_2=1}^n \mathbb{E} \left[ (W_{\alpha\gamma_1}^{i,1})^4 \right] \mathbb{E} \left[ (W_{\alpha\gamma_2}^{i,2})^2 \right] \\
&\quad + 3 \sum_{i=1}^r \sum_{(j,j') \in J} \sum_{\gamma_1, \gamma_2=1}^n \sum_{\gamma_3 \neq \gamma_2} \mathbb{E} \left[ (W_{\alpha\gamma_1}^{i,j})^4 \right] \mathbb{E} \left[ (W_{\alpha\gamma_2}^{i,j'})^2 \right] \mathbb{E} \left[ (W_{\alpha\gamma_3}^{i,j'})^2 \right] \\
&\quad + 9 \sum_{i=1}^r \sum_{\gamma_1, \gamma_2=1}^n \sum_{\gamma_3 \neq \gamma_1} \sum_{\gamma_4 \neq \gamma_2} \mathbb{E} \left[ (W_{\alpha\gamma_1}^{i,1})^2 \right] \mathbb{E} \left[ (W_{\alpha\gamma_2}^{i,1})^2 \right] \mathbb{E} \left[ (W_{\alpha\gamma_3}^{i,2})^2 \right] \mathbb{E} \left[ (W_{\alpha\gamma_4}^{i,2})^2 \right] \\
&= \sum_{i=1}^r \left( \frac{9\sigma_{W^{i,1}}^4 \sigma_{W^{i,2}}^4}{n^2} + 3 \sum_{(j,j') \in J} \frac{3(n-1)\sigma_{W^{i,j}}^4 \sigma_{W^{i,j'}}^4}{n^2} + 9 \frac{(n-1)^2 \sigma_{W^{i,1}}^4 \sigma_{W^{i,2}}^4}{n^2} \right) \\
&= 9 \sum_{i=1}^r \sigma_{W^{i,1}}^4 \sigma_{W^{i,2}}^4 \lesssim \sigma^8.
\end{aligned}$$

919 Likewise, we have

$$\begin{aligned}
A_2 &= \sum_{i_1=1}^r \sum_{i_2 \neq i_1} \sum_{(j,j') \in J} \sum_{\gamma_1, \dots, \gamma_8=1}^n \mathbb{E} \left[ W_{\alpha\gamma_1}^{i_1,1} W_{\alpha\gamma_2}^{i_1,1} W_{\alpha\gamma_3}^{i_2,j} W_{\alpha\gamma_4}^{i_2,j'} \right] \mathbb{E} \left[ W_{\alpha\gamma_5}^{i_1,2} W_{\alpha\gamma_6}^{i_1,2} W_{\alpha\gamma_7}^{i_2,j'} W_{\alpha\gamma_8}^{i_2,j} \right] d_{(i_1,1)}^{(i_2,j)} d_{(i_1,2)}^{(i_2,j')} \\
&= 9 \sum_{i_1=1}^r \sum_{i_2 \neq i_1} \sum_{(j,j') \in J} \sigma_{W^{i_1,1}}^4 \sigma_{W^{i_1,2}}^4 d_{(i_1,1)}^{(i_2,j)} d_{(i_1,2)}^{(i_2,j')} \lesssim \sigma^8
\end{aligned}$$

920 and

$$\begin{aligned}
A_3 &= \sum_{i_1=1}^r \sum_{i_2 \neq i_1} \sum_{(j,j') \in J} \sum_{\gamma_1, \dots, \gamma_8=1}^n \mathbb{E} \left[ W_{\alpha\gamma_1}^{i_1,1} W_{\alpha\gamma_2}^{i_1,1} W_{\alpha\gamma_3}^{i_2,j} W_{\alpha\gamma_4}^{i_2,j'} \right] \mathbb{E} \left[ W_{\alpha\gamma_5}^{i_1,2} W_{\alpha\gamma_6}^{i_1,2} W_{\alpha\gamma_7}^{i_2,j'} W_{\alpha\gamma_8}^{i_2,j} \right] d_{(i_1,1)}^{(i_2,j)} d_{(i_1,2)}^{(i_2,j')} \\
&= 9 \sum_{i_1=1}^r \sum_{i_2 \neq i_1} \sum_{(j,j') \in J} \sigma_{W^{i_1,1}}^4 \sigma_{W^{i_1,2}}^4 d_{(i_1,1)}^{(i_2,j)} d_{(i_1,2)}^{(i_2,j')} \lesssim \sigma^8.
\end{aligned}$$

921 Applying a similar argument, we can also show that  $A_4 \lesssim \sigma^8$  and  $A_5 \lesssim \sigma^8$  holds. Therefore, we  
922 conclude that  $\mathbb{E}(X_\alpha^4) \lesssim \sigma^8$  holds, which completes the proof.  $\square$



923 **Lemma C.9.** Take  $k_1, \dots, k_8 \in \{(i, j) : i \in [r], j \in [2]\}$  arbitrarily. Then, the following statements  
 924 hold as  $n \rightarrow \infty$ .

925 (i) For  $(x^{k_1}, \dots, x^{k_8})$ , we have

$$\begin{aligned} & \mathbb{E} \left[ \left( \frac{1}{n} \sum_{\gamma_1=1}^n x_{\gamma_1}^{k_1} x_{\gamma_1}^{k_2} \right) \left( \frac{1}{n} \sum_{\gamma_2=1}^n x_{\gamma_2}^{k_3} x_{\gamma_2}^{k_4} \right) \left( \frac{1}{n} \sum_{\gamma_3=1}^n x_{\gamma_3}^{k_5} x_{\gamma_3}^{k_6} \right) \left( \frac{1}{n} \sum_{\gamma_4=1}^n x_{\gamma_4}^{k_7} x_{\gamma_4}^{k_8} \right) \right] \\ & \rightarrow \mathbb{E} \left[ Z^{x^{k_1}} Z^{x^{k_2}} \right] \mathbb{E} \left[ Z^{x^{k_3}} Z^{x^{k_4}} \right] \mathbb{E} \left[ Z^{x^{k_5}} Z^{x^{k_6}} \right] \mathbb{E} \left[ Z^{x^{k_7}} Z^{x^{k_8}} \right]. \end{aligned}$$

926 (ii) For  $(x^{k_1}, \dots, x^{k_4})$ , we have

$$\mathbb{E} \left[ \left( \frac{1}{n} \sum_{\gamma_1=1}^n x_{\gamma_1}^{k_1} x_{\gamma_1}^{k_2} \right) \left( \frac{1}{n} \sum_{\gamma_2=1}^n x_{\gamma_2}^{k_3} x_{\gamma_2}^{k_4} \right) \right] \rightarrow \mathbb{E} \left[ Z^{x^{k_1}} Z^{x^{k_2}} \right] \mathbb{E} \left[ Z^{x^{k_3}} Z^{x^{k_4}} \right].$$

927 *Proof.* Define the residual term  $R_\ell$  by

$$R_\ell = \frac{1}{n} \sum_{\gamma_\ell=1}^n x_{\gamma_\ell}^{k_{2\ell-1}} x_{\gamma_\ell}^{k_{2\ell}} - \mathbb{E} \left[ Z^{x^{k_{2\ell-1}}} Z^{x^{k_{2\ell}}} \right]$$

928 for each  $i \in [4]$ . Define constants  $C$  and  $\tilde{R}$  by

$$C = \max_{\ell \in [4]} \left| \mathbb{E} \left[ Z^{x^{k_{2\ell-1}}} Z^{x^{k_{2\ell}}} \right] \right|, \quad \tilde{R} = \left( \max_{\ell \in [4]} \mathbb{E} (R_\ell^4) \right)^{\frac{1}{4}}.$$

929 Note that  $C$  is bounded by the boundedness of  $x^{i,j}$  for all  $i \in [r]$  and  $j \in [2]$  (see Definition 3.1).  
 930 Then, we can write

$$\begin{aligned} & \left| \mathbb{E} \left[ \prod_{\ell=1}^4 \left( \frac{1}{n} \sum_{\gamma_\ell=1}^n x_{\gamma_\ell}^{k_{2\ell-1}} x_{\gamma_\ell}^{k_{2\ell}} \right) \right] - \prod_{\ell=1}^4 \mathbb{E} \left[ Z^{x^{k_{2\ell-1}}} Z^{x^{k_{2\ell}}} \right] \right| \\ & = \left| \mathbb{E} \left[ \prod_{\ell=1}^4 \left( R_\ell + \mathbb{E} \left[ Z^{x^{k_{2\ell-1}}} Z^{x^{k_{2\ell}}} \right] \right) \right] - \prod_{\ell=1}^4 \mathbb{E} \left[ Z^{x^{k_{2\ell-1}}} Z^{x^{k_{2\ell}}} \right] \right| \\ & \leq 4C^3 \tilde{R} + 6C^2 \tilde{R}^2 + 4C \tilde{R}^3 + \tilde{R}^4, \end{aligned}$$

931 where the last inequality follows from the Cauchy–Schwarz inequality and the Lyapunov inequality.  
 932 Likewise, we have

$$\begin{aligned} & \left| \mathbb{E} \left[ \prod_{\ell=1}^2 \left( \frac{1}{n} \sum_{\gamma_\ell=1}^n x_{\gamma_\ell}^{k_{2\ell-1}} x_{\gamma_\ell}^{k_{2\ell}} \right) \right] - \prod_{\ell=1}^2 \mathbb{E} \left[ Z^{x^{k_{2\ell-1}}} Z^{x^{k_{2\ell}}} \right] \right| \\ & = \left| \mathbb{E} \left[ \prod_{\ell=1}^2 \left( R_\ell + \mathbb{E} \left[ Z^{x^{k_{2\ell-1}}} Z^{x^{k_{2\ell}}} \right] \right) \right] - \prod_{\ell=1}^2 \mathbb{E} \left[ Z^{x^{k_{2\ell-1}}} Z^{x^{k_{2\ell}}} \right] \right| \\ & \leq 2C \tilde{R} + \tilde{R}^2. \end{aligned}$$

933 Thus, it remains only to prove that  $\tilde{R}$  converges to 0 as  $n \rightarrow \infty$ . This can be achieved by showing  
 934 that  $\mathbb{E}(R_\ell^4)$  converges to 0 for all  $\ell \in [4]$ . By Fact 3.1 and Lemma B.4 for all  $\ell \in [4]$ , we know  $R_\ell$   
 935 converges almost surely to 0. The continuous mapping theorem then yields

$$R_\ell^4 \xrightarrow{a.s.} 0.$$

936 To upgrade this to convergence in expectation, Facts B.1 and B.2 imply it suffices to show the existence  
 937 of a constant  $\delta > 1$  that satisfies

$$\sup_n \mathbb{E}(R_\ell^{4+\delta}) < \infty.$$

938 But since each  $x^{i,j}$  and its infinite-width limit  $Z^{x^{i,j}}$  are bounded, such a  $\delta$  exists. Therefore, we  
 939 conclude that  $\mathbb{E}(R_\ell^4)$  converges to 0 for all  $\ell \in [4]$ , and consequently,  $\tilde{R}$  does as well.  $\square$

### 940 C.1.2 $S_1$ Converges to 0

941 We study the convergence of the term  $S_1$  defined in Section 4. Specifically, we show

$$S_1 = \left| \mathbb{E}f\left(\frac{1}{n} \sum_{\alpha=1}^n \psi(\mathbf{g}_\alpha^{1:M}, \mathbf{p}_{1:r})\right) - \mathbb{E}f\left(\mathbb{E}\left[\psi(Z^{\mathbf{g}^{1:M}}, \mathbf{p}_{1:r}) \mid \mathbf{p}_{1:r}\right]\right) \right| \rightarrow 0.$$

942 Fix a small  $\epsilon > 0$ . Since  $\hat{\mathbf{p}}_{1:r}$  is a Gaussian vector (see Definition 3.1), it is tight. Hence there exists a  
 943 compact set  $K \subset \mathbb{R}^r$  that satisfies

$$\mathbb{P}(\hat{\mathbf{p}}_{1:r} \in K) > 1 - \epsilon.$$

944 Set

$$\mathbf{y} = \frac{1}{n} \sum_{\alpha=1}^n \psi(\mathbf{g}_\alpha^{1:M}, \mathbf{p}_{1:r}), \quad \mathbf{z} = \mathbb{E}\left[\psi(Z^{\mathbf{g}^{1:M}}, \mathbf{p}_{1:r}) \mid \mathbf{p}_{1:r}\right].$$

945 Then, we have

$$\begin{aligned} S_1 &= |\mathbb{E}f(\mathbf{y}) - \mathbb{E}f(\mathbf{z})| \\ &\leq |\mathbb{E}[(f(\mathbf{y}) - f(\mathbf{z}))1\{\mathbf{p}_{1:r} \in K\}]| + |\mathbb{E}[(f(\mathbf{y}) - f(\mathbf{z}))1\{\mathbf{p}_{1:r} \notin K\}]| \\ &\leq |\mathbb{E}[(f(\mathbf{y}) - f(\mathbf{z}))1\{\mathbf{p}_{1:r} \in K\}]| + \mathbb{E}[|f(\mathbf{y}) - f(\mathbf{z})|1\{\mathbf{p}_{1:r} \notin K\}]. \end{aligned} \quad (5)$$

946 By the boundedness of  $f$ , the second term of Eq. (5) is bounded as

$$\mathbb{E}[|f(\mathbf{y}) - f(\mathbf{z})|1\{\mathbf{p}_{1:r} \notin K\}] \leq C\mathbb{P}(\mathbf{p}_{1:r} \notin K)$$

947 with some constant  $C$ . Furthermore, noting that  $K$  is a Borel set, applying Lemma B.2 and  
 948 Proposition C.1 gives

$$\lim_{n \rightarrow \infty} \mathbb{P}(\mathbf{p}_{1:r} \notin K) = \mathbb{P}(\hat{\mathbf{p}}_{1:r} \notin K) < \epsilon.$$

949 Therefore, for large enough  $n$ , we have

$$\mathbb{E}[|f(\mathbf{y}) - f(\mathbf{z})|1\{\mathbf{p}_{1:r} \notin K\}] < C\epsilon.$$

950 For the first term of Eq. (5), we have

$$\begin{aligned} &|\mathbb{E}[(f(\mathbf{y}) - f(\mathbf{z}))1\{\mathbf{p}_{1:r} \in K\}]| \\ &\leq \sup_{\mathbf{a}_{1:r} \in K} \left| \mathbb{E}\left[f\left(\frac{1}{n} \sum_{\alpha=1}^n \psi(\mathbf{g}_\alpha^{1:M}, \mathbf{a}_{1:r})\right)\right] - \mathbb{E}\left[f\left(\mathbb{E}\left[\psi(Z^{\mathbf{g}^{1:M}}, \mathbf{a}_{1:r})\right]\right)\right] \right| \\ &= \sup_{\mathbf{a}_{1:r} \in K} \Phi_n(\mathbf{a}_{1:r}), \end{aligned}$$

951 where  $\Phi_n : \mathbb{R}^r \rightarrow \mathbb{R}$  is defined by

$$\Phi_n(\mathbf{a}_{1:r}) = \left| \mathbb{E}\left[f\left(\frac{1}{n} \sum_{\alpha=1}^n \psi(\mathbf{g}_\alpha^{1:M}, \mathbf{a}_{1:r})\right)\right] - \mathbb{E}\left[f\left(\mathbb{E}\left[\psi(Z^{\mathbf{g}^{1:M}}, \mathbf{a}_{1:r})\right]\right)\right] \right|.$$

952 Define  $\tilde{\psi}_{\mathbf{a}_{1:r}} : \mathbb{R}^M \rightarrow \mathbb{R}$  by

$$\tilde{\psi}_{\mathbf{a}_{1:r}}(\mathbf{u}_{1:M}) = \psi(\mathbf{u}_{1:M}, \mathbf{a}_{1:r}). \quad (6)$$

953 Observe that  $\tilde{\psi}_{\mathbf{a}_{1:r}}$  is bounded by the boundedness of  $\psi$ . We later show that  $\tilde{\psi}_{\mathbf{a}_{1:r}}$  is also pseudo-  
 954 Lipschitz (Lemma C.10). Therefore, by Eq. (3) and Lemma B.2, we have

$$\lim_{n \rightarrow \infty} \Phi_n(\mathbf{a}_{1:r}) = 0 \quad (\mathbf{a}_{1:r} \in \mathbb{R}^r).$$

955 Moreover, noting that  $f$  and  $\psi$  are bounded and continuous (see Fact B.3), we can apply the  
 956 (unconditional and conditional) bounded convergence theorem to show that  $\Phi_n$  is continuous at every  
 957 point. Therefore, we can apply Lemma C.11 to show that  $\sup_{\mathbf{a}_{1:r} \in K} \Phi_n(\mathbf{a}_{1:r})$  converges to 0 as  
 958  $n \rightarrow \infty$ .

959 **Lemma C.10.** Define  $\tilde{\psi}_{\mathbf{a}_{1:r}} : \mathbb{R}^M \rightarrow \mathbb{R}$  by Eq. (6). Then, it is pseudo-Lipschitz.

960 *Proof.* Suppose  $\psi$  is pseudo-Lipschitz of order  $d + 1$  with  $d \geq 1$ . Then, we have

$$\begin{aligned} |\tilde{\psi}_{\mathbf{a}_{1:r}}(\mathbf{u}_{1:M}) - \tilde{\psi}_{\mathbf{a}_{1:r}}(\mathbf{u}'_{1:M})| &= |\psi(\mathbf{u}_{1:M}, \mathbf{a}_{1:r}) - \psi(\mathbf{u}'_{1:M}, \mathbf{a}_{1:r})| \\ &\lesssim \|(\mathbf{u}_{1:M}) - (\mathbf{u}'_{1:M})\| (1 + \|(\mathbf{u}_{1:M}, \mathbf{a}_{1:r})\|^d + \|(\mathbf{u}'_{1:M}, \mathbf{a}_{1:r})\|^d). \end{aligned}$$

961 By Lemma B.1, we can bound  $\|(\mathbf{u}_{1:M}, \mathbf{a}_{1:r})\|^d$  as

$$\|(\mathbf{u}_{1:M}, \mathbf{a}_{1:r})\|^d \leq (\|(\mathbf{u}_{1:M})\| + \|\mathbf{a}_{1:r}\|)^d \leq 2^{d-1} (\|(\mathbf{u}_{1:M})\|^d + \|\mathbf{a}_{1:r}\|^d).$$

962 Thus, we have

$$\begin{aligned} 1 + \|(\mathbf{u}_{1:M}, \mathbf{a}_{1:r})\|^d + \|(\mathbf{u}'_{1:M}, \mathbf{a}_{1:r})\|^d &\leq 1 + 2^{d-1} (\|(\mathbf{u}_{1:M})\|^d + \|(\mathbf{u}'_{1:M})\|^d) + 2^d \|\mathbf{a}_{1:r}\|^d \\ &\lesssim 1 + \|(\mathbf{u}_{1:M})\|^d + \|(\mathbf{u}'_{1:M})\|^d, \end{aligned}$$

963 where the last inequality holds because  $\mathbf{a}_{1:r}$  is fixed.  $\square$

964 **Lemma C.11.** Let  $K \subset \mathbb{R}^r$  be a compact set. Suppose for each  $n \in \mathbb{N}$ ,  $f_n : \mathbb{R} \rightarrow \mathbb{R}$  is a continuous  
965 function that satisfies

$$\lim_{n \rightarrow \infty} f_n(\mathbf{a}_{1:r}) = 0 \quad (7)$$

966 for any constant  $\mathbf{a}_{1:r} \in K$ . Then, we have

$$\lim_{n \rightarrow \infty} \sup_{\mathbf{a}_{1:r} \in K} |f_n(\mathbf{a}_{1:r})| = 0.$$

967 *Proof.* Fix  $\epsilon > 0$ . Let  $B(\mathbf{a}_{1:r}, \epsilon)$  denote the ball of radius  $\epsilon$  centered at  $\mathbf{a}_{1:r} \in \mathbb{R}^r$ . By the continuity  
968 of  $f_n$ , for each  $\mathbf{a}_{1:r} \in K$ , there exists  $\delta_{\mathbf{a}_{1:r}} > 0$  such that

$$|f_n(\mathbf{a}_{1:r}) - f_n(\mathbf{b}_{1:r})| < \epsilon/2 \quad (\mathbf{b}_{1:r} \in B(\mathbf{a}_{1:r}, \delta_{\mathbf{a}_{1:r}}))$$

969 holds. Note that  $K$  is covered by the union of balls as

$$K \subset \bigcup_{\mathbf{a}_{1:r} \in K} B(\mathbf{a}_{1:r}, \delta_{\mathbf{a}_{1:r}}).$$

970 Since  $K$  is compact, we can cover  $K$  with finitely many balls, say

$$K \subset \bigcup_{i=1}^I B_i = \bigcup_{i=1}^I B(\mathbf{a}_{1:r}^i, \delta_{\mathbf{a}_{1:r}^i}),$$

971 where  $B_i$  is given by  $B_i = B(\mathbf{a}_{1:r}^i, \delta_{\mathbf{a}_{1:r}^i})$ . By Eq. (7), for each  $i \in [I]$ , there exists  $N_i \in \mathbb{N}$  such that

$$|f_n(\mathbf{a}_{1:r}^i)| < \epsilon/2 \quad (n \geq N_i)$$

972 holds. Let  $N$  denote the maximum of  $\{N_i : i \in [I]\}$ . Then, for any  $n \geq N$ , we have

$$\begin{aligned} \sup_{\mathbf{a}_{1:r} \in K} |f_n(\mathbf{a}_{1:r})| &\leq \max_{i \in [I]} \sup_{\mathbf{a}_{1:r} \in B_i} |f_n(\mathbf{a}_{1:r})| \leq \max_{i \in [I]} \sup_{\mathbf{a}_{1:r} \in B_i} (|f_n(\mathbf{a}_{1:r}) - f_n(\mathbf{a}_{1:r}^i)| + |f_n(\mathbf{a}_{1:r}^i)|) \\ &< \max_{i \in [I]} \epsilon = \epsilon, \end{aligned}$$

973 which implies the convergence of  $\sup_{\mathbf{a}_{1:r} \in K} |f_n(\mathbf{a}_{1:r})|$  to zero as  $n$  goes to infinity.  $\square$

### 974 C.1.3 $S_2$ Converges to 0

975 We study the convergence of the term  $S_2$  also defined in Section 4. Specifically, we prove

$$S_2 = \left| \mathbb{E} f \left( \mathbb{E} \left[ \psi(Z^{g^{1:M}}, \mathbf{p}_{1:r}) \mid \mathbf{p}_{1:r} \right] \right) - \mathbb{E} f \left( \mathbb{E} [\psi(Z^{g^{1:M}}, \hat{\mathbf{p}}_{1:r}) \mid \hat{\mathbf{p}}_{1:r}] \right) \right| \rightarrow 0.$$

976 Define a function  $\Psi : \mathbb{R}^r \rightarrow \mathbb{R}$  by

$$\Psi(\mathbf{a}_{1:r}) = \mathbb{E} \left[ \psi(Z^{g^{1:M}}, \mathbf{a}_{1:r}) \right],$$

977 then it is bounded since  $\psi$  is bounded. By the bounded convergence theorem, it is also continuous.  
 978 Therefore, by the continuous mapping theorem, we have

$$\Psi(p_{1:r}) \xrightarrow{d} \Psi(\hat{p}_{1:r}).$$

979 Since  $Z^{g^{1:M}}$  is independent of  $p_{1:r}$ , we have

$$\mathbb{E} \left[ \psi(Z^{g^{1:M}}, p_{1:r}) \mid p_{1:r} \right] = \Psi(p_{1:r}).$$

980 Therefore, we have

$$\mathbb{E} \left[ \psi(Z^{g^{1:M}}, p_{1:r}) \mid p_{1:r} \right] \xrightarrow{d} \Psi(\hat{p}_{1:r}).$$

981 Note that  $\Psi(\hat{p}_{1:r})$  can be expressed as

$$\Psi(\hat{p}_{1:r}) = \mathbb{E} \left[ \psi(Z^{g^{1:M}}, \hat{p}_{1:r}) \mid \hat{p}_{1:r} \right],$$

982 where  $Z^{g^{1:M}}$  is independent of  $\hat{p}_{1:r}$ . By Lemma B.2, this implies  $S_2 \rightarrow 0$ .

## 983 C.2 Proof of Corollary 3.2

984 Let  $\psi : \mathbb{R}^J \rightarrow \mathbb{R}$  be a bounded and Lipschitz function that satisfies  $|\psi| \leq C$ . Note that by Fact B.3, it  
 985 is also pseudo-Lipschitz. Define a bounded and continuous function  $f : \mathbb{R} \rightarrow \mathbb{R}$  by

$$f(x) = -C1\{x < -C\} + x1\{x \in [-C, C]\} + C\{x \geq C\}.$$

986 Then, Lemma B.2 and Theorem 3.1 imply that

$$\begin{aligned} \mathbb{E}[\psi(h_\alpha^1, \dots, h_\alpha^J)] &= \mathbb{E} \left[ \frac{1}{n} \sum_{\alpha=1}^n \psi(h_\alpha^1, \dots, h_\alpha^J) \right] = \mathbb{E} \left[ f \left( \frac{1}{n} \sum_{\alpha=1}^n \psi(h_\alpha^1, \dots, h_\alpha^J) \right) \right] \\ &\rightarrow \mathbb{E} \left[ f \left( \mathbb{E}[\psi(Z^{h^1}, \dots, Z^{h^J}) \mid \hat{p}_1, \dots, \hat{p}_r] \right) \right] = \mathbb{E}[\psi(Z^{h^1}, \dots, Z^{h^J})] \end{aligned}$$

987 holds as  $n \rightarrow \infty$ . Since the above convergence holds for all bounded and Lipschitz function  $\psi$ , by  
 988 Lemma B.2, this implies the desired convergence in distribution.

## 989 NeurIPS Paper Checklist

### 990 1. Claims

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

993 Answer: [Yes]

994 Justification: We clearly stated the main claim in the abstract and introduction.

995 Guidelines:

- 996 • The answer NA means that the abstract and introduction do not include the claims made  
 997 in the paper.
- 998 • The abstract and/or introduction should clearly state the claims made, including the  
 999 contributions made in the paper and important assumptions and limitations. A No or  
 1000 NA answer to this question will not be perceived well by the reviewers.
- 1001 • The claims made should match theoretical and experimental results, and reflect how  
 1002 much the results can be expected to generalize to other settings.
- 1003 • It is fine to include aspirational goals as motivation as long as it is clear that these goals  
 1004 are not attained by the paper.

### 1005 2. Limitations

1006 Question: Does the paper discuss the limitations of the work performed by the authors?

1007 Answer: [Yes]

1008 Justification: We stated the limitation in the last section for discussion.

Guidelines:

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

### 3. Theory assumptions and proofs

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

Answer: [\[Yes\]](#)

Justification: We have described the full assumption, statements, and proof in the main body and the appendix.

Guidelines:

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

### 4. Experimental result reproducibility

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

Answer: [\[Yes\]](#)

Justification: We have described the detailed experimental details in the corresponding section in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.

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

## 5. Open access to data and code

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

Answer: [Yes]

Justification: We will open the source code.

Guidelines:

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

1117	<b>6. Experimental setting/details</b>
1118	Question: Does the paper specify all the training and test details (e.g., data splits, hyper-
1119	parameters, how they were chosen, type of optimizer, etc.) necessary to understand the
1120	results?
1121	Answer: [Yes]
1122	Justification: We have clearly described the experimental details in the appendix.
1123	Guidelines:
1124	• The answer NA means that the paper does not include experiments.
1125	• The experimental setting should be presented in the core of the paper to a level of detail
1126	that is necessary to appreciate the results and make sense of them.
1127	• The full details can be provided either with the code, in appendix, or as supplemental
1128	material.
1129	<b>7. Experiment statistical significance</b>
1130	Question: Does the paper report error bars suitably and correctly defined or other appropriate
1131	information about the statistical significance of the experiments?
1132	Answer: [Yes]
1133	Justification: We have repeated the experimental multiple times and report the standard
1134	deviation comes from the replication.
1135	Guidelines:
1136	• The answer NA means that the paper does not include experiments.
1137	• The authors should answer "Yes" if the results are accompanied by error bars, confidence
1138	intervals, or statistical significance tests, at least for the experiments that support the
1139	main claims of the paper.
1140	• The factors of variability that the error bars are capturing should be clearly stated (for
1141	example, train/test split, initialization, random drawing of some parameter, or overall
1142	run with given experimental conditions).
1143	• The method for calculating the error bars should be explained (closed form formula,
1144	call to a library function, bootstrap, etc.)
1145	• The assumptions made should be given (e.g., Normally distributed errors).
1146	• It should be clear whether the error bar is the standard deviation or the standard error of
1147	the mean.
1148	• It is OK to report 1-sigma error bars, but one should state it. The authors should
1149	preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis
1150	of Normality of errors is not verified.
1151	• For asymmetric distributions, the authors should be careful not to show in tables or
1152	figures symmetric error bars that would yield results that are out of range (e.g. negative
1153	error rates).
1154	• If error bars are reported in tables or plots, The authors should explain in the text how
1155	they were calculated and reference the corresponding figures or tables in the text.
1156	<b>8. Experiments compute resources</b>
1157	Question: For each experiment, does the paper provide sufficient information on the computer
1158	resources (type of compute workers, memory, time of execution) needed to reproduce the
1159	experiments?
1160	Answer: [No]
1161	Justification: Our experiments are small-scale and implementable by a small laptop. Also,
1162	we do not pursue the computational cost in this study, so the computational resource is out
1163	of our focus.
1164	Guidelines:
1165	• The answer NA means that the paper does not include experiments.
1166	• The paper should indicate the type of compute workers CPU or GPU, internal cluster,
1167	or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

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

Answer: [Yes]

Justification: We have checked the code.

Guidelines:

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

## 10. Broader impacts

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

Answer: [No]

Justification: A main focus of this study is fundamental, so there is almost no effect on social impacts.

Guidelines:

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

## 11. Safeguards

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

Answer: [No]

Justification: Since this paper is theoretical, the outcome does not have a high risk for misuse.

Guidelines:



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

## 12. Licenses for existing assets

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

Answer: [NA]

Justification: We have not used any existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

## 13. New assets

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

Answer: [No]

Justification: We have not created a new asset throughout this study.

Guidelines:

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

## 14. Crowdsourcing and research with human subjects

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

Answer: [No]

1271 Justification: We have not performed the crowdsourcing experiments and others.

1272 Guidelines:

- 1273 • The answer NA means that the paper does not involve crowdsourcing nor research with
- 1274 human subjects.
- 1275 • Including this information in the supplemental material is fine, but if the main
- 1276 contribution of the paper involves human subjects, then as much detail as possible
- 1277 should be included in the main paper.
- 1278 • According to the NeurIPS Code of Ethics, workers involved in data collection, curation,
- 1279 or other labor should be paid at least the minimum wage in the country of the data
- 1280 collector.

1281 **15. Institutional review board (IRB) approvals or equivalent for research with human**

1282 **subjects**

1283 Question: Does the paper describe potential risks incurred by study participants, whether

1284 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)

1285 approvals (or an equivalent approval/review based on the requirements of your country or

1286 institution) were obtained?

1287 Answer: [No]

1288 Justification: Since this study is fundamental, there is no potential risk on this point.

1289 Guidelines:

- 1290 • The answer NA means that the paper does not involve crowdsourcing nor research with
- 1291 human subjects.
- 1292 • Depending on the country in which research is conducted, IRB approval (or equivalent)
- 1293 may be required for any human subjects research. If you obtained IRB approval, you
- 1294 should clearly state this in the paper.
- 1295 • We recognize that the procedures for this may vary significantly between institutions
- 1296 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
- 1297 guidelines for their institution.
- 1298 • For initial submissions, do not include any information that would break anonymity (if
- 1299 applicable), such as the institution conducting the review.

1300 **16. Declaration of LLM usage**

1301 Question: Does the paper describe the usage of LLMs if it is an important, original, or

1302 non-standard component of the core methods in this research? Note that if the LLM is used

1303 only for writing, editing, or formatting purposes and does not impact the core methodology,

1304 scientific rigorousness, or originality of the research, declaration is not required.

1305 Answer: [No]

1306 Justification: We have used LLM only for the formatting purposes.

1307 Guidelines:

- 1308 • The answer NA means that the core method development in this research does not
- 1309 involve LLMs as any important, original, or non-standard components.
- 1310 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
- 1311 for what should or should not be described.