LEARNING TO RECALL WITH TRANSFORMERS BEYOND ORTHOGONAL EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern large language models (LLMs) excel at tasks that require storing and retrieving knowledge, such as factual recall and question answering. Transformers are central to this capability because they can encode information during training and retrieve it at inference. Existing theoretical analyses typically study transformers under idealized assumptions such as infinite data or orthogonal embeddings. In realistic settings, however, models are trained on finite datasets with non-orthogonal (random) embeddings. We address this gap by analyzing a single-layer transformer with random embeddings trained with (empirical) gradient descent on a simple token-retrieval task, where the model must identify an informative token within a length-L sequence and learn a one-to-one mapping from tokens to labels. Our analysis tracks the "early phase" of gradient descent and yields explicit formulas for the model's storage capacity—revealing a multiplicative dependence between sample size N, embedding dimension d, and sequence length L. We validate these scalings numerically and further complement them with a lower bound for the underlying statistical problem, demonstrating that this multiplicative scaling is intrinsic under non-orthogonal embeddings.

1 INTRODUCTION

Large language models (LLMs) routinely answer knowledge questions with little or no external context, indicating that substantial factual information is stored in parameters and can be retrieved by suitable prompts (Petroni et al., 2019; Jiang et al., 2020; Roberts et al., 2020). A sharper theoretical account of how such parametric memories are learned and accessed is increasingly important: it can guide scaling choices (e.g., trading off memory capacity against compute budgets, Carlini et al., 2022; Allen-Zhu & Li, 2024) and illuminate failure modes (e.g., hallucination, Zucchet et al., 2025; Huang et al., 2025). Motivated by empirical results documenting the prevalence of parametric factual recall and its scaling with model size (Allen-Zhu & Li, 2024; Morris et al., 2025), recent theoretical works have begun to analyze the capacity and learning dynamics of transformers on controlled factual-recall tasks (Cabannes et al., 2024a; Nichani et al., 2025).

Many theoretical studies of transformer optimization work in population-dynamics settings and adopt simplifying assumptions such as treating token embeddings as orthogonal or one-hot vectors (see, e.g., Tian et al. 2023b; Chen et al. 2024; Ghosal et al. 2024). These choices do not always reflect practical applications, but make the math—particularly gradient calculations—more manageable. Furthermore, such population analyses do not characterize the statistical and computational complexity of gradient-based learning. Moreover, in factual-recall setups, it is known that strictly orthogonal embeddings are not capacity-optimal, whereas random/non-orthogonal embeddings (i.e., *superposition*) enable near-optimal factual storage (Nichani et al., 2025). At the same time, abandoning the orthogonality assumption introduces token interference that leads to intricate optimization behavior (e.g., oscillatory trajectories Cabannes et al., 2024b); in practice, superposition-based, memory-efficient solutions can also be more challenging to train (Elhage et al., 2022), highlighting a fundamental trade-off between optimization/statistical efficiency and optimal storage capacity.

Motivated by the above gaps, we aim to address the following question.

Can we characterize the optimization and sample complexity of a transformer with non-orthogonal embeddings trained by gradient descent in the learning of a factual recall task?

1.1 Our Contributions

In this paper, we analyze gradient-based learning of a single-layer transformer with an attention+MLP block and random embeddings on a synthetic task inspired by Nichani et al. (2025): the model must retrieve an informative token from a context containing many noisy tokens via attention, then map it to the correct label via factual recall. To mitigate the complex optimization dynamics arising from non-orthogonal embeddings, we follow Bietti et al. (2023); Oymak et al. (2023) and consider a simplified training regime involving only a few gradient steps with finite samples on the attention and value matrices. This perspective effectively zooms in on the "early phase" dynamics of gradient descent, a common focus in the feature-learning literature (Ba et al., 2022; Damian et al., 2022; Dandi et al., 2023; Wang et al., 2025).

Our analysis provides a fine-grained characterization of how vocabulary size V, sample size N, embedding dimension d, sequence length L, and MLP width m interact to permit successful gradient-based learning of the recall mechanism. Our main result states that

- The success of learning depends on (V, N, d, L, m) in a *multiplicative* manner: learning becomes easier as (N, d, m) increase — reflecting benefits from more data, higher-dimensional (hence more orthogonal) embeddings, and larger MLP width — whereas learning becomes harder as (V, L) increase, i.e., the task is more difficult with a larger vocabulary or longer sequences. This multiplicative relation is visualized in Figure 1, where we examine how the parameter size $m \times d$ depend on the vocabulary size V at different sequence lengths L.
- Consequently, while optimal capacity and sample complexity can be achieved jointly for short sequences, successful learning on long sequences requires either larger embedding dimension (thus sacrificing capacity) or larger sample sizes (worse statistical complexity).

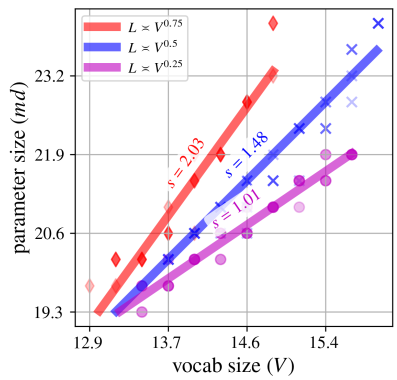


Figure 1: Empirical scaling of parameter count required for GD-trained one-layer transformer to learn factual recall. While the trained model achieves optimal capacity $V \asymp md$ for small L, increasing the sequence length L alters the scaling, suggesting a multiplicative rate.

The multiplicative rate above formalizes the "tradeoff" intuition that smaller embedding dimension d — which increases superposition and thereby improves storage capacity — simultaneously yields a harder learning problem, as reflected in the required sample size. We complement this with a statistical lower bound showing that the trade-off is inherent for any estimator that accesses only gradient information from the initialized transformer. Finally, although our theory is derived for a specific three-step training algorithm, we empirically observe qualitatively similar multiplicative scaling when the transformer is optimized by gradient descent to low empirical risk.

1.2 Related Work

Learning dynamics of transformers. A growing line of theory analyzes how transformers acquire specific behaviors from gradient-based training. Much of this literature imposes population-level assumptions and orthogonal/one-hot embeddings to make gradients tractable, often on discrete synthetic tasks (Li et al., 2023; Bietti et al., 2023; Tian et al., 2023a; Nichani et al., 2024; Chen et al., 2024; Ghosal et al., 2024; Chen et al., 2025; Wang et al., 2025). Several works study few-step training regimes as a lens on the "early phase" of feature learning (Bietti et al., 2023; Wang et al., 2025). Beyond discrete settings, related analyses investigate attention learning for continuous inputs and sparse-signal retrieval (Oymak et al., 2023; Marion et al., 2025). A complementary thread focuses on the emergence of in-context learning and induction mechanisms: single- and two-layer attention trained on linear-regression or Markov data provably implements gradient-descent-like updates and generalized induction heads (Von Oswald et al., 2023; Zhang et al., 2024; Chen et al., 2024; Nichani et al., 2024). These results typically rely on simplified settings and do not address storage capacity. In contrast, our work analyzes finite-sample training with random (non-orthogonal) embeddings in an attention+MLP architecture with a particular focus on factual recall.

Associative memories and storage capacity. Classical associative memories (Hopfield-type models) study recall of vector patterns and established foundational capacity results (Hopfield, 1982; Amit et al., 1985; McEliece et al., 1988; Krotov & Hopfield, 2016; Demircigil et al., 2017; Ramsauer et al., 2020; Schlag et al., 2021). Recent works adapt associative-memory viewpoints to transformers, modeling inner weights as superpositions of outer products and deriving scaling laws and optimization behaviors (Bietti et al., 2023; Cabannes et al., 2024a;b). In factual recall specifically, random (non-orthogonal) embeddings enable near-parameter-count storage, whereas strictly orthogonal embeddings are not capacity-optimal (Nichani et al., 2025). Various empirical works have studied the mechanisms and scaling behaviors of LLMs in factual association tasks (Petroni et al., 2019; Jiang et al., 2020; Geva et al., 2020; Allen-Zhu & Li, 2024). We provide a theoretical analysis of such mechanisms and quantify how vocabulary size, sequence length, embedding dimension, and MLP width jointly govern learning efficiency. Our work operates in a setting similar to (Nichani et al., 2025) but allows finite samples and explicitly considers gradient descent dynamics. Our result is similar to the finite-sample results in (Oymak et al., 2023), where the required sample size grows with the dimensionality and sparsity level of informative tokens, while we allow non-orthogonal embeddings and show optimal capacity as in (Nichani et al., 2025) under certain conditions.

2 Problem Setting

Our goal is to understand the capacity of transformers trained on finite data with non-orthogonal embeddings, in a setting where the relevant information is hidden in a potentially large sequence of non-informative noisy tokens. The attention operation should then identify the relevant token, while the subsequent linear or MLP block can then recall the correct label via an associative memory mechanism. This is similar to the factual recall task studied by Nichani et al. (2025), with simplifications that make the analysis more tractable, as detailed below.

Notation. σ denotes the softmax function. $\mathbb{1}_V \coloneqq (1, \dots, 1)^\top \in \mathbb{R}^V$ is the V-dimensional all-ones vector; $\boldsymbol{e}_i$ is the one-hot vector with a 1 in the i-th position (dimension understood from context). We use $\gtrsim$ (resp. $\lesssim$) to mean "$\geq$" (resp. "$\leq$") up to polylogarithmic factors in V: $f_V \gtrsim g_V \iff f_V \geq \mathrm{poly}(\log V) g_V$ and $f_V \lesssim g_V \iff f_V \leq \mathrm{poly}(\log V) g_V$, for some fixed polynomial. Lastly, $\|\cdot\|_2$ denotes the Euclidean norm for vectors and the operator (spectral) norm for matrices.

Problem setup. Let the input/output tokens take values from a finite alphabet $[V] \coloneqq \{1, \cdots, V\}$. For notational convenience, we represent the alphabet by the one-hot vocabulary $\mathcal{V} = \{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_V\}$. Each example in the data consists of a length-L input sequence $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_L] \in \mathcal{V}^L$ and a label $\boldsymbol{p} \in \mathcal{V}$ generated as follows:

- *Input* tokens are sampled independently and uniformly: $[\boldsymbol{x}_1, \dots, \boldsymbol{x}_L] \sim \mathrm{Unif}(\mathcal{V}^L)$.
- *Informative position* is a random index $\ell \sim \mathrm{Unif}([L])$ independent of $\boldsymbol{X}$.
- *Ground-truth function* is a permutation matrix $\boldsymbol{\Pi}_* \in \{0,1\}^{V \times V}$. Labels are generated as the permuted informative token, $\boldsymbol{p} = \boldsymbol{\Pi}_* \boldsymbol{x}_\ell$, while the remaining tokens are non-informative.

The goal is to identify the correct token position ℓ and learn the target function (permutation) $\boldsymbol{\Pi}_*$.

Transformer architecture. We consider a basic transformer block which first maps input tokens into a d-dimensional embedding space where $d < V$. The embedding layer is parameterized by $(\boldsymbol{Z}_{\text{in}}, \boldsymbol{Z}_{\text{out}}, \boldsymbol{z}_{\text{trig}}, \boldsymbol{z}_{\text{EOS}}) \in \mathbb{R}^{d \times V} \times \mathbb{R}^{d \times V} \times \mathbb{R}^d \times \mathbb{R}^d$, where

- The input tokens are embedded by the columns of the matrix $\boldsymbol{Z}_{\text{in}} \in \mathbb{R}^{d \times V}$.
- Output tokens are associated with unembedding vectors, which are collected in $\boldsymbol{Z}_{\text{out}} \in \mathbb{R}^{d \times V}$.
- $\boldsymbol{z}_{\text{trig}}$ is a trigger vector that marks the informative token.
- $\boldsymbol{z}_{\text{EOS}}$ is the special embedding vector that marks the end-of-sequence.

Given the embedding parameters, we define the self-attention head, parameterized by the key-query matrix $\boldsymbol{W}_{\text{KQ}} \in \mathbb{R}^{d \times d}$, which operates on the embedded sequence of inputs $\boldsymbol{Z}_{\text{in}} \boldsymbol{X} \in \mathbb{R}^{d \times L}$:

$$\mathrm{attn}(\boldsymbol{X}; \boldsymbol{W}_{\text{KQ}}) \coloneqq \boldsymbol{Z}_{\text{in}} \boldsymbol{X} \sigma\Big((\boldsymbol{z}_{\text{trig}} \boldsymbol{e}_\ell^\top + \boldsymbol{Z}_{\text{in}} \boldsymbol{X})^\top \boldsymbol{W}_{\text{KQ}} \boldsymbol{z}_{\text{EOS}} \Big). \tag{1}$$

The trigger embedding z_{trig} is used to "mark" the informative token with a special direction, mimicking the behavior of previous transformer layers that may learn to flag particular tokens by adding to its residual stream[1] (note that the number of trainable parameters inside softmax can be reduced to d by collapsing $W_{\text{KQ}} z_{\text{EOS}}$ into a vector). We consider two different learning models: an *Attention-only* model and a width-m, two-layer neural network model *Attention-MLP*, defined as:

$$\hat{\boldsymbol{p}}(\boldsymbol{X}; \boldsymbol{V}, \boldsymbol{W}_{\text{KQ}}) = \begin{cases} \sigma\Big(\boldsymbol{Z}_{\text{out}}^\top \boldsymbol{V} \operatorname{attn}(\boldsymbol{X}; \boldsymbol{W}_{\text{KQ}})\Big), & \text{Attention only} \\ \sigma\Big(\boldsymbol{Z}_{\text{out}}^\top \boldsymbol{V} \phi(\boldsymbol{W}_{\text{in}} \operatorname{attn}(\boldsymbol{X}; \boldsymbol{W}_{\text{KQ}})\Big), & \text{Attention-MLP} \end{cases} \tag{2}$$

where $\boldsymbol{V} \in \mathbb{R}^{d\times d}$ for the *Attention-only* and $\boldsymbol{V} \in \mathbb{R}^{d\times m}$, $\boldsymbol{W}_{\text{in}} \in \mathbb{R}^{m\times d}$ for the *Attention-MLP* model. Note that compared with *Attention-only* model, the *Attention-MLP* model contains an additional set of trainable parameters and nonlinear activation function ϕ before the value matrix. Constructions of the two models for a related factual recall task can be found in (Nichani et al., 2025, Figure 3).

For the *Attention-MLP*, we keep $\boldsymbol{W}_{\text{in}}$ fixed at its random initialization. The trainable parameters for both of our models are $(\boldsymbol{V}, \boldsymbol{W}_{\text{KQ}})$. We use cross-entropy loss to train our model:

$$\mathcal{L}\big((\boldsymbol{V}, \boldsymbol{W}_{\text{KQ}}), (\boldsymbol{X}, \boldsymbol{p})\big) = -\textstyle\sum_{i=1}^{V} p_i \log \hat{p}_i.$$

Training algorithm. Following Oymak et al. (2023), we consider a 3-step gradient-based algorithm with dataset $\{(\boldsymbol{X}_i, \boldsymbol{p}_i)\}_{i=1}^N$ with a sample size of N. We initialize our parameters as $\boldsymbol{V}^{(0)} = 0$, $\boldsymbol{W}_{\text{KQ}}^{(0)} = 0$ and use the learning rates $\eta, \gamma > 0$:

$$\boldsymbol{V}^{(1)} = \boldsymbol{V}^{(0)} - \eta \cdot \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \nabla_{\boldsymbol{V}} \mathcal{L}\big((\boldsymbol{V}^{(0)}, \boldsymbol{W}_{\text{KQ}}^{(0)}); (\boldsymbol{X}_i, \boldsymbol{p}_i)\big) \tag{3}$$

$$\boldsymbol{W}_{\text{KQ}}^{(1)} = \boldsymbol{W}_{\text{KQ}}^{(0)} - \gamma \cdot \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \nabla_{\boldsymbol{W}_{\text{KQ}}} \mathcal{L}\big((\boldsymbol{V}^{(1)}, \boldsymbol{W}_{\text{KQ}}^{(0)}); (\boldsymbol{X}_i, \boldsymbol{p}_i)\big) \tag{4}$$

$$\boldsymbol{V}^{(2)} = \boldsymbol{V}^{(1)} - \gamma \cdot \tfrac{1}{N} \textstyle\sum_{i=1}^{N} \nabla_{\boldsymbol{V}} \mathcal{L}\big((\boldsymbol{V}^{(1)}, \boldsymbol{W}_{\text{KQ}}^{(1)}); (\boldsymbol{X}_i, \boldsymbol{p}_i)\big). \tag{5}$$

Network prediction and storage. Given our model and training method, we use argmax decoding at inference and define the test accuracy as

$$\text{Accuracy} := \mathbb{P}_{(\boldsymbol{X},\boldsymbol{p})}\big[\boldsymbol{p} = \boldsymbol{e}_{\text{pred}(\boldsymbol{X})}\big], \quad \text{where} \quad \text{pred}(\boldsymbol{X}) := \arg\max_{j\in[V]} \hat{p}_j(\boldsymbol{X}; \boldsymbol{V}^{(2)}, \boldsymbol{W}_{\text{KQ}}^{(1)}),$$

where $\hat{\boldsymbol{p}}(\boldsymbol{X}; \boldsymbol{V}^{(2)}, \boldsymbol{W}_{\text{KQ}}^{(1)})$ is the network output defined in (2). In what follows, we characterize conditions under which the model stores the informative tokens asymptotically, i.e., $\text{Accuracy} \to 1$ as $V \to \infty$, in terms of the relevant parameters (V, N, d, L, m).

3 Main Results

We first present our general theorem on learnability via gradient descent, and then specialize into different regimes to derive more interpretable scaling behaviors in Section 4. We provide a proof sketch in Section C.1, and defer the full proof to Appendix C.

3.1 Technical Assumptions

We first state generic assumptions that apply to both the *Attention-only* and *Attention-MLP* models.

Assumption 1.

- ***Parameter range:*** *Let $L = V^c$ for $c \in (0, 1)$, $\Omega(V \log V) \le N = o(VL)$, and $V \ge \Omega(1)$.*
- ***Learning rate:*** *We use a sufficiently small learning rate $\eta = o(1)$ for the initial step (3), and sufficiently large learning rate $\gamma = \omega(1)$ for the remaining steps (4)-(5) that satisfy Assumption 4.*
- ***Embeddings:*** *Let $\boldsymbol{Z}_{\text{in}}, \boldsymbol{Z}_{\text{out}} \in \mathbb{R}^{d\times V}$ be independent Gaussian matrices, and let $\boldsymbol{z}_{\text{trig}}, \boldsymbol{z}_{\text{EOS}} \in \mathbb{R}^d$ be independent Gaussian vectors, all with i.i.d. entries distributed as $\mathcal{N}(0, 1/d)$.*

[1] The "trigger" terminology is borrowed from (Bietti et al., 2023), where a special previous token "triggers" a retrieval operation in the context of induction heads. Our setup resembles learning only the "induction head" layer assuming the first "previous token head" layer is already in place.

We assume $c \in (0, 1)$ since in many practical pretraining setups, the context length is smaller than the vocabulary size, and the condition $L \ll V$ simplifies several terms in the proofs. The lower bound $N \gtrsim V \log V$ is required so that each element from the alphabet of size V is seen at least once with high probability. The learning rates follow prior analyses (Oymak et al., 2023; Nichani et al., 2024): a small η ensures that the network's predictions remain close to uniform after the first step, whereas a large γ is needed to push the attention scores and predictions toward one-hot vectors.

In addition to the above assumptions, we require the transformer model to have sufficient capacity to reach perfect test accuracy. Such conditions are characterized by Nichani et al. (2025). For the *Attention-only* model, we have the following condition (see Nichani et al., 2025, Theorem 3).

Assumption 2 (Attention-only)**.** *For the* Attention-only *model, we require* $d \gtrsim \sqrt{V}$.

With a nonlinear MLP layer, a smaller embedding dimension can suffice if the width is large enough. Hence for *Attention-MLP* we require the following condition.

Assumption 3 (Attention-MLP)**.** *For the* Attention-MLP *model, we assume that*

- ***Polynomial activation:*** $\phi : \mathbb{R} \to \mathbb{R}$ *satisfies* $\phi(0), \phi'(0), \phi''(0) \neq 0$.
- ***MLP width:*** $md \gtrsim V$ *and* $d \gtrsim V^{\frac{1}{k_\star+1}}$, *where* $k_\star$ *denotes the smallest nonzero Hermite mode of* ϕ, *i.e.,* $k_\star := \min\{k > 0 : \mathbb{E}_{Z\sim\mathcal{N}(0,1)}[\phi(Z)h_k(Z)] \neq 0\}$ *where* h_k *is the* k^{th} *Hermite polynomial.*
- ***Initialization:*** $\boldsymbol{W}_{\text{in}} \in \mathbb{R}^{m\times d}$ *are fixed with entries i.i.d. distributed as* $\mathcal{N}(0, 1)$.

The nonlinear MLP layer allows us to compensate for the embedding dimension and go beyond the $d \gtrsim \sqrt{V}$ lower bound required by the *Attention-only* model (Assumption 2). Note that $md \gtrsim V$ is a necessary condition for capacity as shown in (Nichani et al., 2025). The additional requirements imposed on the polynomial activation function appear to be artifacts of our three-step GD analysis, and we anticipate that they could be relaxed when considering a longer training horizon.

3.2 Learnability Statement

Now we are ready to present our main theorem on the complexity of learning the factual recall task. Specifically, transformer learns the desired mechanism when the signal term dominates the noise and bias terms as stated below.

Theorem 1. *Let Assumptions 1 and 3 hold for* Attention-MLP*, and 1 and 2 hold for* Attention-only. *The* Attention-MLP *model achieves* $\text{Accuracy} = 1 - o_V(1)$ *with probability* $1 - o_V(1)$ *whenever*

$$\underbrace{\frac{1}{VL^2}}_{\text{Signal}} \gtrsim \underbrace{\frac{1}{N\sqrt{L}d(d\wedge L)}}_{\text{Gradient noise}} + \underbrace{\frac{1}{N\sqrt{V}d(d\wedge L)}}_{\text{Mean bias}} + \underbrace{\frac{1}{Nd\sqrt{m}}}_{\text{MLP noise}}. \tag{6}$$

For the Attention-only *model, the same holds with the last MLP noise term removed.*

Theorem 1 characterizes learnability as a function of (V, N, d, L, m) and identifies the following terms that impact the gradient signal-to-noise ratio:

1. *Signal* measures the alignment between the key–query weights $\boldsymbol{W}_{\text{KQ}}^{(1)}$ and the trigger $\boldsymbol{z}_{\text{trig}}$.
2. *Gradient noise* is due to the concentration error in the update of $\boldsymbol{W}_{\text{KQ}}^{(1)}$.
3. *Mean bias* arises from the nonzero mean of token vectors $\{\boldsymbol{X}_i\}_{i=1}^N$.
4. *MLP noise* reflects the randomness in the MLP weight matrix $\boldsymbol{W}_{\text{in}}$ in *Attention–MLP*.

We make the following observations.

- **Multiplicative scaling.** Note that the parameters (V, N, d, L, m) interact in a multiplicative fashion. For example, the noise and bias terms in (6) all decay with $(N \times d)$, suggesting that increasing the embedding dimensions d can lower the statistical complexity of learning the correct recall mechanism. While the full 5-parameter trade-off can be opaque, in Section 4 we focus on specific regimes that lead to simplification of the scaling relationship and validate the rate empirically.

- **Optimal storage & sample complexity.** Recall that the capacity-optimal construction for the factual recall task requires $md \gtrsim V$ parameters (or $d^2 \gtrsim V$ for *Attention–only*); and as discussed earlier, a sample size $N \asymp V \log V$ is necessary to observe all distinct tokens. (6) implies that in the small-L regime, the optimized transformer achieve optimal capacity and sample complexity simultaneously. For longer sequences, however, these two conditions may not be achieved at the same time, i.e., one must increase either the network width or sample size beyond optimality to learn the task — this confirms the empirical observation in Figure 1.

3.3 Statistical Lower Bound

Theorem 1 provides an upper bound (i.e., sufficient condition) on the model and sample size for learning factual recall under a 3-gradient-step optimization procedure. We complement this sufficient condition with a lower bound indicating that the multiplicative dependence on the problem parameters is partly statistical; that is, the scaling behavior will be observed in any model satisfying the broader conditions stated below. Our lower bound applies to statistical methods that can query the dataset through the attention outputs at initialization, $\boldsymbol{h}_i := \mathrm{attn}(\boldsymbol{X}_i, \boldsymbol{W}_{\mathrm{KQ}}^{(0)})$. In particular, we consider queries of the form $\{\boldsymbol{h}_i, \boldsymbol{h}_i\boldsymbol{h}_i^\top\}_{i=1}^N$ as the gradient with respect to the key–query matrix $\boldsymbol{W}_{\mathrm{KQ}}$ depends on these quantities (see (10)). The statement is given below:

Theorem 2 (Informal)**.** *Any method that relies on the noisy version of the queries* $\{\boldsymbol{h}_i,\ \boldsymbol{h}_i\boldsymbol{h}_i^\top\}_{i=1}^N$ *fails, i.e.,* $\mathrm{Accuracy} \not\to 1$ *with finite probability, if* $N \lesssim V\min\{1, L/d^2\}$.

The complete statement of Theorem 2 is deferred to Theorem 4 in Appendix D. We observe that the lower bound does not exactly match our upper bound in Theorem 1, as *Signal* $\lesssim$ *Gradient Noise* in (6) is stronger than the stated lower bound. This being said, Theorem 2 also confirms the multiplicative scaling, hence suggesting the trade-off between capacity and sample efficiency is present in a boarder class of learning algorithms. A stronger computational lower bound for transformers and gradient-based optimization is an interesting problem we leave for future work.

4 Implications and Empirical Verifications

In this section, we leverage our main theorem to obtain more concrete scalings between parameters, and present empirical evidence on the derived multiplicative rate.

4.1 Attention-only Model

We start with the *Attention-only model* which gives a simpler phase diagram.

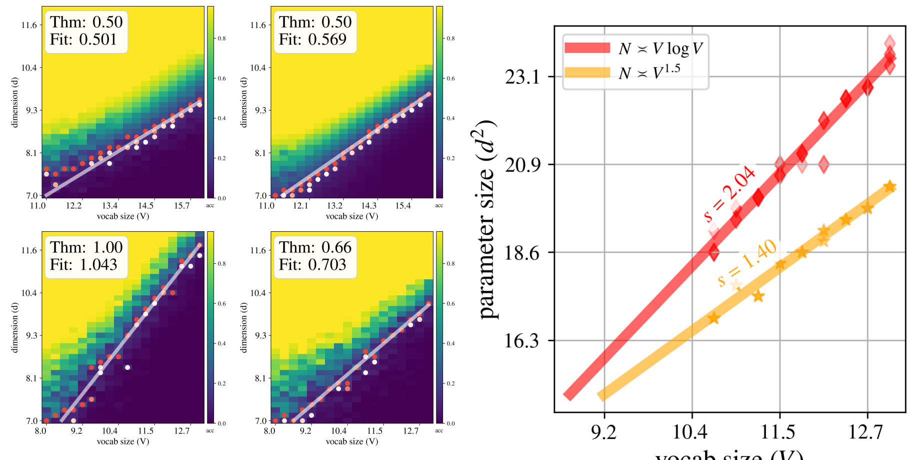
Corollary 1. *For the* Attention-only *model, the bottleneck term in* (6) *is the* Mean bias *term, and Theorem* (1) *is equivalent to requiring* $d \gtrsim \max\{\sqrt{V}, V^{\frac{1}{3}}L^{\frac{4}{3}}/N^{\frac{2}{3}}\}$.

We make the following observations:

- The condition in Corollary 1 is the maximum of two terms, where $d \gtrsim \sqrt{V}$ is due to the capacity requirement in Assumption 2, whereas the second term ensures *Signal* $\gtrsim$ *Mean bias* and implies a multiplicative scaling between the sample size N and embedding dimension d (i.e., increasing one of the parameters can compensate for the other).
- Note that the *Mean bias* term arises from a nonzero token mean, which can potentially be alleviated by centering the tokens, as is effectively done by normalization layer. Exploring the effect of applying normalization in this model is an interesting direction for future work.

Empirical Findings. We run the three-step gradient descent algorithm on an *Attention-only* model over varying V and d, and report the accuracies in the heatmaps (Figure 2). The plots are in log-log scale; therefore, the slopes give the exponent s in $d \asymp V^s$. As shown in the top row of Figures 2a-2b, the slope for relatively small L (where $L \asymp \sqrt{V}$) matches the optimal capacity condition $d \asymp \sqrt{V}$. By contrast, when the context window is larger ($L \asymp V$), the requirement becomes $d \asymp V$, which is also reflected in the experimental results, as observed in the bottom panel of Figure 2a.

In Figure 2b we run experiments with increasing sample size to observe the multiplicative trade-off. As seen in the bottom figure of Figure 2b, increasing the sample size from $V \log V$ to $V^{1.5}$ reduces the dimension exponent from 1 to 0.7 (the theoretical value is $s = 0.66$). Finally, the learnability thresholds for $L \asymp V$ in Figures 2a and 2b are plotted together in Figure 2c, to illustrate that increasing the sample size can compensate for the number of parameters in the network.



(a) $N \asymp V \log V$ (b) $N \asymp V^{1.5}$ (c) Attention only, $L \asymp V$

Figure 2: Empirical scaling of embedding dimension (left) and parameter count (right) via three-step GD for the *Attention-only* model. In (a) and (b), top-left and top-right use $L \asymp \sqrt{V}$; bottom-left and bottom-right use $L \asymp V$. In the right panel, the $L \asymp V$ case is shown under two sample-size regimes, $N \asymp V \log V$ and $N \asymp V^{1.5}$. *Line fitting:* We identify in the heatmaps the smallest embedding dimension that achieves accuracies $\{0.1, 0.125, 0.15\}$ and perform a least squares fit. The slopes of the fitted lines and their theoretical counterparts are reported on the heatmaps. Differences in transparency in (c) are due to overlapping points.

4.2 Attention-MLP Model

For the attention-MLP model, the nonlinear MLP layer introduces additional phases as stated below.

Corollary 2. *For the* Attention-MLP *model, Theorem 1 translates to* $md \gtrsim V$ *and*

$$\text{Signal} \gtrsim \begin{cases} \text{MLP noise}, & m = o(d^2 L) \text{ and } m = o(dV) \\ \text{Gradient noise 2}, & V \gtrsim dL \text{ and } m \gtrsim d^2 L \\ \text{Mean Bias}, & V = o(dL) \text{ and } m \gtrsim dV, \end{cases}$$

where

- Signal $\gtrsim$ MLP noise *is equivalent to* $Nd \gtrsim VL^2/\sqrt{m}$.
- Signal $\gtrsim$ Gradient noise 2 *is equivalent to* $d\sqrt{N} \gtrsim VL^{\frac{1}{4}}$
- Signal $\gtrsim$ Mean Bias *is equivalent to* $dN^{\frac{2}{3}} \gtrsim L^{\frac{4}{3}} V^{\frac{1}{3}}$.

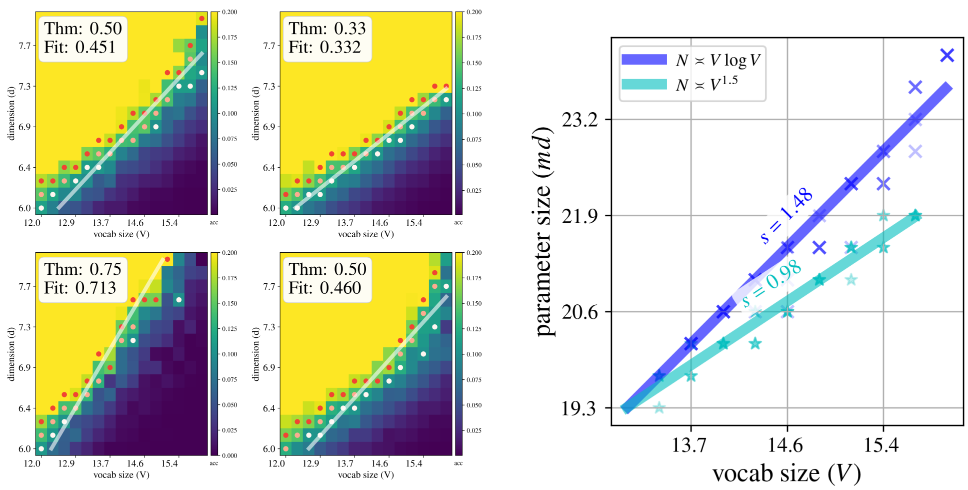
The phase diagram for the Attention-MLP model is richer than *Attention-only*, as we can trade off m and d and hence use a smaller embedding dimension; this results in potentially different dominant terms in the gradient. In particular, since large L and d entails larger magnitude of *Mean Bias* (as in the *Attention-only* setting), we know that by increasing the MLP width m and thereby reducing the required embedding dimension d, we may suppress this bias term.

Empirical Findings. We run the 3-step gradient descent algorithm on an Attention-MLP network over varying V and d and plot the accuracies in Figures 3 and 4. We take the nonlinearity to be the mixture of two Hermite polynomials $\phi = 0.7h_2 + 0.3h_3$, satisfying the conditions in Assumption 3. We run experiments with width $m \asymp d^2$ and $m \asymp d^3$. Due to the prohibitive cost of increasing the width further, we restrict ourselves to the *MLP noise*-dominated region.

In Figure 1, we plot the scaling of the number of parameters (md) as a function of vocabulary size V for different sequence-length regimes in L. We observe that $L \asymp V^{0.25}$ requires $md \asymp V$, which is the optimal capacity, as predicted by our theory. As L increases, we need more parameters to achieve the same capacity, as observed in the $L \asymp V^{0.5}$ and $L \asymp V^{0.75}$ cases in Figure 1, where the slopes agree with our theoretical predictions as well (see Figures 3a and 3b).

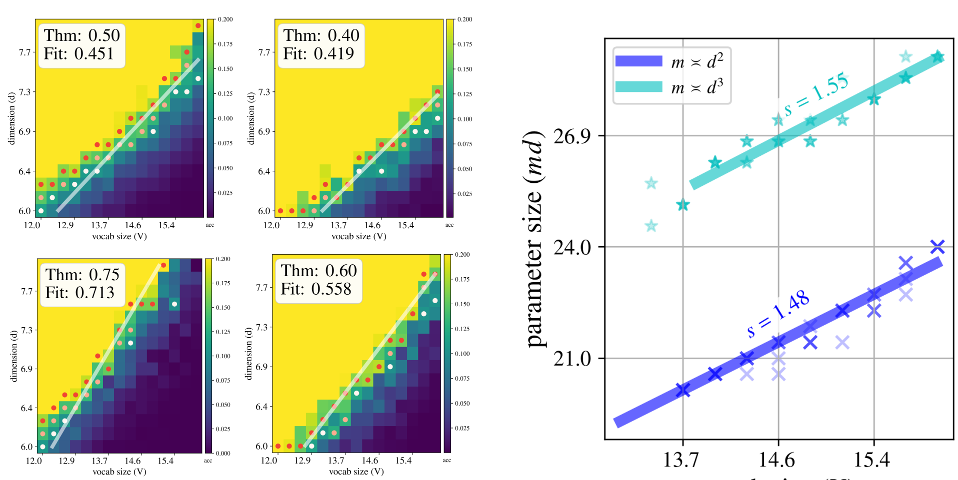
We further test the effect of sample size in Figure 3, where we use $L \asymp V^{0.5}$ and $m \asymp d^2$. We plot both heat maps in Figures 3a and 3b, and the fitted lines for $L \asymp V^{0.5}$ together in Figure 3c. Note that we state the plot in terms of parameter count, which scales as $md \asymp d^3$, so the slopes from the heat map are scaled accordingly. We observe that increasing N from $N \asymp V \log V$ to $N \asymp V^{1.5}$ reduces the network size to the optimal level, aligning with our theoretical prediction. The heatmap versions of these experiments are shown in Figures 3a and 3b.

Lastly, we probe the width scaling by keeping the sample size $N \asymp V \log V$ and $L \asymp V^{0.5}$ fixed in Figure 4. Here, we observe that we can reduce the embedding-dimension requirement by increasing m (Figures 4a and 4b), though it increases the total parameter count overall, as seen in Figure 4c, since width must grow proportionally more than d to achieve the same accuracy. This is also consistent with our theoretical prediction.



(a) $N \asymp V \log V$ (b) $N \asymp V^{1.5}$ (c) $L \asymp V^{0.5}$ and $m \asymp d^2$

Figure 3: Empirical scaling of embedding dimension (left) and parameter count (right) for the *Attention-MLP* model under $N \asymp V \log V$ and $N \asymp V^{1.5}$. In (a) and (b), top-row uses $L \asymp V^{0.5}$; bottom-row uses $L \asymp V^{0.75}$. The right panel also shows $L \asymp V^{0.5}$ under both sample-size regimes.



(a) $m \asymp d^2$ (b) $m \asymp d^3$ (c) $L \asymp V^{0.5}$ and $N \asymp V \log V$

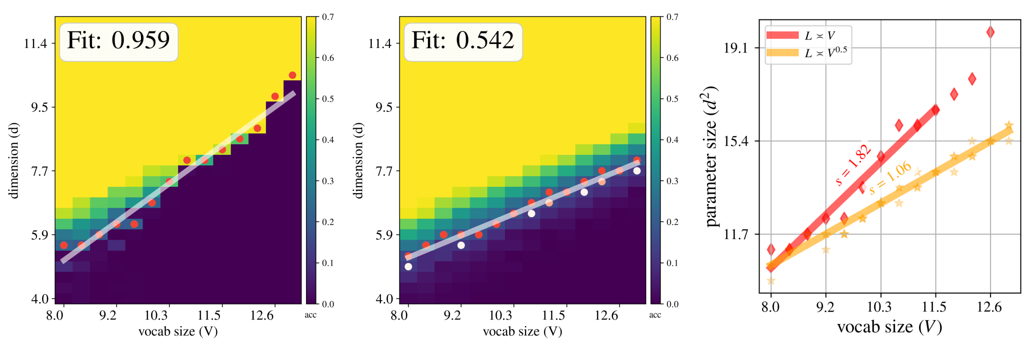
Figure 4: Empirical scaling of embedding dimension (left) and parameter count (right) for the *Attention-MLP* model under two width regimes, $m \asymp d^2$ and $m \asymp d^3$. In (a) and (b), top-row uses $L \asymp V^{0.5}$; bottom-row uses $L \asymp V^{0.75}$. The right panel also shows $L \asymp V^{0.5}$ under both width regimes.

4.3 Beyond Early Phase of Training

While our theoretical analysis handles a particular 3-gradient-step training procedure, we empirically observe qualitatively similar multiplicative scalings when the transformer model is optimized beyond the "early phase". Specifically, we train our *Attention-only* model for multiple steps using (i) full-batch gradient descent and (ii) Adam (Kingma & Ba, 2015) with mini-batch gradients. Throughout this section, we use a sample size $N \asymp V \log V$.

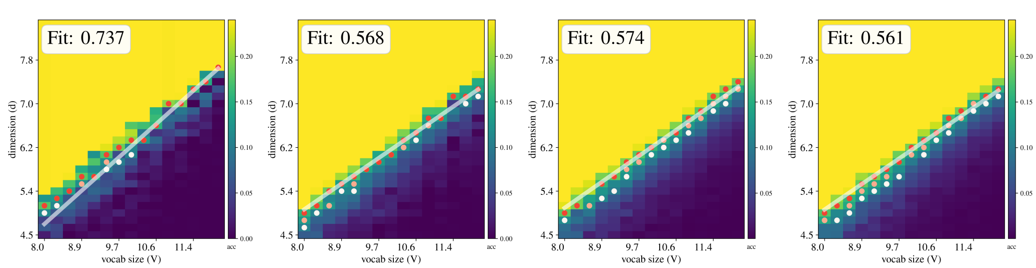
Full-batch gradient descent: We use learning rate $\eta = 0.5$ and continue training until the test accuracy does not improve by more than 0.01 for 10 consecutive checks. In Figures 5a and 5b, we provide heatmaps for $L \asymp V$ and $L \asymp \sqrt{V}$. We observe that when $L \asymp \sqrt{V}$, the slope indicates the network is at the optimal capacity condition; this is also reflected by the slope 1.06 in Figure 5c. By contrast, for large L the slope significantly shifts and becomes suboptimal, confirming the multiplicative relation established in Section 3.

Adam with mini-batch gradients: We use layer normalization in both the attention and output layers and choose learning rate $\eta = 0.005$. We specifically use batch size $\lfloor N/40 \rfloor$ and run the algorithm for 3 epochs. In Figure 6, we provide heatmaps for $L \asymp V$ and $L = 8$ at the end of epochs 2 and 3. We observe that for $L \asymp V$ in early training, the slope is suboptimal, while training the network for one more epoch improves the capacity condition to a near optimal level. In contrast, for $L = 8$, we observe that the network does not exhibit a suboptimal phase at the end of epoch 2, which is in line with our theoretical findings. A rigorous analysis of the full gradient descent dynamics is left for future work.



(a) $L \asymp V$ (b) $L \asymp \sqrt{V}$ (c) Comparison of the fitted lines

Figure 5: Empirical scaling of embedding dimension (a,b) and parameter count (c) for the *Attention-only* model trained by multiple-step GD.



(a) $L \asymp V$, Epoch $= 2$ (b) $L \asymp V$, Epoch $= 3$ (c) $L = 8$, Epoch $= 2$ (d) $L = 8$, Epoch $= 3$

Figure 6: Empirical scaling of embedding dimension for the *Attention-only* model trained by Adam.

5 Proof Overview: Population Analysis

In this section, we outline the ideas for the proof of Theorem 1. For presentation, we consider the *Attention-only* model with population dynamics and orthogonal embeddings. Since we do not use positional encoding in the model, without loss of generality, we fix the correct position to $\ell = 1$.

In the proof, we study the attention scores in (1) and characterize the conditions under which they align with the trigger vector. Once attention can distinguish informative tokens, the remaining part reduces to learning a linearly separable problem, which is well understood. The pre-softmax scores evaluated on a fresh sequence $\boldsymbol{X}_{\text{in}}$, with the key-query matrix given by the first gradient-descent iterate $\boldsymbol{W}_{\text{KQ}}^{(1)}$, is given as

$$\text{scores} := \left(\boldsymbol{z}_{\text{trig}}\boldsymbol{e}_1^\top + \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_{\text{in}}\right)^\top \boldsymbol{W}_{\text{KQ}}^{(1)} \boldsymbol{z}_{\text{EOS}}. \tag{7}$$

For the proof overview part, we analyze the following simplified form of the scores (for the full expression, see (10)):

$$\text{scores} \approx \gamma \boldsymbol{X}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}^\top \underbrace{\left(\frac{1}{NL}\sum_{i=1}^{N} \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i\boldsymbol{X}_i^\top \boldsymbol{Z}_{\text{in}}^\top (\boldsymbol{V}^{(1)})^\top \boldsymbol{Z}_{\text{out}}(\boldsymbol{p}_i - \tfrac{1}{V}\mathbb{1}_V)\right)}_{\text{Non-informative}} \tag{8}$$

$$+\gamma\|\boldsymbol{z}_{\text{trig}}\|_2^2 \boldsymbol{e}_1 \underbrace{\left(\frac{1}{NL}\sum_{i=1}^{N} \boldsymbol{x}_{i,1}^\top \boldsymbol{Z}_{\text{in}}^\top (\boldsymbol{V}^{(1)})^\top \boldsymbol{Z}_{\text{out}}(\boldsymbol{p}_i - \tfrac{1}{V}\mathbb{1}_V)\right)}_{\text{Informative}}. \tag{9}$$

Here $\boldsymbol{V}^{(1)}$ denotes the first iteration of the value matrix given in (3). The informative term in (9) captures the alignment between the trigger vector in the fresh input and the one in the learned weights $\boldsymbol{W}_{\text{KQ}}^{(1)}$, and therefore contains the position information of the informative token. By contrast, the non-informative term in (8) reflects correlations between tokens and does not carry any information about the token's position. The proof characterizes conditions under which the informative term in (9) dominates, which is sufficient for attention to identify the correct position.

To study population dynamics with orthogonal embeddings, we set $\boldsymbol{Z}_{\text{in}} = \boldsymbol{Z}_{\text{out}} = \boldsymbol{I}_V$ and take $N \to \infty$ while other parameters remain fixed. Under population dynamics, we first observe that $\boldsymbol{V}^{(1)} = O(\eta)\boldsymbol{\Pi}_*$, where η is the learning rate of the first step in (3). Then, we can write

$$\begin{aligned}\text{Non-informative} &= \boldsymbol{X}_{\text{in}}^\top \frac{O(\eta\gamma)}{NL}\sum_{i=1}^{N} \boldsymbol{X}_i\boldsymbol{X}_i^\top \boldsymbol{\Pi}_*^\top(\boldsymbol{p}_i - \tfrac{1}{V}\mathbb{1}_V)\\ &\xrightarrow[N\to\infty]{} \frac{O(\eta\gamma)}{L}\boldsymbol{X}_{\text{in}}^\top \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top(\boldsymbol{x}_1 - \tfrac{1}{V}\mathbb{1}_V)] = (1-\frac{1}{V})\frac{O(\eta\gamma)}{VL}\mathbb{1}_L,\end{aligned}$$

where the last equality follows since $\boldsymbol{X}_{\text{in}}$ has one-hot columns. On the other hand, we have

$$\text{Informative} = O(\eta\gamma)\boldsymbol{e}_1\left(\frac{1}{NL}\sum_{i=1}^{N} \boldsymbol{x}_{i,1}^\top \boldsymbol{\Pi}_*^\top(\boldsymbol{p}_i - \tfrac{1}{V}\mathbb{1}_V)\right) = (1-\frac{1}{V})\frac{O(\eta\gamma)}{L}\boldsymbol{e}_1,$$

where we used $\boldsymbol{p}_i = \boldsymbol{\Pi}_*\boldsymbol{x}_{i,1}$, in the last step. By choosing learning rates that guarantee $\eta\gamma \to \infty$, we can show that the attention probabilities align with $\boldsymbol{e}_1$ and eventually select the correct position. The reader may refer to Appendix C.1 for an extended proof overview of the empirical dynamics with non-orthogonal embeddings, where we detail how each term in Theorem 1 arises from the terms in (8)-(9).

6 Conclusion

In this paper, we derived precise asymptotic rates for learning with gradient descent on transformers trained on a simple recall task with random embeddings and finite samples. Our analysis and experiments reveal a rich picture of multiplicative scalings between various problem parameters, showing that parameter count is not the only important factor controlling capacity when learning with finite samples on large noisy sequences. Our results suggest that finer control of the data distribution may be necessary for learning efficiently at optimal capacity, for instance by ensuring sequences are less noisy and more informative, hoping that the discovered mechanisms are robust to harder settings. This is reminiscent of the procedures used for long context extension in LLMs, where most of training happens on shorter sequences, but the final models are extended to work with very long sequences, and empirically do well on retrieval tasks such as "needle-in-a-haystack" (e.g., Gemini Team, 2024), which resembles our theoretical setup. Analyzing similar scalings in more structured data distributions and architectures is thus an interesting avenue for future work.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. *arXiv preprint arXiv:2404.05405*, 2024.

Daniel J Amit, Hanoch Gutfreund, and Haim Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, 55(14):1530, 1985.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Vivien Cabannes, Elvis Dohmatob, and Alberto Bietti. Scaling laws for associative memories. In *International Conference on Learning Representations (ICLR)*, 2024a.

Vivien Cabannes, Berfin Simsek, and Alberto Bietti. Learning associative memories with gradient descent. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024b.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2022.

Lei Chen, Joan Bruna, and Alberto Bietti. Distributional associations vs in-context reasoning: A study of feed-forward and attention layers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Unveiling induction heads: Provable training dynamics and feature learning in transformers. *Advances in Neural Information Processing Systems*, 37:66479–66567, 2024.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pp. 5413–5452. PMLR, 2022.

Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. Learning two-layer neural networks, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*, 2023.

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168:288–299, 2017.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020.

Gaurav Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. Understanding finetuning for factual knowledge extraction. *arXiv preprint arXiv:2406.14785*, 2024.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

Yixiao Huang, Hanlin Zhu, Tianyu Guo, Jiantao Jiao, Somayeh Sojoudi, Michael I Jordan, Stuart Russell, and Song Mei. Generalization or hallucination? understanding out-of-context reasoning in transformers. *arXiv preprint arXiv:2506.10887*, 2025.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR*, 2015.

Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. *Advances in neural information processing systems*, 29, 2016.

Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Pierre Marion, Raphael Berthier, Gerard Biau, and Claire Boyer. Attention layers provably solve single-location regression. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Robert J McEliece, Edward C Posner, Eugene R Rodemich, and Santosh S Venkatesh. The capacity of the hopfield associative memory. *IEEE transactions on Information Theory*, 33(4):461–482, 1988.

John X Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G Edward Suh, Alexander M Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. How much do language models memorize? *arXiv preprint arXiv:2505.24832*, 2025.

Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Eshaan Nichani, Jason D Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Samet Oymak, Ankit Singh Rawat, Mahdi Soltanolkotabi, and Christos Thrampoulidis. On the role of attention in prompt-tuning. In *International Conference on Machine Learning*, 2023.

Yang Peng, Yuchen Xin, and Zhihua Zhang. Matrix rosenthal and concentration inequalities for markov chains with applications in statistical learning. *arXiv preprint arXiv:2508.04327*, 2025.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel. Language models as knowledge bases? *arXiv preprint arXiv:1909.01066*, 2019.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, et al. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.

Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*, 2020.

Jonathan Scarlett and Volkan Cevher. An introductory guide to fano's inequality with applications in statistical estimation. *arXiv preprint arXiv:1901.00555*, 2019.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon S Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Yuandong Tian, Yiping Wang, Zhenyu Zhang, Beidi Chen, and Simon Du. Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention. *arXiv preprint arXiv:2310.00535*, 2023b.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Zixuan Wang, Eshaan Nichani, Alberto Bietti, Alex Damian, Daniel Hsu, Jason D Lee, and Denny Wu. Learning compositional functions with transformers from easy-to-hard data. In *Conference on Learning Theory (COLT)*, 2025.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

Nicolas Zucchet, Jörg Bornschein, Stephanie Chan, Andrew Lampinen, Razvan Pascanu, and Soham De. How do language models learn facts? dynamics, curricula and hallucinations. *arXiv preprint arXiv:2503.21676*, 2025.

A APPENDIX

CONTENTS

LLM Usage. Large language models are used to polish the abstract and find relevant references for the related work section.

B PRELIMINARIES

Proof organization. We combine the proof for both models: Using $\phi(x) = x$ and $m = \infty$ is valid for applying the arguments for *Attention-only* model in the proof. As the network succeeds in storing all informative tokens only when attention selects the correct position, we focus on how attention learns the correct index and under what conditions. This is the bottleneck in our analysis under Assumptions 1, 2, and 3. Accordingly, we study the pre-softmax scores in (1). Theorem 3 characterizes the scaling of these terms and yields (6). Because the proof involves lengthy expressions, we provide a proof sketch in Section C.1 and refer readers to the corresponding parts of the formal proof.

Additional Notation. For a vector $\boldsymbol{x} \in \mathbb{R}^V$, we use $\text{diag}(\boldsymbol{x}) \in \mathbb{R}^{V\times V}$ denotes the diagonal matrix which has the same diagonal entries with $\boldsymbol{x}$, while for a matrix $\boldsymbol{A}$, $\text{diag}(\boldsymbol{A}) \in \mathbb{R}^V$ denotes the column vector whose elements coincide with the diagonal entries of $\boldsymbol{A}$. For a random variable $\boldsymbol{w}$, $\mathbb{E}_{\boldsymbol{w}}[\cdot]$ denotes taking expectation with respect to $\boldsymbol{w}$ and keeping the remaining independent terms fixed. Similarly, we use $\mathbb{E}[\cdot|\boldsymbol{w}]$ for conditional expectation, conditioned on $\boldsymbol{w}$. We use $\mathbb{1}_{\text{Event}}$ as an indicator function, which takes values $\{0, 1\}$ depending on the event holds or not. We use C to denote any constant in the upper-bound, which might depend on ϕ.

Since we do not use positional encoding in the model, without loss of generality we can fix the informative index $\ell = 1$. We define the sequence of non-informative tokens as $\boldsymbol{N}_i := [\boldsymbol{x}_{i,2}, \cdots, \boldsymbol{x}_{i,L}]^\top$. We will denote the rows of $\boldsymbol{W}_{\text{in}}$ with $\{\boldsymbol{w}_k\}_{k=1}^m$. For compact representation the attention with the trigger we define

$$\mathsf{Z}_{\text{in}} =: [\boldsymbol{Z}_{\text{in}} \quad \boldsymbol{z}_{\text{trig}}] \text{ and } \mathsf{X}_i = \begin{bmatrix} \boldsymbol{x}_{i,1}^\top & 1 \\ \boldsymbol{N}_i & 0 \end{bmatrix} \in \mathbb{R}^{L\times(V+1)}$$

With this notation, we can write the iterates in three-step GD. Let

$$\boldsymbol{\alpha}^{(0)} := \sigma\Big((\boldsymbol{z}_{\text{trig}}\boldsymbol{e}_\ell^\top + \boldsymbol{Z}_{\text{in}}\boldsymbol{X})^\top \boldsymbol{W}_{\text{KQ}}^{(0)} \boldsymbol{z}_{\text{EOS}}\Big).$$

We have

$$\boldsymbol{V}^{(1)} = \boldsymbol{Z}_{\text{out}}\Big(\frac{\eta}{N}\sum_{i=1}^N (\boldsymbol{p}_i - \hat{\boldsymbol{p}}_i^{(0)})\phi\big((\boldsymbol{\alpha}_i^{(0)})^\top \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{W}_{\text{in}}^\top\big)\Big)$$

$$\begin{aligned} \boldsymbol{W}_{\text{KQ}}^{(1)} = \mathsf{Z}_{\text{in}}\, \frac{\gamma}{N}\sum_{i=1}^N \mathsf{X}_i^\top \big(\text{diag}(\boldsymbol{\alpha}_i^{(0)}) - \boldsymbol{\alpha}_i^{(0)}(\boldsymbol{\alpha}_i^{(0)})^\top\big)\boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{W}_{\text{in}}^\top \\ \times \text{diag}\Big(\phi'\big(\boldsymbol{W}_{\text{in}}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\boldsymbol{\alpha}_i^{(0)}\big)\Big)(\boldsymbol{V}^{(1)})^\top \boldsymbol{Z}_{\text{out}}(\boldsymbol{p}_i - \hat{\boldsymbol{p}}_i^{(1)})\boldsymbol{z}_{\text{EOS}}^\top \end{aligned} \quad (10)$$

For notational convenience, we define the noise due to finite width as

$$\begin{aligned} \text{FW}(\boldsymbol{W}_{\text{in}}; \boldsymbol{Z}_{\text{in}}, \boldsymbol{X}_i, \boldsymbol{X}_j) := \frac{1}{m}\Big(\boldsymbol{W}_{\text{in}}^\top \text{diag}\Big(\phi'\big(\tfrac{1}{L}\boldsymbol{W}_{\text{in}}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big)\Big)\phi\Big(\tfrac{1}{L}\boldsymbol{W}_{\text{in}}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top \mathbb{1}_L\Big) \\ - \mathbb{E}_{\boldsymbol{W}_{\text{in}}}\Big[\boldsymbol{W}_{\text{in}}^\top \text{diag}\Big(\phi'\big(\tfrac{1}{L}\boldsymbol{W}_{\text{in}}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big)\Big)\phi\Big(\tfrac{1}{L}\boldsymbol{W}_{\text{in}}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top \mathbb{1}_L\Big)\Big]\Big). \end{aligned}$$

and the terms arising in the expected value term as

$$\alpha_{ij} := \mathbb{E}_{\boldsymbol{w}}\Big[\phi'\big(\tfrac{1}{L}\boldsymbol{w}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big)\phi'\big(\tfrac{1}{L}\boldsymbol{w}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top \mathbb{1}_L\big)\Big],$$

$$\beta_{ij} := \mathbb{E}_{\boldsymbol{w}}\Big[\phi''\big(\tfrac{1}{L}\boldsymbol{w}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big)\phi\big(\tfrac{1}{L}\boldsymbol{w}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top \mathbb{1}_L\big)\Big].$$

C Proof of Theorem 1

C.1 Proof sketch for Theorem 1

For convenience, we fix $\boldsymbol{\Pi}_* = \boldsymbol{I}_V$ and, accordingly, $\boldsymbol{p}_i = \boldsymbol{x}_{i,1}$, and consider the *Attention-only* model unless stated otherwise; however the derivations for *Attention-only* model holds also for *Attention-MLP*. We study the pre-attention scores given in (7), with the explicit formula in (8)–(9). We derive the terms in (6) in three parts:

- In the first part, we analyze the informative component (9) and derive the scaling of the *Signal* term in (6).
- In the second part, we analyze the non-informative component (8) and derive the scaling of *Gradient noise* and *Mean bias* in (6), corresponding to its mean and bias components.
- In the third part, we consider the *Attention-MLP* model and derive the scaling of *MLP noise* in (8).

Before proceeding, we note that both the informative and non-informative terms in (8)–(9) depend on the first iterate of the output layer, $\boldsymbol{V}^{(1)}$, which can be decomposed into mean, bias and gradient noise components as

$$\boldsymbol{V}^{(1)} \approx \boldsymbol{Z}_{\text{out}}\Big(\frac{1}{NL}\sum_{i=1}^{N}(\boldsymbol{x}_{i,1} - \tfrac{1}{V}\mathbb{1}_V)(\boldsymbol{X}_i\mathbb{1}_L)^\top\Big)\boldsymbol{Z}_{\text{in}}^\top \tag{11}$$

$$\approx \boldsymbol{Z}_{\text{out}}\Big(\underbrace{\frac{1}{VL}(\boldsymbol{I}_V - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)}_{\text{Mean}} + \underbrace{\frac{1}{VN}\sum_{i=1}^{N}(\boldsymbol{x}_{i,1} - \tfrac{1}{V}\mathbb{1}_V)\mathbb{1}_V^\top}_{\text{Bias}} + \underbrace{\frac{1}{\sqrt{LVN}}\boldsymbol{\Xi}}_{\text{Gradient noise}}\Big)\boldsymbol{Z}_{\text{in}}^\top \tag{12}$$

where the gradient noise component is given by

$$\boldsymbol{\Xi} := \sqrt{\frac{V}{LN}}\Big(\sum_{i=1}^{N}(\boldsymbol{x}_{i,1} - \frac{1}{V}\mathbb{1}_V)(\boldsymbol{X}_i\mathbb{1}_L - \tfrac{L}{V}\mathbb{1}_V)^\top - \frac{1}{V}(\boldsymbol{I}_V - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\Big).$$

Here, the bias term arises from aggregating tokens at initialization: the aggregate-token averages $\frac{1}{L}\boldsymbol{X}_i\mathbb{1}_L$ concentrate around their mean $\frac{1}{V}\mathbb{1}_V$ as L grows, so this effect appears as the bias term. The gradient-noise term captures finite-sample fluctuations of tokens around this mean. We explicitly factor out the typical size $1/\sqrt{VLN}$ in (12) so that the remaining matrix $\boldsymbol{\Xi}$ stays of constant size on average, i.e., $\mathbb{E}[\|\boldsymbol{\Xi}\|_2^2] = O(1)$. We are now ready to consider the cases listed above.

Informative term. With the decomposition in (12), the informative term in (9) can be written as the sum of two contributions:

$$\text{Informative} \approx \frac{1}{VL^2}\underbrace{\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_{i,1}^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{I}_V - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\boldsymbol{Z}_{\text{out}}^\top\boldsymbol{Z}_{\text{out}}(\boldsymbol{x}_{i,1} - \frac{1}{V}\mathbb{1}_V)}_{=O(1)}$$

The first term is due to the mean component; the second term is due to the gradient-noise $\boldsymbol{\Xi}$ component in (12). The bias-related terms are ignored, as they do not contribute. By standard concentration arguments for Gaussian matrices, the first term remains $O(1)$, whereas the second term concentrates within $\pm(\log V)/d$, yielding the *Signal* component (6) (the noise component is weaker than the remaining terms in (6)). See the"*Concentration bound for* $\texttt{score}_{12}$" part for the formal proof.

Non-informative term. In this part, we consider large L regime where

$$\frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i\boldsymbol{X}_i^\top\boldsymbol{Z}_{\text{in}}^\top \approx \frac{1}{d}\boldsymbol{I}_d. \tag{13}$$

We consider an arbitrary row of $\boldsymbol{X}_{\text{in}}$, which we denote with $\boldsymbol{x}_{\text{in}}$. With this approximation, we can write the non-informative term as

$$\text{Non-informative} \approx \frac{1}{d\sqrt{LNV}}\underbrace{\boldsymbol{x}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{\Xi}\boldsymbol{Z}_{\text{out}}^\top\boldsymbol{Z}_{\text{out}}\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_{i,1} - \frac{1}{V}\mathbb{1}_V)}_{\in\sqrt{\frac{V}{N}}\left[\frac{-\log V}{d}, \frac{\log V}{d}\right]}$$

$$+\frac{1}{Vd}\underbrace{\boldsymbol{x}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \mathbb{1}_V}_{\in\sqrt{\frac{V}{d}}[-\log V, \log V]} \underbrace{\left\|\boldsymbol{Z}_{\text{out}} \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_{i,1} - \frac{1}{V}\mathbb{1}_V)\right\|_2^2}_{\approx \frac{1}{N}},$$

where we ignore terms depending on the mean component in (11), as they do not contribute. Here, the first term is due to the gradient-noise component Ξ; by standard concentration arguments, this yields the scaling of the *Gradient noise* term. The second term arises from the bias term in (12); using standard concentration, this gives the scaling of the *Mean bias* term in (6). See the "*Concentration bound for* $\texttt{score}_{11}$" part for the formal proof.

MLP-noise in Attention-MLP. We denote the rows of $\boldsymbol{W}_{\text{in}}$ by $\{\boldsymbol{w}_k\}_{k=1}^m$, where $\boldsymbol{w}_k \sim \mathcal{N}(0, \boldsymbol{I}_d)$. We work in the large-$L$ regime for illustration, adopting the approximation in (13); however, the result extends to general L. Under this approximation, MLP-noise can be written as

$$\text{MLP-noise} \approx \boldsymbol{x}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}^\top \frac{1}{N^2 d}\sum_{i,j=1}^{N}\Bigg(\frac{1}{m}\sum_{k=1}^{m}\boldsymbol{w}_k \phi'\Big(\tfrac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i \mathbb{1}_L\Big)\phi\Big(\tfrac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j \mathbb{1}_L\Big) \times \big(\boldsymbol{x}_{i,1} - \tfrac{1}{V}\mathbb{1}_V\big)\boldsymbol{Z}_{\text{out}}^\top \boldsymbol{Z}_{\text{out}}\big(\boldsymbol{x}_{j,1} - \tfrac{1}{V}\mathbb{1}_V\big)\Bigg).$$

For large L, we have $\left\|\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i \mathbb{1}_L\right\|_2 \approx L^{-1/2} \to 0$, hence

$$\phi'\Big(\tfrac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i \mathbb{1}_L\Big)\phi\Big(\tfrac{1}{L}\boldsymbol{w}k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j \mathbb{1}_L\Big) \to \underbrace{\phi(0)\phi'(0)}_{\text{nonzero constant}},$$

where Assumption 4 ensures $\phi(0)\phi'(0) \neq 0$. Replacing the ϕ-dependent factors by this constant yields

$$\text{MLP noise} \approx \frac{\phi(0)v\phi'(0)}{d}\underbrace{\frac{1}{m}\sum_{k=1}^{m}\boldsymbol{x}_{\text{input}}^\top \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{w}_k;}_{\in\left[-\frac{\log V}{\sqrt{m}}, \frac{\log V}{\sqrt{m}}\right]}\underbrace{\left\|\boldsymbol{Z}_{\text{out}}\frac{1}{N}\sum_{i=1}^{N}\big(\boldsymbol{x}_{i,1} - \tfrac{1}{V}\mathbb{1}_V\big)\right\|_2^2}_{\approx\frac{1}{N}},$$

which, by standard concentration arguments, gives the scaling of the *MLP noise* term in (6). See the "*Concentration bound for* $\mathbf{s_3}$" part for the formal proof.

C.2 Attention scores and their asymptotic scaling

Assumption 4 (Technical conditions)**.** *We work under the following conditions:*

- ***Permutation.*** *Without loss of generality, assume* $\boldsymbol{\Pi} = \boldsymbol{I}_V$.
- ***Learning rates.*** *Take* $\eta = o_V(1)$, *chosen sufficiently small so that any* $o_\eta(1)$ *terms are negligible; in particular, we may write* $\hat{\mathbf{p}}_1 = \frac{1}{V}\mathbb{1}_V + o_\eta(1)$.
- ***Activation.*** *We consider a polynomial activation* ϕ *with a degree of* $p_\star$ *satisfying:*
  - $\phi(0), \phi'(0), \phi''(0) \neq 0$
  - *The smallest non-zero Hermite component of* ϕ *has index* $q_\star$, *i.e,* $q^\star := \min\{k > 0 | \mathbb{E}[\phi(Z)H_{e_k}] \neq 0\}$, *for* $Z \sim N(0, 1)$.

By using the technical condition above and ignoring the vanishing terms due to learning rate, we decompose the attention scores in to three terms $\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3 \in \mathbb{R}^L$.:

$$\mathsf{X}\mathsf{Z}_{\text{in}}^\top \boldsymbol{W}_{\text{KQ}}^{(1)}$$

$$= \frac{\eta\gamma}{N^2 L^2}\mathsf{X}\mathsf{Z}_{\text{in}}^\top \mathsf{Z}_{\text{in}}\sum_{i,j=1}^{N}\alpha_{ij}\mathsf{X}_i^\top\big(\boldsymbol{I}_L - \tfrac{1}{L}\mathbb{1}_L\mathbb{1}_L^\top\big)\boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top \mathbb{1}_L$$

$$
\begin{aligned}
&\qquad\qquad\qquad\qquad\qquad \times (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top \boldsymbol{Z}_{\text{out}}^\top \boldsymbol{Z}_{\text{out}} (\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V) \\
&+ \frac{\eta\gamma}{N^2L^2} \mathsf{X}\mathsf{Z}_{\text{in}}^\top \mathsf{Z}_{\text{in}} \sum_{i,j=1}^N \beta_{ij} \mathsf{X}_i^\top \big(\boldsymbol{I}_L - \tfrac{1}{L}\mathbb{1}_L\mathbb{1}_L^\top\big) \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L \\
&\qquad\qquad\qquad\qquad\qquad \times (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top \boldsymbol{Z}_{\text{out}}^\top \boldsymbol{Z}_{\text{out}} (\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V) \\
&+ \frac{\eta\gamma}{N^2L} \mathsf{X}\mathsf{Z}_{\text{in}}^\top \mathsf{Z}_{\text{in}} \sum_{i,j=1}^N \mathsf{X}_i^\top \big(\boldsymbol{I}_L - \tfrac{1}{L}\mathbb{1}_L\mathbb{1}_L^\top\big) \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \mathrm{FW}(\boldsymbol{W}_{\text{in}}; \boldsymbol{Z}_{\text{in}}, \boldsymbol{X}_i, \boldsymbol{X}_j) \\
&\qquad\qquad\qquad\qquad\qquad \times (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top \boldsymbol{Z}_{\text{out}}^\top \boldsymbol{Z}_{\text{out}} (\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V) \\
&=: \eta\gamma\big(\boldsymbol{s}_1 + \boldsymbol{s}_2 + \boldsymbol{s}_3\big).
\end{aligned}
$$

The following theorem characterizes the scaling of each term:

Theorem 3. *With probability at least $1 - o_V(1)$, we have the following:*

$$
\begin{aligned}
\boldsymbol{e}_l^\top \boldsymbol{s}_1 &\asymp \mathbb{1}_{l=1}\Big(\frac{1}{VL^2} \pm \frac{1}{\sqrt{NV}L^{3/2}d}\Big) \pm \frac{1}{N\sqrt{L}d(d \wedge L^2)^{1/2}(d \wedge L)^{1/2}} \\
\boldsymbol{e}_l^\top \boldsymbol{s}_2 &\asymp \pm\Big(\frac{1}{N\sqrt{L}d(L \wedge d)} + \frac{1}{NLd(L \wedge d)^{1/2}}\Big) \\
\boldsymbol{e}_l^\top \boldsymbol{s}_3 &\asymp \frac{\pm 1}{Nd\sqrt{m}}.
\end{aligned}
$$

Moreover, for notational convenience, we define

$$
\begin{aligned}
\boldsymbol{A}_{1,ir} &\coloneqq \boldsymbol{Z}_{\text{in}} \Big(\frac{1}{LN} \sum_{j=1}^N \alpha_{ij} (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top\Big) \\
&\qquad\qquad \times \Big(\frac{1}{LN} \sum_{j=1}^N \alpha_{rj} (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top\Big) \boldsymbol{Z}_{\text{in}}^\top \qquad (14) \\
\boldsymbol{A}_{2,ir} &\coloneqq \boldsymbol{Z}_{\text{in}} \Big(\frac{1}{LN} \sum_{j=1}^N \alpha_{ij} (\boldsymbol{N}_j^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top\Big) \\
&\qquad\qquad \times \Big(\frac{1}{LN} \sum_{j=1}^N \alpha_{rj} (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big) \boldsymbol{Z}_{\text{in}}^\top \qquad (15) \\
\boldsymbol{A}_{3,ir} &\coloneqq \frac{1}{L^2V^2} \Big(\frac{1}{N} \sum_{j=1}^N \alpha_{ij} (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)\Big)^\top \Big(\frac{1}{N} \sum_{j=1}^N \alpha_{rj} (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)\Big) \boldsymbol{Z}_{\text{in}} \mathbb{1}_V \mathbb{1}_V^\top \boldsymbol{Z}_{\text{in}}^\top
\end{aligned}
$$

and

$$
\begin{aligned}
\boldsymbol{S}_1 &\coloneqq \Big(\frac{1}{LN} \sum_{j=1}^N (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top\Big)\Big(\frac{1}{LN} \sum_{j=1}^N (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top\Big) \qquad (16) \\
\boldsymbol{S}_2 &\coloneqq \Big(\frac{1}{LN} \sum_{j=1}^N (\mathsf{N}_j^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top\Big) \\
&\qquad\qquad \times \Big(\frac{1}{LN} \sum_{j=1}^N (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big) \qquad (17) \\
\boldsymbol{S}_3 &\coloneqq \frac{1}{L^2V^2} \Big(\frac{1}{N} \sum_{j=1}^N (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)\Big)^\top \Big(\frac{1}{N} \sum_{j=1}^N (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)\Big) \mathbb{1}_V \mathbb{1}_V^\top. \qquad (18)
\end{aligned}
$$

We first make an observation that we will frequently rely on in the following:

Proposition 1. *For any $p \in \mathbb{N}$, we have*

$$\mathbb{E}[\|\boldsymbol{A}_{1,ir}\|_2^p] \vee \mathbb{E}[\|\boldsymbol{A}_{2,ir}\|_2^p] \vee \mathbb{E}[\|\boldsymbol{A}_{3,ir}\|_2^p] \le \mathrm{poly}_{p,p_\star}(d,V,L).$$

where $\mathrm{poly}_{p,p_\star}(N,d,V,L)$ *denotes a polynomial function of* (d,V,L) *whose degree depends on* $(p,p_\star)$.

Proof. By Proposition 12, we observe that $\alpha_{ij} \le \mathrm{poly}_{p_\star}(d,V,L)$. Therefore, we have

$$\|\boldsymbol{A}_{1,ir}\|_2 \vee \|\boldsymbol{A}_{2,ir}\|_2 \vee \|\boldsymbol{A}_{3,ir}\|_2 \le \mathrm{poly}_{p_\star}(d,V,L)\|\mathsf{Z}_{\mathrm{in}}\mathsf{Z}_{\mathrm{in}}^\top\|_2.$$

from which the result follows. □

C.3 Proof of Theorem 3

We observe that

$$\begin{aligned}\mathsf{X}\mathsf{Z}_{\mathrm{in}}^\top\mathsf{Z}_{\mathrm{in}}\mathsf{X}_i^\top &= \left(\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}^\top + \delta\mathsf{Z}_{\mathrm{in}}\boldsymbol{e}_{V+1}\boldsymbol{e}_1^\top\right)^\top\left(\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top + \delta\mathsf{Z}_{\mathrm{in}}\boldsymbol{e}_{V+1}\boldsymbol{e}_1^\top\right)\\ &= \boldsymbol{X}\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top + \delta\boldsymbol{e}_1\mathsf{z}_{\mathrm{trig}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top + \delta\boldsymbol{X}\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\mathrm{trig}}\boldsymbol{e}_1^\top + \delta^2\|\mathsf{z}_{\mathrm{trig}}\|_2^2\boldsymbol{e}_1\boldsymbol{e}_1^\top. \end{aligned} \tag{19}$$

In the following, we will consider $\boldsymbol{x}_l = \boldsymbol{e}_\nu$, for $\nu \in [V]$. We will write

$$\delta\boldsymbol{e}_1^\top\boldsymbol{e}_l\mathsf{z}_{\mathrm{trig}} + \boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{X}^\top\boldsymbol{e}_l = \mathsf{z}_\nu + \mathbb{1}_{k=1}\delta\mathsf{z}_{\mathrm{trig}} =: \mathsf{z}_{\nu,\delta} \tag{20}$$

and

$$\delta\boldsymbol{e}_1\mathsf{z}_{\mathrm{trig}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}^\top\boldsymbol{e}_l + \delta^2\|\mathsf{z}_{\mathrm{trig}}\|_2^2\boldsymbol{e}_1\boldsymbol{e}_1^\top\boldsymbol{e}_l = \delta\mathsf{z}_{\nu,\delta}^\top\mathsf{z}_{\mathrm{trig}}\boldsymbol{e}_1 =: \delta s_{\nu,\delta}\boldsymbol{e}_1. \tag{21}$$

In the following, we will consider the event.

$$\texttt{Event} \coloneqq (E.1) \cap (E.2).$$

C.3.1 Concentration bound for s_1

By (19)-(20)-(21), we can write that

$$\begin{aligned}\boldsymbol{e}_l^\top\boldsymbol{s}_1 &= \frac{1}{N^2L^2}\sum_{i,j=1}^N \alpha_{ij}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top\big(\boldsymbol{I}_L - \frac{1}{L}\mathbb{1}_L\mathbb{1}_L^\top\big)\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_j^\top\mathbb{1}_L(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top\boldsymbol{Z}_{\mathrm{out}}^\top\boldsymbol{Z}_{\mathrm{out}}(\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)\\ &+ \frac{\delta s_{\nu,\delta}}{N^2L^2}\sum_{i,j=1}^N \alpha_{ij}\big(\boldsymbol{e}_1 - \frac{1}{L}\mathbb{1}_L\big)^\top\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_j^\top\mathbb{1}_L(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top\boldsymbol{Z}_{\mathrm{out}}^\top\boldsymbol{Z}_{\mathrm{out}}(\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)\\ &=: \texttt{score}_{11} + \texttt{score}_{12}.\end{aligned}$$

We will analyze $\texttt{score}_{11}$ and $\texttt{score}_{12}$ separately. We define

$$\boldsymbol{V}_i \coloneqq \boldsymbol{V}_{i,1} + \boldsymbol{V}_{i,2} + \boldsymbol{V}_{i,3},$$

where

$$\boldsymbol{V}_{i,1} \coloneqq \Big(\frac{1}{NL}\sum_{j=1}^N \alpha_{ij}(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top\Big)$$

$$\boldsymbol{V}_{i,2} \coloneqq \Big(\frac{1}{NL}\sum_{j=1}^N \alpha_{ij}\big(\boldsymbol{N}_j^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top\big)\mathbb{1}_{L-1}(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top\Big)$$

$$\boldsymbol{V}_{i,3} \coloneqq \tfrac{1}{V}\mathbb{1}_V\Big(\frac{1}{NL}\sum_{j=1}^N \alpha_{ij}(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top\Big).$$

Concentration bound for $\texttt{score}_{11}$: We start with $\texttt{score}_{11}$. We define

$$\begin{aligned}\boldsymbol{C}_i &\coloneqq \frac{1}{L}(\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\big(\boldsymbol{I}_L - \frac{1}{L}\mathbb{1}_L\mathbb{1}_L^\top\big)\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}} \\ &= \underbrace{\frac{1}{L}(\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}}_{\coloneqq \boldsymbol{C}_{i,1}} - \underbrace{\frac{1}{L^2}(\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}}_{\coloneqq \boldsymbol{C}_{i,2}}\end{aligned}$$

By Chebyshev's inequality, with probability $1 - o_V(1)$,

$$\begin{aligned}\texttt{score}_{11} = \frac{1}{N}\sum_{i=1}^N \operatorname{tr}(\boldsymbol{V}_i\boldsymbol{Z}_{\text{out}}^\top\boldsymbol{Z}_{\text{out}}\boldsymbol{C}_i) &= \operatorname{tr}\Big(\boldsymbol{Z}_{\text{out}}\frac{1}{N}\sum_{i=1}^N \boldsymbol{C}_i\boldsymbol{V}_i\boldsymbol{Z}_{\text{out}}^\top\Big) \\ &= \underbrace{\operatorname{tr}\Big(\frac{1}{N}\sum_{i=1}^N \boldsymbol{C}_i\boldsymbol{V}_i\Big)}_{\texttt{score}_{111}} \pm \underbrace{\frac{1}{\sqrt{d}}\Big\|\frac{\log V}{N}\sum_{i=1}^N \boldsymbol{C}_i\boldsymbol{V}_i\Big\|_F}_{\texttt{score}_{112}}.\end{aligned}$$

We start with bounding $\texttt{score}_{112}$ term. We have

$$\texttt{score}_{112} \le \frac{1}{\sqrt{d}}\Big\|\frac{1}{N}\sum_{i=1}^N \boldsymbol{C}_i\boldsymbol{V}_{i,1}\Big\|_F + \frac{1}{\sqrt{d}}\Big\|\frac{1}{N}\sum_{i=1}^N \boldsymbol{C}_i\boldsymbol{V}_{i,2}\Big\|_F + \frac{1}{\sqrt{d}}\Big\|\frac{1}{N}\sum_{i=1}^N \boldsymbol{C}_i\boldsymbol{V}_{i,3}\Big\|_F$$

We have for $u \in \{1,2\}$

$$\Big\|\frac{1}{N}\sum_{i=1}^N \boldsymbol{C}_i\boldsymbol{V}_{i,u}\Big\|_F^2 \le \frac{2}{N^2}\Big(\sum_{i,r=1}^N \operatorname{tr}(\boldsymbol{C}_{i,1}\boldsymbol{V}_{i,u}\boldsymbol{V}_{r,u}^\top\boldsymbol{C}_{r,1}^\top) + \operatorname{tr}(\boldsymbol{C}_{i,2}\boldsymbol{V}_{i,u}\boldsymbol{V}_{r,u}^\top\boldsymbol{C}_{r,2}^\top)\Big).$$

We define

$$t_1 \coloneqq \frac{\phi'(0)^4}{dL^2}\big(\frac{1}{N} + (1-\frac{1}{V})\frac{1}{V}\big), \qquad t_2 \coloneqq \frac{\phi'(0)^4}{d}(1-\frac{1}{V})^2\frac{L-1}{L^2N},.$$

For $i \neq r$, by using the definition in $\boldsymbol{A}_{1,ir}$ and $\boldsymbol{A}_{2,ir}$ in (14)-(15), we have

$$\begin{aligned}&\operatorname{tr}(\boldsymbol{C}_{i,1}\boldsymbol{V}_{i,u}\boldsymbol{V}_{r,u}^\top\boldsymbol{C}_{r,1}^\top) = \frac{1}{L^2}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{A}_{u,ir}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta} \\ &= \frac{t_u}{L^2}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta} \\ &+ \frac{1}{L^2}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top(\boldsymbol{A}_{u,ir} - t_u\boldsymbol{I}_d)\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta} \\ &\le \frac{t_u}{L^2}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta} \\ &+ \frac{1}{L^2}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \frac{1}{V})\|\boldsymbol{A}_{u,ir} - t_u\boldsymbol{I}_d\|_2\|\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}\|_2\|\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}\|_2\end{aligned}$$

By Proposition 9, we have

$$\frac{t_u}{L^2}\mathbb{E}\left[(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\text{in}}\right] \le \frac{Ct_u}{L^2}\frac{1}{Vd}. \tag{22}$$

Moreover, by using $\texttt{Event}$ and Propositions 1 and 9, we have

$$\begin{aligned}&\frac{1}{L^2}\mathbb{E}\left[(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \frac{1}{V})\|\boldsymbol{A}_{u,ir} - t_u\boldsymbol{I}_d\|_2\|\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}\|_2\|\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}\|_2|\mathsf{Z}_{\text{in}}\right] \\ &\le \tfrac{C}{Vd(L\wedge d)}\begin{cases}\phi'(0)^2\Big(\frac{1}{NdL^3} + \frac{1}{VdL^2}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big) + \phi'(0)^4\Big(\frac{\log V}{L^2V^{3/2}\sqrt{d}} + \frac{\log^2 V}{L^2N\sqrt{V}d}\Big), & u=1 \\ \frac{\sqrt{V}}{d\sqrt{NL}}\Big(\frac{1}{NL^{\frac{3}{2}}} + \frac{1}{V\sqrt{L}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big) + \phi'(0)^4\Big(\frac{\log V}{NL\sqrt{V}d} + \frac{\log^3 V}{N\sqrt{LV}d}\Big), & u=2\end{cases} \\ &\le \frac{C}{N^{3/2}\sqrt{V}d^2L^2}\frac{1}{L\wedge d} + \frac{C}{V^{3/2}\sqrt{N}Ld^2}\frac{1}{L\wedge d}\frac{1}{V\wedge L^2\wedge L\sqrt{d}} + \frac{C\log^3 V}{NV^{3/2}L^{1/2}d^{3/2}}\frac{1}{(L\wedge d)^{3/2}}.\end{aligned} \tag{23}$$

On the other hand,

$$\begin{aligned}
&\operatorname{tr}(\boldsymbol{C}_{i,2}\boldsymbol{V}_{i,u}\boldsymbol{V}_{r,u}^\top\boldsymbol{C}_{r,2}^\top)\\
&= \frac{1}{L^4}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{A}_{u,ir}\boldsymbol{Z}_{\text{in}}\mathsf{X}_r^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\\
&= \frac{t_u}{L^4}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_r^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\\
&+ \frac{1}{L^4}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top (\boldsymbol{A}_{u,ir} - t_u\boldsymbol{I}_d)\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_r^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\\
&\le \frac{t_u}{L^4}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_r^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\\
&+ \frac{1}{L^4}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \tfrac{1}{V})\|\boldsymbol{A}_{u,ir} - t_u\boldsymbol{I}_d\|_2 \|\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_r^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\|_2 \|\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\|_2.
\end{aligned}$$

By using `Event`,

$$\tfrac{t_u}{L^4}\,\mathbb{E}\left[(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_r^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\text{in}}\right] \le \tfrac{Ct_u}{VL}\tfrac{\log^2 V}{V\wedge L^2\wedge L\sqrt{d}}\tfrac{1}{L\wedge d}.$$

Moreover, by using `Event`,

$$\begin{aligned}
&\tfrac{1}{L^4}\,\mathbb{E}\left[(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \tfrac{1}{V})\|\boldsymbol{A}_{u,ir} - t_u\boldsymbol{I}_d\|_2 \|\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_r^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\|_2 \|\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\|_2|\mathsf{Z}_{\text{in}}\right]\\
&\le \tfrac{C}{VL^2(L\wedge d)}\begin{cases} \phi'(0)^2\left(\tfrac{1}{NdL^3} + \tfrac{1}{VdL^2}\tfrac{1}{V\wedge L^2\wedge L\sqrt{d}}\right) + \phi'(0)^4\left(\tfrac{\log V}{L^2V^{3/2}\sqrt{d}} + \tfrac{\log^2 V}{L^2N\sqrt{V}d}\right), & u=1\\ \tfrac{\sqrt{V}}{d\sqrt{NL}}\left(\tfrac{1}{NL^{\frac{3}{2}}} + \tfrac{1}{V\sqrt{L}}\tfrac{1}{V\wedge L^2\wedge L\sqrt{d}}\right) + \phi'(0)^4\left(\tfrac{\log V}{NL\sqrt{Vd}} + \tfrac{\log^3 V}{N\sqrt{LVd}}\right), & u=2 \end{cases}\\
&\le \frac{C}{N^{3/2}\sqrt{V}dL^4}\frac{1}{L\wedge d} + \frac{C}{V^{3/2}\sqrt{N}L^3 d}\frac{1}{L\wedge d}\frac{1}{V\wedge L^2\wedge L\sqrt{d}} + \frac{C\log^3 V}{NV^{3/2}L^{5/2}\sqrt{d}}\frac{1}{(L\wedge d)^{3/2}}
\end{aligned}\tag{24}$$

On the other hand, for $i=r$, we have

$$\begin{aligned}
\operatorname{tr}(\boldsymbol{C}_{i,1}\boldsymbol{V}_{i,u}\boldsymbol{V}_{i,u}^\top\boldsymbol{C}_{i,1}^\top) + \operatorname{tr}(\boldsymbol{C}_{i,2}\boldsymbol{V}_{i,u}\boldsymbol{V}_{i,u}^\top\boldsymbol{C}_{i,2}^\top)
&= \frac{1}{L^2}(1-\frac{1}{V})\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{A}_{u,ii}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\\
&+ \frac{(1-\frac{1}{V})}{L^4}\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{A}_{u,ii}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\\
&\le \frac{t_u}{L^2}(1-\frac{1}{V})\|\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\|_2^2 + \frac{t_u}{L^4}\|\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\|_2^2.
\end{aligned}$$

By using `Event` and Proposition 9, we have

$$\begin{aligned}
\frac{t_u}{L^2}\,\mathbb{E}\left[\|\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\|_2^2|\mathsf{Z}_{\text{in}}\right] + \frac{t_u}{L^4}\,\mathbb{E}\left[\|\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\|_2^2|\mathsf{Z}_{\text{in}}\right]&\\
\le \frac{Ct_u}{L^2}\left(\frac{L}{d} + \frac{L^2}{d^2}\right) + \frac{Ct_u}{L^2}\frac{1}{L\wedge d}.&
\end{aligned}\tag{25}$$

Therefore, we have by (22)-(23)-(24)-(25) and using $N \ll VL$ and $L \ll V$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \boldsymbol{C}_i\boldsymbol{V}_{i,1}\right\|_F^2|\mathsf{Z}_{\text{in}}\right] + \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \boldsymbol{C}_i\boldsymbol{V}_{i,2}\right\|_F^2|\mathsf{Z}_{\text{in}}\right]\\
&\le \frac{C}{N^2dL(d\wedge L^2)(d\wedge L)} + \frac{C}{N^{3/2}\sqrt{V}dL^2(d\wedge L^2)(L\wedge d)}\\
&+ \frac{C}{V^{3/2}\sqrt{N}Ld(d\wedge L^2)(L\wedge d)}\frac{1}{V\wedge L^2\wedge L\sqrt{d}} + \frac{C\log^3 V}{NV^{3/2}\sqrt{Ld}(d\wedge L^2)(L\wedge d)^{3/2}}\\
&\le \frac{C}{N^2dL(d\wedge L^2)(d\wedge L)} + \frac{C\log^3 V}{NV^{3/2}\sqrt{Ld}(d\wedge L^2)(L\wedge d)^{3/2}}.
\end{aligned}\tag{26}$$

On the other hand, we have

$$\Big\|\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{C}_i\boldsymbol{V}_{i,3}\Big\|_F^2 \leq \frac{2}{N^2}\Big(\sum_{i,r=1}^{N}\mathrm{tr}(\boldsymbol{C}_{i,1}\boldsymbol{V}_{i,3}\boldsymbol{V}_{r,3}^\top\boldsymbol{C}_{r,1}^\top)+\mathrm{tr}(\boldsymbol{C}_{i,2}\boldsymbol{V}_{i,3}\boldsymbol{V}_{r,3}^\top\boldsymbol{C}_{r,2}^\top)\Big).$$

We define $t_3 \coloneqq \frac{\phi'(0)^4}{NV^2L^2}$ and

$$\tilde{\Delta}_{3,ir} \coloneqq \frac{1}{L^2V^2}\Big(\frac{1}{N}\sum_{j=1}^{N}\alpha_{ij}(\boldsymbol{x}_j-\tfrac{1}{V}\mathbb{1}_V)\Big)^\top\Big(\frac{1}{N}\sum_{j=1}^{N}\alpha_{rj}(\boldsymbol{x}_j-\tfrac{1}{V}\mathbb{1}_V)\Big)-\frac{\phi'(0)^4}{NV^2L^2}.$$

We have for $i\neq r$,

$$\begin{aligned}
&\mathrm{tr}(\boldsymbol{C}_{i,1}\boldsymbol{V}_{i,3}\boldsymbol{V}_{r,3}^\top\boldsymbol{C}_{r,1}^\top)+\mathrm{tr}(\boldsymbol{C}_{i,2}\boldsymbol{V}_{i,3}\boldsymbol{V}_{r,3}^\top\boldsymbol{C}_{r,2}^\top)\\
&=\frac{1}{L^2}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r}-\tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{A}_{3,ir}\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\nu,\delta}\\
&+\frac{1}{L^4}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r}-\tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{A}_{3,ir}\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_r^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_r\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\nu,\delta}\\
&\leq\frac{t_3}{L^2}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r}-\tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\nu,\delta}\\
&+\frac{t_3}{L^4}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r}-\tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_r^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_r\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\nu,\delta}\\
&+\frac{\tilde{\Delta}_{3,ir}}{L^2}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r}-\tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\nu,\delta}\\
&+\frac{\tilde{\Delta}_{3,ir}}{L^4}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r}-\tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_r^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_r\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\nu,\delta}
\end{aligned}$$

For the first term, by Proposition 9,

$$\frac{t_3}{L^2}\mathbb{E}\Big[(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r}-\tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\mathrm{in}}\Big]\leq\frac{C\phi'(0)^4}{NV^2L^4}\frac{\log^2V}{d^2}.$$

For the second term, by using `Event`, we have

$$\begin{aligned}
&\frac{t_3}{L^4}\mathbb{E}\Big[(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r}-\tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_r^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_r\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\mathrm{in}}\Big]\\
&\leq\frac{\phi'(0)^4}{NV^3L^4}\frac{1}{d}\Big(L\vee\frac{V}{d}\Big)
\end{aligned}$$

For the third item and fourth items, by using `Event` and Proposition 9,

$$\begin{aligned}
&\tfrac{1}{L^2}\mathbb{E}\Big[(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r}-\tfrac{1}{V})|\tilde{\Delta}_{3,ir}|\|\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\nu,\delta}\|_2\|\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\nu,\delta}\|_2|\mathsf{Z}_{\mathrm{in}}\Big]\\
&+\tfrac{1}{L^4}\mathbb{E}\Big[(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r}-\tfrac{1}{V})|\tilde{\Delta}_{3,ir}|\|\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\nu,\delta}\|_2\|\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_r^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_r\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\nu,\delta}\|_2|\mathsf{Z}_{\mathrm{in}}\Big]\\
&\leq\frac{C}{V^3L^4}\Big(\frac{L}{d}+\frac{V}{d^2}+\frac{L^2}{d^2}\Big)\Bigg(\frac{\phi'(0)^4\log^2V}{N\sqrt{V}}+\frac{\phi'(0)^2}{N}\Big(\frac{1}{NL}+\frac{1}{\sqrt{N}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big)\Bigg)\\
&+\frac{C}{V^3L^4}\Big(\frac{L}{d}+\frac{V}{d^2}+\frac{L^2}{d^2}\Big)\Big(\frac{1}{NL}+\frac{1}{\sqrt{N}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big)^2\\
&\leq\frac{C}{NV^3L^2d(L\wedge d)}\frac{\phi'(0)^4\log^2V}{\sqrt{V}\wedge L^4\wedge L^2d}+\frac{C}{NV^2L^4d^2}\frac{\phi'(0)^4\log^2V}{\sqrt{V}\wedge L^4\wedge L^2d}.
\end{aligned}\tag{27}$$

For $i=r$, by using `Event`, we have

$$\begin{aligned}
&\mathrm{tr}(\boldsymbol{C}_{i,1}\boldsymbol{V}_{i,3}\boldsymbol{V}_{i,3}^\top\boldsymbol{C}_{i,1}^\top)+\mathrm{tr}(\boldsymbol{C}_{i,2}\boldsymbol{V}_{i,3}\boldsymbol{V}_{i,3}^\top\boldsymbol{C}_{i,2}^\top)\\
&\qquad=\frac{1}{L^2}(1-\tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{A}_{3,ir}\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\nu,\delta}\\
&\qquad+\frac{1}{L^4}(1-\frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{A}_{3,ir}\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\nu,\delta}
\end{aligned}$$

$$\leq \frac{2t_3}{L^2}|\mathbb{1}_V^\top \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}|^2 + \frac{2t_3}{L^4}|\mathbb{1}_V^\top \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L \mathbb{1}_L^\top \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}|^2.$$

Then, by using Event and Proposition 9,

$$\mathbb{E}\left[\operatorname{tr}(\boldsymbol{C}_{i,1}\boldsymbol{V}_{i,3}\boldsymbol{V}_{i,3}^\top \boldsymbol{C}_{i,1}^\top) + \operatorname{tr}(\boldsymbol{C}_{i,2}\boldsymbol{V}_{i,3}\boldsymbol{V}_{i,3}^\top \boldsymbol{C}_{i,2}^\top)|\mathsf{Z}_{\text{in}}\right] \leq \frac{C\phi'(0)^4 \log^2 V}{NVd^2L^3}\Big(1+\frac{L}{d}\Big). \tag{28}$$

Therefore, by using (27)-(28) and using $L \ll V$ and $N \ll VL$, we have

$$\mathbb{E}\left[\Big\|\frac{1}{N}\sum_{i=1}^N \boldsymbol{C}_i \boldsymbol{V}_{i,3}\Big\|_F^2 |\mathsf{Z}_{\text{in}}\right] \ll \frac{1}{N^2 dL(d\wedge L^2)(d\wedge L)}. \tag{29}$$

Therefore, by (26)-(29), we have

$$\mathtt{score}_{112} \leq \frac{C\log V}{N\sqrt{L}d(d\wedge L^2)^{1/2}(d\wedge L)^{1/2}} + \frac{C\log^{5/2} V}{\sqrt{N}(Vd)^{3/4}L^{1/4}(d\wedge L^2)^{1/2}(L\wedge d)^{3/4}}.$$

For $\mathtt{score}_{111}$, we write

$$\mathtt{score}_{111} = \underbrace{\operatorname{tr}\Big(\frac{1}{N}\sum_{i=1}^N \boldsymbol{C}_i\boldsymbol{V}_{i,1}\Big) + \operatorname{tr}\Big(\frac{1}{N}\sum_{i=1}^N \boldsymbol{C}_i\boldsymbol{V}_{i,2}\Big)}_{:=\mathtt{score}_{1111}} + \underbrace{\operatorname{tr}\Big(\frac{1}{N}\sum_{i=1}^N \boldsymbol{C}_i\boldsymbol{V}_{i,3}\Big)}_{:=\mathtt{score}_{1112}}.$$

We have

$$\begin{aligned}
&\mathtt{score}_{1111}\\
&= \frac{1}{N^2L^2}\sum_{i,j=1}^N \alpha_{ij}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \tfrac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\big(\boldsymbol{I}_L - \frac{1}{L}\mathbb{1}_L\mathbb{1}_L^\top\big)\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_j^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\\
&= \frac{\phi'(0)^2}{N^2L^2}\sum_{j=1}^N \mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\Big(\sum_{\substack{i=1\\i\neq j}}^N(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \tfrac{1}{V}\mathbb{1}_V)\boldsymbol{X}_i^\top\big(\boldsymbol{I}_L - \tfrac{1}{L}\mathbb{1}_L\mathbb{1}_L^\top\big)\boldsymbol{X}_i\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_j^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\\
&+ \frac{(1-\frac{1}{V})}{N^2L^2}\sum_{j=1}^N \alpha_{jj}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\big(\boldsymbol{I}_L - \frac{1}{L}\mathbb{1}_L\mathbb{1}_L^\top\big)\boldsymbol{X}_j\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_j^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\\
&+ \frac{1}{N^2L^2}\sum_{j=1}^N\sum_{\substack{i=1\\i\neq j}}^N(\alpha_{ij}-\phi'(0)^2)(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\big(\boldsymbol{I}_L - \tfrac{1}{L}\mathbb{1}_L\mathbb{1}_L^\top\big)\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_j^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\\
&=: \mathtt{score}_{11111} + \mathtt{score}_{11112} + \mathtt{score}_{11113}.
\end{aligned}$$

We start with the last term. By using Cauchy-Schwartz inequality,

$$\begin{aligned}
|\mathtt{score}_{11113}| &\leq \Big(\frac{1}{N^2L^2}\sum_{j=1}^N\sum_{\substack{i=1\\i\neq j}}^N|\alpha_{ij}-\phi'(0)^2||\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \frac{1}{V}|\Big)\\
&\qquad\times \sup_{i\neq j\in[N]}|\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\big(\boldsymbol{I}_L - \frac{1}{L}\mathbb{1}_L\mathbb{1}_L^\top\big)\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_j^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L|\\
&\leq \Big(\frac{1}{N^2L^2}\sum_{j=1}^N\sum_{\substack{i=1\\i\neq j}}^N|\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \frac{1}{V}|\Big)\sup_{i\neq j\in[N]}|\alpha_{ij}-\phi'(0)^2|\\
&\qquad\times \sup_{i\neq j\in[N]}|\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\big(\boldsymbol{I}_L - \frac{1}{L}\mathbb{1}_L\mathbb{1}_L^\top\big)\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L|
\end{aligned}$$

$$\leq \frac{C \log V}{V\sqrt{Ld}(L \wedge d)} \frac{1}{V \wedge L^2 \wedge L\sqrt{d}}. \tag{30}$$

where we used Event in (30). Next, we consider $\texttt{score}_{11112}$:

$$|\texttt{score}_{11112}|$$
$$= \frac{(1-\frac{1}{V})}{N^2L^2} \sum_{j=1}^{N} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \Big(\alpha_{jj} \boldsymbol{X}_j^\top \boldsymbol{X}_j \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L - \mathbb{E}\left[\alpha_{jj} \boldsymbol{X}_j^\top \boldsymbol{X}_j \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L \,\middle|\, \mathsf{Z}_{\text{in}}\right] \Big)$$
$$+ \frac{(1-\frac{1}{V})}{NL^2} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \mathbb{E}\left[\alpha_{11} \boldsymbol{X}_1^\top \boldsymbol{X}_1 \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_1^\top \mathbb{1}_L \,\middle|\, \mathsf{Z}_{\text{in}}\right]$$
$$- \frac{(1-\frac{1}{V})}{N^2LV} \sum_{j=1}^{N} \alpha_{jj} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \boldsymbol{X}_j \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \mathbb{1}_V$$
$$- \frac{(1-\frac{1}{V})}{N^2L^3} \sum_{j=1}^{N} \alpha_{jj} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L \mathbb{1}_L^\top \boldsymbol{X}_j \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} (\boldsymbol{X}_j^\top - \frac{1}{V} \mathbb{1}_V \mathbb{1}_L^\top) \mathbb{1}_L$$

By using Event, we have

- For the first summand,

$$\mathbb{E}\left[\left(\frac{(1-\frac{1}{V})}{N^2L^2} \sum_{j=1}^{N} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \Big(\alpha_{jj} \boldsymbol{X}_j^\top \boldsymbol{X}_j \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L - \mathbb{E}\left[\alpha_{jj} \boldsymbol{X}_j^\top \boldsymbol{X}_j \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L\right] \Big) \right)^2 \,\middle|\, \mathsf{Z}_{\text{in}} \right]$$

$$\leq \frac{(1-\frac{1}{V})^2}{N^3L^4} \mathbb{E}\left[\alpha_{jj}^2 \Big(\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_1^\top \boldsymbol{X}_1 \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_1^\top \mathbb{1}_L \Big)^2 \,\middle|\, \mathsf{Z}_{\text{in}} \right]$$
$$\leq \frac{C\phi'(0)^4}{N^3L^3} \mathbb{E}\left[\left\| \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_1^\top \boldsymbol{X}_1 \boldsymbol{Z}_{\text{in}}^\top \right\|_2^2 \,\middle|\, \mathsf{Z}_{\text{in}} \right] + o_V(1)$$
$$\leq \frac{C\phi'(0)^4}{N^3Ld(L \wedge d)}$$

By Chebyshev's inequality with probability $1 - o_V(1)$, we have

$$\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \frac{(1-\frac{1}{V})}{N^2L^2} \sum_{j=1}^{N} \Big(\alpha_{jj} \boldsymbol{X}_j^\top \boldsymbol{X}_j \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L - \mathbb{E}\left[\alpha_{jj} \boldsymbol{X}_j^\top \boldsymbol{X}_j \quad \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L | \mathsf{Z}_{\text{in}}\right] \Big)$$
$$\leq \frac{C\phi'(0)^2 \log V}{N^{\frac{3}{2}} \sqrt{Ld} \sqrt{L \wedge d}}.$$

- For the second summand,

$$\frac{(1-\frac{1}{V})}{NL^2} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \mathbb{E}\left[\alpha_{11} \boldsymbol{X}_1^\top \boldsymbol{X}_1 \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_1^\top \mathbb{1}_L \,\middle|\, \mathsf{Z}_{\text{in}}\right]$$
$$= \frac{(1-\frac{1}{V})\phi'(0)^2}{NL^2} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \mathbb{E}\left[\boldsymbol{X}_1^\top \boldsymbol{X}_1 \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_1^\top \boldsymbol{X}_1 \,\middle|\, \mathsf{Z}_{\text{in}}\right] \mathbb{1}_V$$
$$+ \frac{(1-\frac{1}{V})}{NL^2} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \mathbb{E}\left[(\alpha_{11} - \phi'(0)^2) \boldsymbol{X}_1^\top \boldsymbol{X}_1 \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_1^\top \boldsymbol{X}_1 \,\middle|\, \mathsf{Z}_{\text{in}}\right] \mathbb{1}_V$$
$$= \frac{(1-\frac{1}{V})\phi'(0)^2}{NL} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \mathbb{E}\left[\boldsymbol{x}_1 \boldsymbol{x}_1^\top \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{x}_1 \boldsymbol{x}_1^\top \,\middle|\, \mathsf{Z}_{\text{in}}\right] \mathbb{1}_V$$
$$+ \frac{(1-\frac{1}{V})\phi'(0)^2}{NV^2} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \mathbb{1}_V$$
$$+ \frac{(1-\frac{1}{V})}{NLV} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \mathbb{E}\left[(\alpha_{11} - \phi'(0)^2) \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_1^\top \mathbb{1}_L | \mathsf{Z}_{\text{in}}\right]$$
$$+ \frac{(1-\frac{1}{V})}{NL^2} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \mathbb{E}\left[(\alpha_{11} - \phi'(0)^2) \big(\boldsymbol{X}_1^\top \boldsymbol{X}_1 - \frac{L}{V} \boldsymbol{I}_V \big) \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_1^\top \mathbb{1}_L | \mathsf{Z}_{\text{in}}\right]$$

$$\leq C\log V\Big(\frac{1}{NL\sqrt{V}d}+\frac{1}{N\sqrt{V}d^{3/2}}+\frac{1}{NL^{3/2}d}\Big).$$

- For the third summand,

$$\frac{1}{N^2LV}\sum_{j=1}^{N}\alpha_{jj}\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\boldsymbol{X}_j\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V$$
$$=\frac{\phi'(0)^2}{NV^2}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}^\top\mathbb{1}_V+\frac{\phi'(0)^2}{N^2LV}\sum_{j=1}^{N}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{X}_j^\top\boldsymbol{X}_j-\frac{L}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V$$
$$+\frac{1}{N^2LV}\sum_{j=1}^{N}(\alpha_{jj}-\phi'(0)^2)\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\boldsymbol{X}_j\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V.$$

The first term:

$$\Big|\frac{\phi'(0)^2}{NV^2}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}^\top\mathbb{1}_V\Big|\leq\frac{C\log V}{N\sqrt{V}d^{\frac{3}{2}}}.$$

The second term: By using

$$\mathbb{E}\Big[\Big(\sum_{j=1}^{N}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{X}_j^\top\boldsymbol{X}_j-\frac{L}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\Big)^2|\mathsf{Z}_{\text{in}}\Big]$$
$$=\sum_{j=1}^{N}\mathbb{E}\Big[\Big(\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{X}_j^\top\boldsymbol{X}_j-\frac{L}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\Big)^2|\mathsf{Z}_{\text{in}}\Big]=\frac{LVN}{d^2}.$$

Therefore, by Chebyshev's inequality, we have

$$\Big|\frac{\phi'(0)^2}{N^2LV}\sum_{j=1}^{N}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{X}_j^\top\boldsymbol{X}_j-\frac{L}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\Big|\leq\frac{\phi'(0)^2}{N^{3/2}\sqrt{VL}d}.$$

Finally,

$$\Big|\frac{1}{N^2LV}\sum_{j=1}^{N}(\alpha_{jj}-\phi'(0)^2)\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\boldsymbol{X}_j\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\Big|$$
$$\leq\frac{1}{NV}\|\mathsf{z}_{\nu,\delta}\|_2\Big\|\frac{1}{NL}\sum_{j=1}^{N}(\alpha_{jj}-\phi'(0)^2)\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\boldsymbol{X}_j\boldsymbol{Z}_{\text{in}}^\top\Big\|_2\Big\|\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\Big\|_2\leq\frac{C}{N\sqrt{VL}d}.$$

Therefore,

$$\Big|\frac{1}{N^2LV}\sum_{j=1}^{N}\alpha_{jj}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\boldsymbol{X}_j\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\Big|\leq C\log V\Big(\frac{1}{N\sqrt{V}d^{\frac{3}{2}}}+\frac{1}{N\sqrt{VL}d}\Big).$$

- For the last summand,

$$\frac{(1-\frac{1}{V})}{N^2L^3}\sum_{j=1}^{N}\alpha_{jj}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_j\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L$$
$$=\frac{(1-\frac{1}{V})}{N^2L^3}\sum_{j=1}^{N}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}$$
$$\times\Big(\alpha_{jj}\boldsymbol{X}_j^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_j\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L-\mathbb{E}\left[\alpha_{jj}\boldsymbol{X}_j^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_j\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\;|\mathsf{Z}_{\text{in}}\right]\Big)$$
$$+\frac{(1-\frac{1}{V})\phi'(0)^2}{NL^3}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\,\mathbb{E}\left[\boldsymbol{X}_1^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_1\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_1^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\;|\mathsf{Z}_{\text{in}}\right]$$

$$+\frac{(1-\frac{1}{V})}{NL^3}\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\,\mathbb{E}\left[(\alpha_{11}-\phi'(0)^2)\boldsymbol{X}_1^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_1\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_1^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\;\middle|\mathsf{Z}_{\text{in}}\right]. \quad (31)$$

We have

$$\begin{aligned}
&\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\,\mathbb{E}\left[\left(\alpha_{jj}\boldsymbol{X}_1^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_1\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_1^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\right)^2\;\middle|\mathsf{Z}_{\text{in}}\right]\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\\
&\quad\le C\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\,\mathbb{E}\left[\boldsymbol{X}_1^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_1\right]\boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}+o_V(1)\\
&\quad\le C\Big(\frac{L}{d}+\frac{L^2}{Vd}\Big).
\end{aligned}$$

Moreover, by using Proposition 6

$$\begin{aligned}
&\mathbb{E}\left[\boldsymbol{X}_1^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_1\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_1^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\;\middle|\mathsf{Z}_{\text{in}}\right]\\
&=\mathbb{E}\left[\boldsymbol{X}_1^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_1\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_1^\top \mathbb{1}_L\;\middle|\mathsf{Z}_{\text{in}}\right]-\frac{L}{V}\,\mathbb{E}\left[\boldsymbol{X}_1^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_1\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}\mathbb{1}_V\;\middle|\mathsf{Z}_{\text{in}}\right]\\
&=L\;\mathbb{E}\left[\boldsymbol{x}_1\boldsymbol{x}_1^\top \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{x}_1\;\middle|\mathsf{Z}_{\text{in}}\right]+\Big(\frac{L(L-1)}{V^2}\text{tr}(\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}})-\frac{2L(L-1)}{V^3}\mathbb{1}_V^\top \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}\mathbb{1}_V\Big)\mathbb{1}_V\\
&+\frac{L(L-2)}{V^2}\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}\mathbb{1}_V
\end{aligned}$$

Lastly,

$$\begin{aligned}
&\Big|\frac{1}{NL^3}\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\,\mathbb{E}\left[(\alpha_{11}-\phi'(0)^2)\boldsymbol{X}_1^\top \mathbb{1}_L\mathbb{1}_L^\top \boldsymbol{X}_1\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_1^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\;\middle|\mathsf{Z}_{\text{in}}\right]\Big|\\
&\quad\le\frac{1}{NL^3}\,\mathbb{E}\left[|\alpha_{11}-\phi'(0)^2||\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_1^\top \mathbb{1}_L|\;|\mathbb{1}_L^\top \boldsymbol{X}_1\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_1^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L|\;\middle|\mathsf{Z}_{\text{in}}\right]\\
&\quad\le\frac{C\log V}{NL^{5/2}\sqrt{d}}.
\end{aligned}$$

Therefore, by Chebyshev's inequality, with probability $1-o_V(1)$, we have

$$(31)\le C\log V\Big(\frac{1}{NL^2\sqrt{L\wedge d}}+\frac{1}{NL\sqrt{Vd}}\Big)$$

Therefore, we have

$$|\texttt{score}_{11112}|\le C\log V\Big(\frac{1}{N\sqrt{V(L\wedge d)}\sqrt{d}}+\frac{1}{NL^2\sqrt{L\wedge d}}\Big). \quad (32)$$

Finally, we consider $\texttt{score}_{11111}$:

$$\begin{aligned}
&\texttt{score}_{11111}\\
&=(1-\tfrac{1}{L})\frac{\phi'(0)^2}{NL}\sum_{j=1}^N\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\frac{1}{NL}\sum_{\substack{i=1\\i\neq j}}^N(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\tfrac{1}{V})\boldsymbol{x}_i\boldsymbol{x}_i^\top\Big)\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_j^\top-\tfrac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\\
&-\frac{\phi'(0)^2}{NL^2}\sum_{j=1}^N\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\frac{1}{NL}\sum_{\substack{i=1\\i\neq j}}^N(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\tfrac{1}{V})(\boldsymbol{x}_i\mathbb{1}_L^\top \boldsymbol{X}_i+\boldsymbol{X}_i^\top \mathbb{1}_L\boldsymbol{x}_i^\top)\Big)\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_j^\top-\tfrac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\\
&+\frac{\phi'(0)^2}{NL}\sum_{j=1}^N\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\frac{1}{NL}\sum_{\substack{i=1\\i\neq j}}^N(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\tfrac{1}{V})\boldsymbol{N}_i^\top \boldsymbol{N}_i\Big)\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}(\boldsymbol{x}_j-\tfrac{1}{V}\mathbb{1}_V)\\
&+\frac{\phi'(0)^2}{NL}\sum_{j=1}^N\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\frac{1}{NL}\sum_{\substack{i=1\\i\neq j}}^N(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\tfrac{1}{V})\boldsymbol{N}_i^\top \boldsymbol{N}_i\Big)\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}(\boldsymbol{N}_j^\top-\tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}
\end{aligned}$$

$$- \frac{\phi'(0)^2}{NL^2} \sum_{j=1}^{N} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \Big(\frac{1}{NL} \sum_{\substack{i=1 \\ i \neq j}}^{N} (\mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} - \tfrac{1}{V}) \boldsymbol{N}_i^\top \mathbb{1}_L \mathbb{1}_L^\top \boldsymbol{N}_i \Big) \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} (\boldsymbol{N}_j^\top - \tfrac{1}{V} \mathbb{1}_V \mathbb{1}_{L-1}^\top) \mathbb{1}_{L-1}$$

$$=: \texttt{score}_{V1} + \texttt{score}_{V2} + \texttt{score}_{V3} + \texttt{score}_{V4} + \texttt{score}_{V5}.$$

For the first summand, we write

$$\texttt{score}_{V1} := \underbrace{(1 - \frac{1}{L}) \frac{\phi'(0)^2}{NL} \sum_{j=1}^{N} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \Big(\frac{1}{NL} \sum_{\substack{i=1 \\ i \neq j}}^{N} (\mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} - \frac{1}{V}) \boldsymbol{x}_i \boldsymbol{x}_i^\top \Big) \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} (\boldsymbol{x}_j - \frac{1}{V} \mathbb{1}_V)}_{:= \texttt{score}_{V11}}$$

$$+ \underbrace{(1 - \frac{1}{L}) \frac{\phi'(0)^2}{NL} \sum_{j=1}^{N} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \Big(\frac{1}{NL} \sum_{\substack{i=1 \\ i \neq j}}^{N} (\mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} - \frac{1}{V}) \boldsymbol{x}_i \boldsymbol{x}_i^\top \Big) \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} (\boldsymbol{N}_j^\top - \frac{1}{V} \mathbb{1}_V \mathbb{1}_{L-1}^\top) \mathbb{1}_{L-1}}_{:= \texttt{score}_{V12}}.$$

We have by $\texttt{Event}$

$$|\texttt{score}_{V11}| \le \Big(\frac{\phi'(0)^2}{N^2 L^2} \sum_{j=1}^{N} \sum_{\substack{i=1 \\ i \neq j}}^{N} |\mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} - \frac{1}{V}| \Big) \sup_{i \neq j} \Big| \boldsymbol{x}_i \mathbb{1}_{\boldsymbol{x}_i \neq \boldsymbol{e}_\nu} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{x}_i^\top \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} (\boldsymbol{x}_j - \frac{1}{V} \mathbb{1}_V) \Big|$$

$$\le \Big(\frac{\phi'(0)^2}{N^2 L^2} \sum_{j=1}^{N} \sum_{\substack{i=1 \\ i \neq j}}^{N} \Big| \mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} - \frac{1}{V}| \Big) \sup_{i \neq j} | \mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{e}_\nu} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} (\boldsymbol{x}_j - \frac{1}{V} \mathbb{1}_V) \Big|$$

$$\le \frac{C \log V}{V L^2 \sqrt{d}}.$$

Moreover, let

$$\texttt{score}_{V12} =: (1 - \frac{1}{L}) \frac{\phi'(0)^2}{NL} \sum_{j=1}^{N} \texttt{score}_{V12,j}.$$

We have $\mathbb{E}[\texttt{score}_{V12,j}] = 0$ and $\mathbb{E}[\texttt{score}_{V12,j} \texttt{score}_{V12,j'}] = 0$ for $j \neq j'$, and

$$\mathbb{E}[\texttt{score}_{V12,j}^2]$$

$$\le \frac{CL}{d} \, \mathbb{E} \Big[\mathbb{1}_{\boldsymbol{x}_j = \boldsymbol{e}_k} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \Big(\frac{1}{NL} \sum_{\substack{i=1 \\ i \neq j}}^{N} \mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} \boldsymbol{x}_i \boldsymbol{x}_i^\top \Big) \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \Big(\frac{1}{NL} \sum_{\substack{i=1 \\ i \neq j}}^{N} \mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} \boldsymbol{x}_i \boldsymbol{x}_i^\top \Big) \boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta} \Big]$$

$$+ \frac{CL}{d} \, \mathbb{E} \Big[\mathbb{1}_{\boldsymbol{x}_j \neq \boldsymbol{e}_k} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \Big(\frac{1}{NL} \sum_{\substack{i=1 \\ i \neq j}}^{N} \mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} \boldsymbol{x}_i \boldsymbol{x}_i^\top \Big) \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \Big(\frac{1}{NL} \sum_{\substack{i=1 \\ i \neq j}}^{N} \mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} \boldsymbol{x}_i \boldsymbol{x}_i^\top \Big) \boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta} \Big]$$

$$+ \frac{CL}{dV^2} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \Big(\frac{1}{NL} \sum_{\substack{i=1 \\ i \neq j}}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top \Big) \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \Big(\frac{1}{NL} \sum_{\substack{i=1 \\ i \neq j}}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top \Big) \boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}$$

$$\le \frac{C}{V L d^2}.$$

Therefore, by Chebyshev's inequality with probability $1 - o_V(1)$, we have

$$|\texttt{score}_{V12,j}| \le \frac{C \log V}{\sqrt{NV} L^{3/2} d}.$$

Therefore,

$$|\texttt{score}_{V1}| \le \frac{C \log V}{V L^2 \sqrt{d}} + \frac{C \log V}{\sqrt{NV} L^{3/2} d}.$$

Moreover, for the second term, we write

$$
\begin{aligned}
&|\mathtt{score}_{V2}| \\
&\leq \Big(\frac{\phi'(0)^2}{N^2L^3}\sum_{j=1}^{N}\sum_{\substack{i=1\\i\neq j}}^{N}|\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\tfrac{1}{V}|\Big) \\
&\qquad\qquad \times \sup_{i\neq j}|\mathbb{1}_{\boldsymbol{x}_i\neq\boldsymbol{e}_\nu}\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}(\boldsymbol{x}_i\mathbb{1}_L^\top\boldsymbol{X}_i+\boldsymbol{X}_i^\top\mathbb{1}_L\boldsymbol{x}_i^\top\Big)\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}(\boldsymbol{X}_j^\top-\tfrac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L| \\
&\leq \Big(\frac{\phi'(0)^2}{N^2L^3}\sum_{j=1}^{N}\sum_{\substack{i=1\\i\neq j}}^{N}|\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\tfrac{1}{V}|\Big) \\
&\qquad\qquad \times \sup_{i\neq j}|\mathbb{1}_{\boldsymbol{x}_i\neq\boldsymbol{e}_\nu}\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}(\boldsymbol{x}_i\mathbb{1}_L^\top\boldsymbol{X}_i+\boldsymbol{X}_i^\top\mathbb{1}_L\boldsymbol{x}_i^\top\Big)\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}(\boldsymbol{X}_j^\top-\tfrac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L| \\
&\leq \frac{C\log V}{VL^2d}.
\end{aligned}
$$

For the third term, we write

$$
\begin{aligned}
\mathtt{score}_{V3} &= \frac{\phi'(0)^2}{NL^2}\sum_{i=1}^{N}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{N}_i^\top\boldsymbol{N}_i\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\Big(\frac{1}{N}\sum_{\substack{j=1\\j\neq i}}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\tfrac{1}{V})(\boldsymbol{x}_j-\tfrac{1}{V}\mathbb{1}_V)\Big) \\
&= \frac{\phi'(0)^2}{NL^2}\sum_{i=1}^{N}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\mathrm{in}}\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i-\frac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\Big(\frac{1}{N}\sum_{\substack{j=1\\j\neq i}}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\tfrac{1}{V})(\boldsymbol{x}_j-\tfrac{1}{V}\mathbb{1}_V)\Big) \\
&+ \frac{\phi'(0)^2(L-1)}{NL^2V}\sum_{i=1}^{N}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\Big(\frac{1}{N}\sum_{\substack{j=1\\j\neq i}}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\tfrac{1}{V})(\boldsymbol{x}_j-\tfrac{1}{V}\mathbb{1}_V)\Big) \\
&=: \mathtt{score}_{V31}+\mathtt{score}_{V32}.
\end{aligned}
$$

For $\mathtt{score}_{V32}$, we note that

$$
\begin{aligned}
&\mathbb{E}\Big[\Big(\frac{1}{N}\sum_{\substack{j=1\\j\neq i}}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big)\Big(\frac{1}{N}\sum_{\substack{j=1\\j\neq i}}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big)^\top|\boldsymbol{x}_i\Big] \\
&\qquad\qquad \preceq C\Big(\frac{1}{NV}+\frac{1}{V^2}\Big)(\boldsymbol{x}_i-\frac{1}{V}\mathbb{1}_V)(\boldsymbol{x}_i-\frac{1}{V}\mathbb{1}_V)^\top+\frac{C}{NV^3}\boldsymbol{I}_V.
\end{aligned}
$$

Therefore, we have

$$
\mathbb{E}\Big[\Big(\frac{1}{N}\sum_{\substack{j=1\\j\neq i}}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big)\Big(\frac{1}{N}\sum_{\substack{j=1\\j\neq i}}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big)^\top\Big]\preceq\frac{C}{V^3}\boldsymbol{I}_V.
$$

We have

$$
\mathbb{E}[\mathtt{score}_{V32}^2]\leq\frac{C}{N^2L^2V^4d}\sum_{i=1}^{N}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{Z}_{\mathrm{in}}^\top\mathsf{z}_{\nu,\delta}\leq\frac{C}{NL^2V^2d^3}.
$$

Therefore,

$$
|\mathtt{score}_{V32}|\leq\frac{C\log V}{\sqrt{N}VLd^{\frac{3}{2}}}
$$

Moreover, we have

$$
\mathbb{E}[\mathtt{score}_{V32}^2]
$$

$$\leq \frac{C}{N^2V^2L^4d}\sum_{i=1}^{N}\mathbb{E}\left[\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i - \frac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i - \frac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}\right]$$

$$\leq \frac{C}{N^2V^2L^4d}\frac{L-1}{V}\sum_{i=1}^{N}\mathbb{E}\left[\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}\right] - \frac{C}{N^2V^2L^4d}\frac{L-1}{V^2}\sum_{i=1}^{N}\mathbb{E}\left[\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}\right]$$

$$\leq \frac{C}{NV^2L^3d^2}.$$

Therefore, by Chebyshev's inequality, we have

$$|\mathtt{score}_{V32}| \leq \frac{C\log V}{\sqrt{N}VL^{\frac{3}{2}}d}$$

For the fourth term, we have $\mathbb{E}[\mathtt{score}_{V4}] = 0$ and

$$\mathbb{E}[\mathtt{score}_{V4}^2] \leq \frac{C}{VN^4L^4}\sum_{j=1}^{N}\sum_{\substack{i=1\\i\neq j}}^{N}\mathbb{E}\left[\Big(\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{N}_i^\top\boldsymbol{N}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\Big)^2\right]$$

$$\leq \frac{C}{VN^4L^3d}\sum_{j=1}^{N}\sum_{\substack{i=1\\i\neq j}}^{N}\mathbb{E}\left[\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{N}_i^\top\boldsymbol{N}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{N}_i^\top\boldsymbol{N}_i\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}\right]$$

$$\leq \frac{C}{VN^2Ld^2(L\wedge d)}$$

Therefore, by Chebyshev's inequality with probability $1-o_V(1)$, we have

$$|\mathtt{score}_{V4}| \leq \frac{C\log V}{N\sqrt{VL}d\sqrt{L\wedge d}}$$

For the last term, we have $\mathbb{E}[\mathtt{score}_{V5}] = 0$ and

$$\mathbb{E}[\mathtt{score}_{V5}^2] \leq \frac{C}{N^4L^6V}\sum_{j=1}^{N}\sum_{\substack{i=1\\i\neq j}}^{N}\mathbb{E}\left[\Big(\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{N}_i^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{N}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\Big)^2\right]$$

$$\leq \frac{C}{N^4L^4dV}\sum_{j=1}^{N}\sum_{\substack{i=1\\i\neq j}}^{N}\mathbb{E}\left[\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{N}_i^\top\mathbb{1}_{L-1}\mathbb{1}_{L-1}^\top\boldsymbol{N}_i\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}\right]$$

$$\leq \frac{C}{N^2L^4dV}\Big(\frac{L}{d} + \frac{L^2}{Vd}\Big)$$

Therefore, by Chebyshev's inequality with probability $1-o_V(1)$, we have

$$|\mathtt{score}_{V5}| \leq \frac{C\log V}{NLd\sqrt{V}(L\wedge\sqrt{V})}.$$

Overall, we have

$$|\mathtt{score}_{11111}| \leq C\log V\Big(\frac{1}{VL^2\sqrt{d}} + \frac{1}{\sqrt{NV}L^{3/2}d} + \frac{1}{N\sqrt{VL}d\sqrt{L\wedge d}} + \frac{1}{NLd\sqrt{V}(L\wedge\sqrt{V})}\Big). \tag{33}$$

Therefore, by (30)-(32)-(33) and using $N \ll VL$ and $L \ll V$, we have

$$|\mathtt{score}_{1111}| \leq C\log V\Big(\frac{1}{VL^2\sqrt{d}} + \frac{1}{\sqrt{NV}L^{3/2}d} + \frac{1}{N\sqrt{VL}d\sqrt{L\wedge d}} + \frac{1}{NLd\sqrt{V}(L\wedge\sqrt{V})}\Big)$$

$$+ C\log V\Big(\frac{1}{N\sqrt{V}(L\wedge d)\sqrt{d}} + \frac{1}{NL^2\sqrt{L\wedge d}}\Big) + \frac{C\log V}{V\sqrt{Ld}(L\wedge d)}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}$$

$$\leq C\log V\Big(\frac{1}{N\sqrt{V}(L\wedge d)\sqrt{d}}+\frac{1}{\sqrt{NV}L^{3/2}d}+\frac{1}{VL^{3/2}d^2}+\frac{1}{V^2\sqrt{L}d^{3/2}}\Big)+\frac{C\log V}{VL^2(L\wedge d)^{1/2}}. \tag{34}$$

Finally,

$$\begin{aligned}
&\texttt{score}_{1112}\\
&=\frac{1}{N^2L^2V}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_{ij}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\big(\boldsymbol{I}_L-\frac{1}{L}\mathbb{1}_L\mathbb{1}_L^\top\big)\boldsymbol{X}_i\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\\
&=\frac{1}{N^2L^2V}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_{ij}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\Big(\boldsymbol{X}_i^\top\boldsymbol{X}_i-\frac{L}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\\
&+\frac{1}{N^2LV^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_{ij}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\\
&-\frac{1}{N^2L^3V}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_{ij}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V
\end{aligned}$$

By using `Event` and Proposition 9, we have

- For the third summand,

$$\frac{1}{L^2}\mathbb{E}\Big[\big(\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\big)^2|\mathsf{Z}_{\rm in}\Big]\leq\frac{C\log^4 V}{d}\Big(L+\frac{V}{d}\Big).$$

- For the first summand,

$$\begin{aligned}
&\mathbb{E}\Big[\big(\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\big(\boldsymbol{X}_i^\top\boldsymbol{X}_i-\frac{L}{V}\boldsymbol{I}_V\big)\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\big)^2|\mathsf{Z}_{\rm in}\Big]\\
&=\mathbb{E}\Big[\big(\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\big)^2|\mathsf{Z}_{\rm in}\Big]-\frac{L^2}{V^2}\big(\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\big)^2=\frac{CVL\log^2 V}{d^2}
\end{aligned}$$

  Therefore

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\frac{1}{N^2L^2V}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_{ij}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\Big(\boldsymbol{X}_i^\top\boldsymbol{X}_i-\frac{L}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\Big)^2|\mathsf{Z}_{\rm in}\Big]\\
&\quad+\mathbb{E}\Big[\Big(\frac{1}{N^2L^3V}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_{ij}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\Big)^2|\mathsf{Z}_{\rm in}\Big]\\
&\leq\frac{C\log^2 V}{NL^3Vd^2}.
\end{aligned}$$

- For the second summand, by `Event`,

$$\begin{aligned}
&\frac{1}{N^2LV^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_{ij}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\\
&\leq\frac{1}{N^2LV^2}\Big(\sum_{i=1}^{N}\sum_{j=1}^{N}|\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V}|\Big)\sup_{i,j}|\alpha_{ij}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V|\leq\frac{C\log V}{VL\sqrt{V}d^{3/2}}.
\end{aligned}$$

Therefore, by Chebyshev's inequality with probability $1-o_V(1)$, we have

$$|\texttt{score}_{1112}|\leq\frac{C\log^3 V}{NL^3Vd^2}+\frac{C\log^2 V}{VL\sqrt{V}d^{3/2}}. \tag{35}$$

By (34)-(35), overall we have

$$\texttt{score}_{111}\leq C\log V\Big(\frac{1}{VL^2\sqrt{d}}+\frac{1}{\sqrt{NV}L^{3/2}d}+\frac{1}{N\sqrt{VL}d\sqrt{L\wedge d}}+\frac{1}{NLd\sqrt{V}(L\wedge\sqrt{V})}\Big)$$

$$+ C\log V\Big(\frac{1}{N\sqrt{V}(L\wedge d)\sqrt{d}} + \frac{1}{N\sqrt{V}d^{\frac{3}{2}}} + \frac{1}{NL^2\sqrt{L\wedge d}}\Big) + \frac{C\log V}{V\sqrt{Ld}(L\wedge d)}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}$$
$$\leq C\log V\Big(\frac{1}{N\sqrt{V}(L\wedge d)\sqrt{d}} + \frac{1}{\sqrt{NV}L^{3/2}d} + \frac{1}{VL^{3/2}d(L\wedge d)} + \frac{1}{V^2\sqrt{Ld}(L\wedge d)}\Big)$$
$$+ \frac{C\log V}{VL^2(L\wedge d)^{1/2}}.$$

Concentration bound for $\texttt{score}_{12}$: We recall that

$$\texttt{score}_{12} = \frac{\delta s_{\nu,\delta}}{N^2L^2}\sum_{i,j=1}^{N}\alpha_{ij}\big(\boldsymbol{e}_1 - \frac{1}{L}\mathbb{1}_L\big)^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top\boldsymbol{Z}_{\text{out}}^\top\boldsymbol{Z}_{\text{out}}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)$$

In this part, we will focus on the term

$$\frac{1}{N^2L^2}\sum_{i,j=1}^{N}\alpha_{ij}\big(\boldsymbol{e}_1 - \frac{1}{L}\mathbb{1}_L\big)^\top \boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top\boldsymbol{Z}_{\text{out}}^\top\boldsymbol{Z}_{\text{out}}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)$$
$$= \frac{1}{N^2L^2}\sum_{i,j=1}^{N}\text{tr}(\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top\boldsymbol{Z}_{\text{out}}^\top\boldsymbol{Z}_{\text{out}}\alpha_{ij}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\boldsymbol{x}_i^\top)$$
$$- \frac{1}{N^2L^3}\sum_{i,j=1}^{N}\alpha_{ij}\text{tr}\Big(\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top\boldsymbol{Z}_{\text{out}}^\top\boldsymbol{Z}_{\text{out}}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\mathbb{1}_L^\top\boldsymbol{X}_i\Big)$$
$$= \texttt{score}_{121} + \texttt{score}_{122}$$

For the first term, we write

$$\texttt{score}_{121} = \frac{\phi'(0)^2}{N^2L^2}\sum_{i,j=1}^{N}\text{tr}(\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top\boldsymbol{Z}_{\text{out}}^\top\boldsymbol{Z}_{\text{out}}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\boldsymbol{x}_i^\top)$$
$$+ \frac{1}{N^2L^2}\sum_{i,j=1}^{N}\text{tr}\Big(\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top\boldsymbol{Z}_{\text{out}}^\top\boldsymbol{Z}_{\text{out}}(\alpha_{ij} - \phi'(0)^2)(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\boldsymbol{x}_i^\top\Big)$$
$$= \texttt{score}_{1211} + \texttt{score}_{1212}$$

We start with the second term. By Chebyshev's inequality, we have

$$\texttt{score}_{1212} = \frac{1}{N^2L^2}\sum_{i,j=1}^{N}(\alpha_{ij} - \phi'(0)^2)(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \frac{1}{V})\boldsymbol{x}_i^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L$$
$$\pm \frac{1}{N^2L^2\sqrt{d}}\Big\|\sum_{i,j=1}^{N}(\alpha_{ij} - \phi'(0)^2)(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\boldsymbol{x}_i^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top\Big\|_F$$

By $\texttt{Event}$

$$\sum_{i,j=1}^{N}(\alpha_{ij} - \phi'(0)^2)^2(\boldsymbol{x}_i^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L)^2$$
$$\leq \frac{1}{(V\wedge L^2\wedge L\sqrt{d})^2}\sum_{i\neq j=1}^{N}(\boldsymbol{x}_i^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L)^2 + \frac{1}{L^2}\sum_{i=1}^{N}(\boldsymbol{x}_i^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L)^2$$
$$\leq \Big(\frac{N^2}{(V\wedge L^2\wedge L\sqrt{d})^2} + \frac{N}{L^2}\Big)\Big(1 + \frac{L}{d}\Big)$$

Therefore,

$$\frac{1}{N^2L^2\sqrt{d}}\Big\|\sum_{i,j=1}^{N}(\alpha_{ij} - \phi'(0)^2)(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\boldsymbol{x}_i^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top\Big\|_F$$

$$\leq \frac{1}{N^2L^{3/2}\sqrt{d}(L\wedge d)^{1/2}}\Big(\frac{N}{V\wedge L^2\wedge L\sqrt{d}}+\frac{\sqrt{N}}{L}\Big)\Big\|\sum_{i=1}^N(\boldsymbol{x}_i-\frac{1}{V}\mathbb{1}_V)(\boldsymbol{x}_i-\frac{1}{V}\mathbb{1}_V)^\top\Big\|_2$$

$$\leq \frac{1}{NVL^{3/2}\sqrt{d}(L\wedge d)^{1/2}}\Big(\frac{N}{V\wedge L^2\wedge L\sqrt{d}}+\frac{\sqrt{N}}{L}\Big)=\frac{o_V(1)}{VL^2}.$$

Moreover,

$$\|\frac{1}{N^2L^2}\sum_{i,j=1}^N(\alpha_{ij}-\phi'(0)^2)(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})\boldsymbol{x}_i^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L\|$$

$$\leq\frac{1}{N^2L^2}\Big(\sum_{i,j=1}^N|\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V}|\Big)\sup_{i,j\in[N]}|\alpha_{ij}-\phi'(0)^2)\boldsymbol{x}_i^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L|\leq\frac{(1-\frac{1}{V})}{NL^2}\frac{1}{L}\Big(1+\sqrt{\frac{L}{d}}\Big).$$

Therefore, $|\texttt{score}_{1212}|\ll\frac{1}{VL^2}$. Next, we consider $\texttt{score}_{122}$:

$$\begin{aligned}&\texttt{score}_{122}\\&=\frac{\phi'(0)^2}{N^2L^3}\sum_{i,j=1}^N(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L\\&\pm\frac{\phi'(0)^2}{\sqrt{d}}\Big\|\frac{1}{N^2L^3}\sum_{i,j=1}^N(\boldsymbol{x}_i-\frac{1}{V}\mathbb{1}_V)\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)^\top\Big\|_F\\&+\frac{1}{N^2L^3}\sum_{i,j=1}^N(\alpha_{ij}-\phi'(0)^2)(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L\\&\pm\frac{1}{\sqrt{d}}\Big\|\frac{1}{N^2L^3}\sum_{i,j=1}^N(\alpha_{ij}-\phi'(0)^2)(\boldsymbol{x}_i-\frac{1}{V}\mathbb{1}_V)\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)^\top\Big\|_F\\&=:\texttt{score}_{1221}+\texttt{score}_{1222}+\texttt{score}_{1223}+\texttt{score}_{1224}.\end{aligned}$$

For $\texttt{score}_{1224}$, by $\texttt{Event}$,

$$\frac{1}{L^2}\sum_{i,j=1}^N(\alpha_{ij}-\phi'(0)^2)^2(\mathbb{1}_L^\top\boldsymbol{X}_i^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L)^2$$

$$\leq\frac{1}{(V\wedge L^2\wedge L\sqrt{d})^2}\frac{1}{L^2}\sum_{i\neq j=1}^N(\mathbb{1}_L^\top\boldsymbol{X}_i^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L)^2+\frac{1}{L^2}\frac{1}{L^2}\sum_{i=1}^N(\mathbb{1}_L^\top\boldsymbol{X}_i^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L)^2$$

$$\leq\frac{N^2}{(V\wedge L^2\wedge L\sqrt{d})^3}+\frac{N}{L^2}.$$

Therefore,

$$|\texttt{score}_{1224}|\leq\frac{1}{NVL^2\sqrt{d}}\Big(\frac{N}{(V\wedge L^2\wedge L\sqrt{d})^{3/2}}+\frac{\sqrt{N}}{L}\Big)\leq\frac{o_V(1)}{VL^2}.$$

For $\texttt{score}_{1223}$, by $\texttt{Event}$,

$$\begin{aligned}|\texttt{score}_{1223}|\leq\frac{1}{N^2L^3}\Big(\sum_{i\neq j=1}^N|\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V}|\Big)|\sup_{i\neq j}|(\alpha_{ij}-\phi'(0)^2)\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L|\\+\frac{1}{NL^3}\sup_i|(\alpha_{ii}-\phi'(0)^2)\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L|\leq\frac{o_V(1)}{VL^2}.\end{aligned}$$

For the first two terms, we define

$$\boldsymbol{V}_0:=\frac{1}{NL}\sum_{j=1}^N\boldsymbol{X}_j^\top\mathbb{1}_L(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)^\top,\qquad\boldsymbol{V}_{0,1}:=\frac{1}{NL}\sum_{i=1}^N\boldsymbol{x}_i(\boldsymbol{x}_i-\frac{1}{V}\mathbb{1}_V)^\top$$

$$\boldsymbol{V}_{0,2} \coloneqq \frac{1}{NL}\sum_{i=1}^{N}\big(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_{L-1}^\top\big)\mathbb{1}_{L-1}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)^\top,\ \boldsymbol{V}_{0,3} \coloneqq \frac{1}{V}\mathbb{1}_V\frac{1}{NL}\sum_{i=1}^{N}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)^\top.$$

We have by `Event`,

$$|\mathtt{score}_{1222}| \le \frac{\phi'(0)^2}{\sqrt{d}}\Big\|\frac{1}{L}\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{V}_0\boldsymbol{V}_0^\top\boldsymbol{Z}_{\mathrm{in}}^\top\Big\|_F \le \frac{\phi'(0)^2}{NVL^2\sqrt{d}}\Big\|\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{Z}_{\mathrm{in}}^\top\Big\|_F \le \frac{2\phi'(0)^2}{LdLN} \le \frac{o_V(1)}{VL^2}.$$

Lastly,

$$|\mathtt{score}_{1221}| = \frac{\phi'(0)^2}{L}\mathrm{tr}\Big(\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{V}_0\boldsymbol{V}_0^\top\boldsymbol{Z}_{\mathrm{in}}^\top\Big) \le \frac{\phi'(0)^2}{NVL^2\sqrt{d}}\mathrm{tr}\Big(\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{Z}_{\mathrm{in}}^\top\Big) \le \frac{2\phi'(0)^2}{NL^2\sqrt{d}} \le \frac{o_V(1)}{VL^2}.$$

Therefore, $|\mathtt{score}_{122}| \ll \frac{1}{VL^2}$. Lastly, we consider $\mathtt{score}_{121}$. By Chebyshev's inequality, we have

$$\begin{aligned}
&|\mathtt{score}_{121}|\\
&= \phi'(0)^2\mathrm{tr}(\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{V}_{0,1}) + \phi'(0)^2\mathrm{tr}(\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{V}_{0,2}) + \phi'(0)^2(L-1)\mathrm{tr}(\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{V}_{0,3})\\
&\pm \frac{1}{\sqrt{d}}\Big\|\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{V}_{0,1}\Big\|_F \pm \frac{1}{\sqrt{d}}\Big\|\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{V}_{0,2}\Big\|_F \pm \frac{L-1}{\sqrt{d}}\Big\|\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{V}_{0,3}\Big\|_F.
\end{aligned}$$

We have the following:

- The first summand:

$$\mathrm{tr}(\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{V}_{0,1}) = \mathrm{tr}(\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{V}_{0,1}\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\mathrm{in}}^\top) \asymp \frac{1}{VL^2}$$

- The second summand: By Chebyshev's inequality, we have

$$\begin{aligned}
\mathrm{tr}(\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{V}_{0,2}) &= \frac{1}{N^2L^2}\sum_{i,j=1}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \tfrac{1}{V})\boldsymbol{x}_i^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}(\boldsymbol{N}_j^\top - \tfrac{1}{V}\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\\
&= \pm\frac{1}{\sqrt{d}}\Big\|\boldsymbol{V}_{0,2}\boldsymbol{V}_{0,1}^\top\Big\|_F.
\end{aligned}$$

  We have by `Event`,

$$\Big\|\boldsymbol{V}_{0,2}\boldsymbol{V}_{0,1}^\top\Big\|_F \le \frac{C\log V}{VL}\Big\|\boldsymbol{V}_{0,2}\Big\|_F \le \frac{C\log V}{VL\sqrt{LN}}.$$

- The third summand: We have

$$\begin{aligned}
&(L-1)\mathrm{tr}(\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{V}_{0,2})\\
&\quad= \frac{L-1}{N^2L^2V}\sum_{i,j=1}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \tfrac{1}{V})\boldsymbol{x}_i^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\mathbb{1}_V\\
&\quad= \frac{L-1}{N^2L^2V}\sum_{i,j=1}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \tfrac{1}{V}) \pm \frac{L-1}{N^2L^2V\sqrt{d}}\Big\|\sum_{i,j=1}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \tfrac{1}{V})\boldsymbol{x}_i\Big\|_2.
\end{aligned}$$

  We have by `Event`

$$\Big|\frac{L-1}{N^2L^2V}\sum_{i,j=1}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \tfrac{1}{V})\Big| \le \frac{C\log V}{NLV^{3/2}}.$$

  and

$$\frac{L-1}{N^2L^2V\sqrt{d}}\Big\|\sum_{i,j=1}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \tfrac{1}{V})\boldsymbol{x}_i\Big\|_2 = \frac{L-1}{LV\sqrt{d}}\Big\|\boldsymbol{V}_{0,1}\frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)\Big\|_2 \le \frac{1}{\sqrt{N}V^2L\sqrt{d}}$$

- The fourth summand: We have by Event

$$\frac{1}{\sqrt{d}}\left\|\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{V}_{0,1}\right\|_F=\frac{1}{\sqrt{d}}\left\|\boldsymbol{Z}_{\text{in}}\boldsymbol{V}_{0,1}\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\text{in}}^\top\right\|_F\le\frac{C}{VL^2d}$$

- The fifth summand: We have by Event

$$\begin{aligned}\left\|\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{V}_{0,2}\right\|_F^2&\le\operatorname{tr}\Big(\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{V}_{0,2}\boldsymbol{V}_{0,2}^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{V}_{0,1}\Big)\\&\le\frac{1}{NLd}\operatorname{tr}(\boldsymbol{Z}_{\text{in}}\boldsymbol{V}_{0,1}\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\text{in}}^\top)\le\frac{V}{NLd}\frac{1}{V^2L^2}=\frac{1}{NVL^3d}.\end{aligned}$$

Therefore,

$$\frac{1}{\sqrt{d}}\left\|\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{V}_{0,2}\right\|_F\le\frac{1}{\sqrt{NV}L^{3/2}d}$$

- The sixth summand:

$$(L-1)^2\left\|\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{V}_{0,3}\right\|_F^2\le\frac{1}{V^2N}\mathbb{1}_V\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{V}_{0,1}\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\le\frac{1}{V^2NL^2d}$$

Therefore,

$$\frac{L-1}{\sqrt{d}}\left\|\boldsymbol{V}_{0,1}^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{V}_{0,3}\right\|_F\le\frac{C}{VL\sqrt{N}d}.$$

Then,

$$\texttt{score}_{121}=\frac{1\pm o_V(1)}{VL^2}\pm\frac{1}{\sqrt{NV}L^{3/2}d}.$$

Therefore, we have

$$\texttt{score}_{12}=\delta s_{\nu,\delta}\Big(\frac{1\pm o_V(1)}{VL^2}\pm\frac{1}{\sqrt{NV}L^{3/2}d}\Big).$$

C.3.2 Concentration bound for $\mathbf{s_2}$

We have

$$\begin{aligned}&\boldsymbol{e}_l^\top\boldsymbol{s}_2\\&=\frac{1}{N^2L^2}\sum_{i,j=1}^N\beta_{ij}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L(\boldsymbol{x}_i-\tfrac{1}{V}\mathbb{1}_V)^\top\boldsymbol{Z}_{\text{out}}^\top\boldsymbol{Z}_{\text{out}}(\boldsymbol{x}_j-\tfrac{1}{V}\mathbb{1}_V)\\&-\frac{1}{N^2L^3}\sum_{i,j=1}^N\beta_{ij}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L(\boldsymbol{x}_i-\tfrac{1}{V}\mathbb{1}_V)^\top\boldsymbol{Z}_{\text{out}}^\top\boldsymbol{Z}_{\text{out}}(\boldsymbol{x}_j-\tfrac{1}{V}\mathbb{1}_V)\\&+\frac{\delta s_{\nu,\delta}}{N^2L^2}\sum_{i,j=1}^N\beta_{ij}\big(\boldsymbol{e}_1-\frac{1}{L}\mathbb{1}_L\big)^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L(\boldsymbol{x}_i-\tfrac{1}{V}\mathbb{1}_V)^\top\boldsymbol{Z}_{\text{out}}^\top\boldsymbol{Z}_{\text{out}}(\boldsymbol{x}_j-\tfrac{1}{V}\mathbb{1}_V)\\&=:\texttt{score}_{21}+\texttt{score}_{22}+\texttt{score}_{23}.\end{aligned}$$

Concentration for $\texttt{score}_{21}$: We will write $\texttt{score}_{21}$ as follows:

$$\begin{aligned}\texttt{score}_{21}=\frac{1}{N^2L^2}&\sum_{i,j=1}^N(\beta_{ij}-\phi'(0)^2)\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\\&\qquad\times(\boldsymbol{x}_i-\tfrac{1}{V}\mathbb{1}_V)^\top\boldsymbol{Z}_{\text{out}}^\top\boldsymbol{Z}_{\text{out}}(\boldsymbol{x}_j-\tfrac{1}{V}\mathbb{1}_V)\\&+\frac{\phi'(0)^2}{N^2L^2}\sum_{i,j=1}^N\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top-\tfrac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{x}_i+\tfrac{L-1}{V}\mathbb{1}_V\big)\end{aligned}$$

$$\times (\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)^\top \boldsymbol{Z}_{\text{out}}^\top \boldsymbol{Z}_{\text{out}} (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)$$

$$+ \frac{\phi'(0)^2}{N^2L^2} \sum_{i,j=1}^N \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \Big(\boldsymbol{N}_i^\top \boldsymbol{N}_i - \tfrac{L-1}{V} \boldsymbol{I}_V \Big) \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \big(\boldsymbol{x}_i + \tfrac{L-1}{V}\mathbb{1}_V \big)$$

$$\times (\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)^\top \boldsymbol{Z}_{\text{out}}^\top \boldsymbol{Z}_{\text{out}} (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)$$

$$+ \frac{\phi'(0)^2}{N^2L^2} \sum_{i,j=1}^N \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \Big(\boldsymbol{x}_i \boldsymbol{x}_i^\top - \tfrac{1}{V} \boldsymbol{I}_V \Big) \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \big(\boldsymbol{N}_i^\top - \tfrac{1}{V}\mathbb{1}_V \mathbb{1}_{L-1}^\top \big) \mathbb{1}_{L-1}$$

$$\times (\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)^\top \boldsymbol{Z}_{\text{out}}^\top \boldsymbol{Z}_{\text{out}} (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)$$

$$+ \frac{\phi'(0)^2}{N^2L^2} \sum_{i,j=1}^N \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \Big(\boldsymbol{N}_i \boldsymbol{N}_i^\top - \tfrac{L-1}{V} \boldsymbol{I}_V \Big) \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \big(\boldsymbol{N}_i^\top - \tfrac{1}{V}\mathbb{1}_V \mathbb{1}_{L-1}^\top \big)$$

$$\times (\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)^\top \boldsymbol{Z}_{\text{out}}^\top \boldsymbol{Z}_{\text{out}} (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)$$

$$+ \frac{\phi'(0)^2}{N^2LV} \sum_{i,j=1}^N \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L (\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)^\top \boldsymbol{Z}_{\text{out}}^\top \boldsymbol{Z}_{\text{out}} (\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)$$

$$=: \texttt{score}_{211} + \texttt{score}_{212} + \texttt{score}_{213} + \texttt{score}_{214} + \texttt{score}_{215} + \texttt{score}_{216}$$

By Chebyshev's inequality, we have

$$\texttt{score}_{211} = \frac{1}{N^2L^2} \sum_{i,j=1}^N (\beta_{ij} - \phi'(0)^2)(\mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} - \tfrac{1}{V}) \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L$$

$$\pm \frac{1}{N^2L^2\sqrt{d}} \Big\| \sum_{i,j=1}^N (\beta_{ij} - \phi'(0)^2)(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V) \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L (\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)^\top \Big\|_F$$

$$=: \texttt{score}_{2111} + \texttt{score}_{2112}.$$

By using $\texttt{Events}$

$$|\texttt{score}_{2111}|$$

$$\le \frac{1}{N^2L^2} \Bigg(\sum_{i=1}^N \Big(\sum_{j=1}^N (\beta_{ij} - \phi'(0)^2)(\mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} - \tfrac{1}{V}) \Big)^2 \Bigg)^{\frac{1}{2}} \Big(\sum_{i=1}^N (\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L)^2 \Big)^{\frac{1}{2}}$$

$$\le \frac{C \log^2 V}{N^{3/2}L^2} \Big(\frac{\sqrt{N}}{L} + \frac{N^{3/2}}{V} \frac{1}{V \wedge L^2 \wedge L\sqrt{d}} \Big) \Big(\frac{L}{\sqrt{d}(L \wedge d)^{\frac{1}{2}}} \Big)$$
$$+ \frac{C \log^2 V}{N^{3/2}LV} \Big(\frac{\sqrt{N}}{L} + \frac{N^{3/2}}{V} \frac{1}{V \wedge L^2 \wedge L\sqrt{d}} \Big) \Big(\frac{\sqrt{L}V}{d^{3/2}} \Big)$$
$$\le \frac{C \log^2 V}{VL^2\sqrt{d}(L \wedge d)^{\frac{1}{2}}} + \frac{C \log^2 V}{NL^{3/2}d^{3/2}} + \frac{C \log^2 V}{V\sqrt{L}d^{3/2}} \frac{1}{V \wedge L^2 \wedge L\sqrt{d}}.$$

Moreover, by Chebyshev's inequality

$$|\texttt{score}_{2112}| \le \frac{1}{N^{3/2}L^2\sqrt{d}} \Big\| \sum_{i=1}^N (\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)(\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)^\top \Big\|_2$$
$$\times \Big(\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N (\beta_{ij} - \phi'(0)^2)^2 |\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L|^2 \Big)^{\frac{1}{2}}$$
$$\le \frac{C \log^2 V}{N^{3/2}L^2\sqrt{d}} \frac{N}{V} \Big(\frac{\sqrt{N}}{V \wedge L^2 \wedge L\sqrt{d}} + \frac{1}{L} \Big) \Big(\frac{\sqrt{L}}{\sqrt{d}} + \frac{L^{3/2}}{d^{3/2}} \Big)$$
$$= \frac{C \log^2 V}{\sqrt{NL}Vd(L \wedge d)} \Big(\frac{\sqrt{N}}{V \wedge L^2 \wedge L\sqrt{d}} + \frac{1}{L} \Big)$$

$$\leq \frac{C\log^2 V}{V\sqrt{L}d(L\wedge d)}\frac{1}{V\wedge L^2\wedge L\sqrt{d}} + \frac{C\log^2 V}{\sqrt{N}VL^{3/2}d(L\wedge d)}.$$

Therefore,

$$|\texttt{score}_{211}| \leq \frac{1}{VL^2\sqrt{d}} + \frac{C\log^2 V}{NL^{3/2}d^{3/2}} + C\log^2 V\Big(\frac{1}{V\sqrt{L}d(L\wedge d)} + \frac{1}{V\sqrt{L}d^{3/2}}\Big)\frac{1}{V\wedge L^2\wedge L\sqrt{d}}.$$

Moreover, by Chebyshev's inequality, we have

$$\begin{aligned}
&\texttt{score}_{212}\\
&= \frac{(1-\frac{1}{V})}{N^2L^2}\Big(\sum_{j=1}^N(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)\Big)^\top \sum_{i=1}^N(\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \tfrac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{x}_i + \tfrac{L-1}{V}\mathbb{1}_V\big)\\
&\pm \frac{1}{N^2L^2\sqrt{d}}\Big\|\sum_{j=1}^N(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)\Big\|_2\Big\|\sum_{i=1}^N(\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \tfrac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{x}_i + \tfrac{L-1}{V}\mathbb{1}_V\big)\Big\|_2.
\end{aligned}$$

Let $n_i \coloneqq |\{j \leq N | \boldsymbol{x}_j = \boldsymbol{e}_i\}|$. We have

$$\begin{aligned}
&\frac{1}{N^2L^2}\Big\|\sum_{j=1}^N(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)\Big\|_2\Big\|\sum_{i=1}^N(\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \tfrac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{x}_i + \tfrac{L-1}{V}\mathbb{1}_V\big)\Big\|_2\\
&\frac{1}{N^2L^2}\Big\|\sum_{j=1}^N(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)\Big\|_2\Big\|\sum_{i=1}^N(\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \tfrac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{x}_i + \frac{L-1}{V}\mathbb{1}_V\big)\Big\|_2\\
&+ \frac{L-1}{L}\frac{1}{N^2LV}\Big\|\sum_{j=1}^N(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)\Big\|_2\Big\|\sum_{i=1}^N(\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \tfrac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\Big\|_2\\
&\leq \frac{C}{NL^2}\Big(\frac{1}{N}\sum_{i=1}^V n_i^2\Big|\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{e}_i\boldsymbol{e}_i^\top - \tfrac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{e}_i\Big|^2\Big)^{\frac{1}{2}}\\
&+ \frac{C}{\sqrt{N}LV}\Big\|\frac{1}{N}\sum_{i=1}^N(\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \tfrac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\Big\|_2
\end{aligned}$$

By $\texttt{Event}$, we have

$$\Big|\frac{1}{N}\sum_{i=1}^V n_i^2\Big| \leq \frac{CN}{V} \quad\text{and}\quad \sup_{i\leq N}\Big|\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{e}_i\boldsymbol{e}_i^\top - \frac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{e}_i\Big| \leq \frac{C\log V}{\sqrt{d}}.$$

Moreover,

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^N(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \frac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\Big\|_2^2|\mathsf{Z}_{\text{in}}\Big]\\
&\leq \frac{1}{N}\mathbb{E}\Big[\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \frac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \frac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\text{in}}\Big]\\
&+ \frac{1}{N^2}\mathbb{E}\Big[\sum_{i\neq j=1}^N(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \frac{1}{V})\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \frac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_j\boldsymbol{x}_j^\top - \frac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\text{in}}\Big]\\
&\leq (\frac{1}{N} + \frac{N-1}{NV})\mathbb{E}\Big[\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \frac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \frac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\text{in}}\Big]
\end{aligned}$$

$$\leq \Big(\frac{1}{N} + \frac{N-1}{NV}\Big)\frac{CV\log V}{d^2},$$

where we used $\texttt{Events}$ in the last step. Therefore, by Chebyshev's inequality, we have

$$|\texttt{score}_{212}| \leq \frac{C\log V}{\sqrt{NV}L^2\sqrt{d}}\Big(1 + \frac{L}{\sqrt{Vd}}\Big).$$

Moreover, by using Chebyshev's inequality

$$\begin{aligned}
&\texttt{score}_{213}\\
&= \frac{1}{N^2L^2}\Big(\sum_{j=1}^N(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)\Big)^\top\\
&\times \Big(\sum_{i=1}^N(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{x}_i + \frac{L-1}{V}\mathbb{1}_V)\Big)\\
&\pm \frac{1}{N^2L^2\sqrt{d}}\Big\|\sum_{j=1}^N(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)\Big\|_2\\
&\times \Big\|\sum_{i=1}^N(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{x}_i + \frac{L-1}{V}\mathbb{1}_V)\Big\|_2
\end{aligned}$$

We have

$$\begin{aligned}
&\mathbb{E}\left[\Big(\sum_{i,j=1}^N(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{x}_i + \frac{L-1}{V}\mathbb{1}_V)\Big)^2\Big|\mathsf{Z}_{\text{in}}\right]\\
&= \sum_{i=1}^N\mathbb{E}\left[\Big(\sum_{j=1}^N(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{x}_i\Big)^2\Big|\mathsf{Z}_{\text{in}}\right]\\
&\leq 2(1-\frac{1}{V})^2\sum_{i=1}^N\mathbb{E}\left[\Big(\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i - \frac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{x}_i + \frac{L-1}{V}\mathbb{1}_V)\Big)^2\Big|\mathsf{Z}_{\text{in}}\right]\\
&+ \frac{2(1-\frac{1}{V})}{V}\sum_{i=1}^N\sum_{j\neq i}^N\mathbb{E}\left[\Big(\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i - \frac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{x}_i + \frac{L-1}{V}\mathbb{1}_V)\Big)^2\Big|\mathsf{Z}_{\text{in}}\right] \quad (36)
\end{aligned}$$

We have

$$\mathbb{E}\left[\Big(\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i - \frac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{x}_i + \frac{L-1}{V}\mathbb{1}_V)\Big)^2\Big|\mathsf{Z}_{\text{in}}\right] \leq \frac{CL}{d^2}\Big(1 + \frac{L^2}{V}\Big).$$

Therefore, we have (36) $\leq \frac{CN^2L}{Vd^2}$. Also,

$$\mathbb{E}\left[\Big\|\sum_{i=1}^N(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i - \frac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{x}_i + \frac{L-1}{V}\mathbb{1}_V)\Big\|_2^2\,\Big|\mathsf{Z}_{\text{in}}\right] \leq \frac{CL}{d^2}\Big(1 + \frac{L^2}{V}\Big).$$

Therefore, by $\texttt{Events}$ and Chevushev's inequality, we have

$$|\texttt{score}_{213}| \leq C\log V\Big(\frac{1}{N\sqrt{V}L^{3/2}d} + \frac{1}{NVLd} + \frac{1}{NL^{3/2}d^{3/2}} + \frac{1}{N\sqrt{V}\sqrt{L}d^{3/2}}\Big).$$

Moreover, by Chebyshev's inequality

$$\begin{aligned}
&\texttt{score}_{214} =\\
&\frac{1}{N^2L^2}\Big(\sum_{j=1}^N(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)\Big)^\top
\end{aligned}$$

$$\times \Big(\sum_{i=1}^{N}(\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \tfrac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{N}_i^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top\big)\mathbb{1}_{L-1}\Big)$$

$$\pm \frac{1}{N^2L^2\sqrt{d}}\Big\|\sum_{j=1}^{N}(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)\Big\|_2$$

$$\times \Big\|\sum_{i=1}^{N}(\boldsymbol{x}_i - \tfrac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \tfrac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{N}_i^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top\big)\mathbb{1}_{L-1}\Big\|_2.$$

We have

$$\begin{aligned}
&\mathbb{E}\left[\Big(\sum_{i,j=1}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i^\top\boldsymbol{x}_i - \tfrac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{N}_i^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top\big)\mathbb{1}_{L-1}\Big)^2 \,\big|\mathsf{Z}_{\text{in}}\right] \\
&= \sum_{i=1}^{N}\mathbb{E}\left[\Big(\sum_{j=1}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \tfrac{1}{V})\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \tfrac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{N}_i^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top\big)\mathbb{1}_{L-1}\Big)^2 \,\big|\mathsf{Z}_{\text{in}}\right] \\
&\le 2(1-\tfrac{1}{V})^2\sum_{i=1}^{N}\mathbb{E}\left[\Big(\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{N}_i^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top\big)\mathbb{1}_{L-1}\Big)^2 \,\big|\mathsf{Z}_{\text{in}}\right] \\
&+ \tfrac{2(1-\frac{1}{V})}{V}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\mathbb{E}\left[\Big(\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \tfrac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{N}_i^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top\big)\mathbb{1}_{L-1}\Big)^2 \,\big|\mathsf{Z}_{\text{in}}\right]
\end{aligned} \tag{37}$$

We have

$$\begin{aligned}
&\mathbb{E}\left[\Big(\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{N}_i^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top\big)\mathbb{1}_{L-1}\Big)^2 \,\big|\mathsf{Z}_{\text{in}}\right] \\
&\le \frac{CL}{d}\mathsf{z}_{\nu,\delta}^\top\,\mathbb{E}\left[\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \tfrac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \tfrac{1}{V}\boldsymbol{I}_V\Big)\,\big|\mathsf{Z}_{\text{in}}\right]\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta} \le \frac{CL}{d^2}.
\end{aligned}$$

Therefore, (37) $\le \frac{CN^2L}{Vd^2}$. Also,

$$\begin{aligned}
&\mathbb{E}\left[\Big\|\sum_{i=1}^{N}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{N}_i^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top\big)\mathbb{1}_{L-1}\Big\|_2^2 \,\big|\mathsf{Z}_{\text{in}}\right] \\
&= \frac{CL}{d}\sum_{i=1}^{N}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\,\mathbb{E}\left[\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \frac{1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{x}_i\boldsymbol{x}_i^\top - \frac{1}{V}\boldsymbol{I}_V\Big)\,\big|\mathsf{Z}_{\text{in}}\right]\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta} \le \frac{CNL}{d^2}.
\end{aligned}$$

Therefore, by `Events` and by Chebyshev's inequality, we have

$$|\mathtt{score}_{214}| \le C\log V\Big(\frac{1}{N\sqrt{V}L^{3/2}d} + \frac{1}{NL^{3/2}d^{3/2}}\Big)$$

Moreover, let

$$\begin{aligned}
\gamma_i &:= \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{N}_i^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top\big)\mathbb{1}_{L-1} \\
&\quad - \mathbb{E}\Big[\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{N}_i^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top\big)\mathbb{1}_{L-1}|\mathsf{Z}_{\text{in}}\Big].
\end{aligned}$$

We have

$$\begin{aligned}
&\mathtt{score}_{215} = \\
&\frac{1}{N^2L^2}\Big(\sum_{j=1}^{N}(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)\Big)^\top\Big(\sum_{i=1}^{N}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\gamma_i\Big) \\
&\pm \frac{1}{N^2L^2\sqrt{d}}\Big\|\sum_{j=1}^{N}(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)\Big\|_2\Big\|\sum_{i=1}^{N}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\gamma_i\Big\|_2
\end{aligned}$$

$$+\frac{1}{N^2L^2}\Big(\sum_{j=1}^{N}(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big)^\top$$
$$\times\Big(\sum_{i=1}^{N}(\boldsymbol{x}_i-\tfrac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\,\mathbb{E}\Big[\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i-\tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{N}_i^\top-\tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top\big)\mathbb{1}_{L-1}|\mathsf{Z}_{\text{in}}\Big]\Big)$$
$$\pm\frac{1}{N^2L^2\sqrt{d}}\Big\|\sum_{j=1}^{N}(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big\|_2$$
$$\times\Big\|\sum_{i=1}^{N}(\boldsymbol{x}_i-\frac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\,\mathbb{E}\Big[\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i-\tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{N}_i^\top-\tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top\big)\mathbb{1}_{L-1}|\mathsf{Z}_{\text{in}}\Big]\Big\|_2.$$

By Proposition 9

$$\mathbb{E}\Big[\Big(\sum_{i,j=1}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})\gamma_i\Big)^2\Big]=\sum_{i=1}^{N}\mathbb{E}\Big[\Big(\sum_{j=1}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})\gamma_i\Big)^2\Big]$$
$$\le 2(1-\frac{1}{V})^2\sum_{i=1}^{N}\mathbb{E}[\gamma_i^2]+\frac{2(1-\frac{1}{V})}{V}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\mathbb{E}[\gamma_i^2]\le C\log^2 V\frac{N^2}{V}\Big(\frac{L}{d}+\frac{L^2}{d^2}\Big).$$

Then,

$$\mathbb{E}\Big[\Big\|\sum_{i=1}^{N}(\boldsymbol{x}_i-\frac{1}{V}\mathbb{1}_V)\gamma_i\Big\|_2^2\Big]\le\sum_{i=1}^{N}\mathbb{E}[\gamma_i^2]\le CN\log^2 V\Big(\frac{L}{d}+\frac{L^2}{d^2}\Big).$$

Moreover, by Proposition 9 and `Event`, we have

$$\Big\|\sum_{i=1}^{N}(\boldsymbol{x}_i-\frac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\,\mathbb{E}\Big[\Big(\boldsymbol{N}_i^\top\boldsymbol{N}_i-\tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big(\boldsymbol{N}_i^\top-\tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top\big)\mathbb{1}_{L-1}|\mathsf{Z}_{\text{in}}\Big]\Big\|_2$$
$$\le C\log V\frac{L\sqrt{N}}{\sqrt{Vd}}.$$

Therefore, by Chebyshev's inequality, we have

$$|\texttt{score}_{215}|\le C\log^2 V\Big(\frac{1}{NL\sqrt{Vd}(L\wedge d)^{1/2}}+\frac{1}{NLd(L\wedge d)^{1/2}}+\frac{1}{NL\sqrt{Vd}}\Big)$$

Lastly, by Chebyshev's inequality, we have

$$\texttt{score}_{216}=\frac{\phi'(0)^2}{N^2LV}\sum_{i,j=1}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L$$
$$\pm\frac{\phi'(0)^2}{N^2LV\sqrt{d}}\|\sum_{i,j=1}^{N}(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L(\boldsymbol{x}_i-\frac{1}{V}\mathbb{1}_V)^\top\|_F$$
$$=\frac{\phi'(0)^2}{N^2LV}\sum_{i,j=1}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j}-\frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L$$
$$\pm\frac{\phi'(0)^2}{V\sqrt{d}}\|\frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\|_2\|\frac{1}{NL}\sum_{i=1}^{N}(\boldsymbol{x}_i-\frac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\|_2$$

We have

$$\Big\|\frac{1}{NL}\sum_{i=1}^{N}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L(\boldsymbol{x}_i-\frac{1}{V}\mathbb{1}_V)^\top\Big\|_2$$

$$\leq \Big\| \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \frac{1}{NL} \sum_{i=1}^N (\boldsymbol{X}_i^\top - \frac{1}{V} \mathbb{1}_V \mathbb{1}_L^\top) \mathbb{1}_L (\boldsymbol{x}_i - \frac{1}{V} \mathbb{1}_V)^\top \Big\|_2$$

$$+ \frac{1}{V} |\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \mathbb{1}_V| \Big\| \frac{1}{N} \sum_{i=1}^N (\boldsymbol{x}_i - \frac{1}{V} \mathbb{1}_V)^\top \Big\|_2$$

$$\leq \frac{CV}{d} \Big\| \boldsymbol{Z}_{\text{in}} \frac{1}{NL} \sum_{i=1}^N (\boldsymbol{X}_i^\top - \frac{1}{V} \mathbb{1}_V \mathbb{1}_L^\top) \mathbb{1}_L (\boldsymbol{x}_i - \frac{1}{V} \mathbb{1}_V)^\top \Big\|_2 + C \log V \frac{\sqrt{V}}{d^{3/2} \sqrt{N}}$$

$$\leq \frac{CV}{\sqrt{NL} d^{3/2}} + C \log V \frac{\sqrt{V}}{d^{3/2} \sqrt{N}} \leq \frac{CV}{\sqrt{NL} d^{3/2}}.$$

Moreover,

$$\mathbb{E}\Big[\Big(\frac{1}{N^2 LV} \sum_{i,j=1}^N (\mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} - \frac{1}{V}) \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L \Big)^2 | \mathsf{Z}_{\text{in}} \Big]$$

$$= \frac{1}{N^4 L^2 V^2} \sum_{j=1}^N \mathbb{E}\Big[\Big(\sum_{i=1}^N (\mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} - \frac{1}{V}) \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L \Big)^2 | \mathsf{Z}_{\text{in}} \Big]$$

$$\leq \frac{2}{N^4 L^2 V^2} \sum_{j=1}^N \mathbb{E}\Big[\Big(\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L \Big)^2 | \mathsf{Z}_{\text{in}} \Big]$$

$$+ \frac{2}{N^4 L^2 V^2} \sum_{j=1}^N \mathbb{E}\Big[\Big(\sum_{\substack{i=1 \\ i \neq j}}^N (\mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} - \frac{1}{V}) \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L \Big)^2 | \mathsf{Z}_{\text{in}} \Big]$$

$$\leq \frac{2}{N^4 L^2 V^2} \sum_{j=1}^N \mathbb{E}\Big[\Big(\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L \Big)^2 | \mathsf{Z}_{\text{in}} \Big]$$

$$+ \frac{2}{N^2 V^3} \sum_{j=1}^N \mathbb{E}\Big[\Big\| \frac{1}{NL} \sum_{\substack{i=1 \\ i \neq j}}^N \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L (\boldsymbol{x}_i - \frac{1}{V} \mathbb{1}_V)^\top \Big\|_2^2 | \mathsf{Z}_{\text{in}} \Big]$$

$$\leq \frac{2}{N^4 L^2 V^2} \sum_{j=1}^N \mathbb{E}\Big[\Big(\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L \Big)^2 | \mathsf{Z}_{\text{in}} \Big] + \frac{C}{N^2 V L d^3},$$

where we used C.3.2 in the last step. We have

$$\frac{1}{N^4 L^2 V^2} \sum_{j=1}^N \mathbb{E}\Big[\Big(\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L \Big)^2 | \mathsf{Z}_{\text{in}} \Big]$$

$$\leq \frac{1}{N^3 L^2 V^2} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \Big(\frac{L^2}{V^2} \mathbb{1}_V \mathbb{1}_V^\top + \frac{L}{V} \boldsymbol{I}_V \Big) \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}} \mathsf{z}_{\nu,\delta} \leq \frac{C \log^2 V}{N^3 L d^3}$$

Therefore, by Chebyshev's inequality, we have

$$|\texttt{score}_{216}| \leq C \log V \Big(\frac{1}{N\sqrt{L} d^2} + \frac{1}{N\sqrt{VL} d^{3/2}} \Big) \leq \frac{C \log V}{N \sqrt{L} d^2}$$

Overall, by using $N \ll VL$,

$$|\texttt{score}_{21}|$$
$$\leq C \log^2 V \Big(\frac{1}{N\sqrt{L} d (L \wedge d)} + \frac{1}{NLd(L \wedge d)^{1/2}} + \frac{1}{NL\sqrt{Vd}} + \frac{1}{\sqrt{N} VLd} + \frac{1}{\sqrt{NV} L^2 \sqrt{d}} \Big)$$
$$+ C \log^2 V \Big(\frac{1}{VL^2\sqrt{d}} \frac{1}{V^2 \sqrt{L} d^{3/2}} \Big).$$

C.3.3 CONCENTRATION BOUND FOR $\mathbf{s_3}$

We have

$$
\begin{aligned}
&\boldsymbol{e}_l^\top \boldsymbol{s}_3 \\
&= \frac{1}{N^2L}\sum_{i,j=1}^{N} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \\
&\quad \times \Big(\frac{1}{m}\sum_{k=1}^{m} \boldsymbol{w}_k \phi'\big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big)\phi\Big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top \mathbb{1}_L\Big) \\
&\qquad - \mathbb{E}\Big[\boldsymbol{w}_k \phi'\big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big)\phi\Big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top \mathbb{1}_L\Big)\Big]\Big)(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top \boldsymbol{Z}_{\text{out}}^\top \boldsymbol{Z}_{\text{out}}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V) \\
&- \frac{1}{N^2L^2}\sum_{i,j=1}^{N} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L \mathbb{1}_L^\top \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \\
&\quad \times \Big(\frac{1}{m}\sum_{k=1}^{m} \boldsymbol{w}_k \phi'\big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big)\phi\Big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top \mathbb{1}_L\Big) \\
&\qquad - \mathbb{E}\Big[\boldsymbol{w}_k \phi'\big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big)\phi\Big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top \mathbb{1}_L\Big)\Big]\Big)(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top \boldsymbol{Z}_{\text{out}}^\top \boldsymbol{Z}_{\text{out}}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V) \\
&+ \frac{\delta s_{\nu,\delta}}{N^2L}\sum_{i,j=1}^{N} (\boldsymbol{e}_1 - \frac{1}{L}\mathbb{1}_L)^\top \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \\
&\quad \times \Big(\frac{1}{m}\sum_{k=1}^{m} \boldsymbol{w}_k \phi'\big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big)\phi\Big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top \mathbb{1}_L\Big) \\
&- \qquad \mathbb{E}\Big[\boldsymbol{w}_k \phi'\big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big)\phi\Big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top \mathbb{1}_L\Big)\Big]\Big)(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top \boldsymbol{Z}_{\text{out}}^\top \boldsymbol{Z}_{\text{out}}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V) \\
&=: \texttt{score}_{31} + \texttt{score}_{32} + \texttt{score}_{33}.
\end{aligned}
$$

Concentration bound for $\texttt{score}_{31}$: We start with $\texttt{score}_{31}$. We have

$$
\begin{aligned}
\texttt{score}_{31k} &:= \operatorname{tr}\Big(\frac{1}{NL}\sum_{i=1}^{N}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{w}_k \phi'\big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big) \\
&\qquad \times \frac{1}{N}\sum_{j=1}^{N}\phi\Big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top \mathbb{1}_L\Big)(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top \boldsymbol{Z}_{\text{out}}^\top \boldsymbol{Z}_{\text{out}}\Big) \\
&= \operatorname{tr}\Big(\frac{1}{NL}\sum_{i=1}^{N}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{w}_k \phi'\big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big) \\
&\qquad \times \frac{1}{N}\sum_{j=1}^{N}\phi\Big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top \mathbb{1}_L\Big)(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top\Big) \\
&\quad \pm \frac{1}{\sqrt{d}}\Big\|\frac{1}{NL}\sum_{i=1}^{N}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{w}_k \phi'\big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big)\Big\|_2 \\
&\qquad \Big\|\frac{1}{N}\sum_{j=1}^{N}\phi\Big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top \mathbb{1}_L\Big)(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)\Big\|_2 \\
&=: \texttt{score}_{31k_1} + \texttt{score}_{31k_2},
\end{aligned}
$$

where we used Chebyshev's inequality for the second step. We define

$$\phi(t) =: \phi(0) + t\psi(t) \text{ and } \phi'(t) =: \phi(0) + t\psi_1(t) \text{ and } \psi(t) =: \psi(0) + t\psi_2(t).$$

and write

$$\begin{aligned}
\texttt{score}_{31k_1} &= \phi(0)\phi'(0)\,\mathrm{tr}\Big(\frac{1}{NL}\sum_{i=1}^{N}\boldsymbol{x}_i \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{w}_k \frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top\Big) \\
&+ \phi(0)\mathrm{tr}\Big(\frac{1}{NL^2}\sum_{i=1}^{N}\boldsymbol{x}_i \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{x}_i\boldsymbol{x}_i^\top \boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{w}_k\boldsymbol{w}_k^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top \mathbb{1}_L \psi_1\big(\tfrac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big)\frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top\Big) \\
&+ \phi(0)\mathrm{tr}\Big(\frac{1}{NL^2}\sum_{i=1}^{N}\boldsymbol{x}_i \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{N}_i^\top \boldsymbol{N}_i \boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{w}_k\boldsymbol{w}_k^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top \mathbb{1}_L \psi_1\big(\tfrac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big)\frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{x}_j - \tfrac{1}{V}\mathbb{1}_V)^\top\Big) \\
&+ \mathrm{tr}\Big(\frac{1}{NL}\sum_{i=1}^{N}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{w}_k \phi'\big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_i^\top \mathbb{1}_L\big) \\
&\qquad \times \frac{1}{N}\sum_{j=1}^{N}\psi\Big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_j^\top \mathbb{1}_L\Big)\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{X}_j^\top \mathbb{1}_L(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top\Big) \\
&=: \texttt{score}_{31k_{11}} + \texttt{score}_{31k_{12}} + \texttt{score}_{31k_{13}} + \texttt{score}_{31k_{14}}.
\end{aligned}$$

In the following, we bound each term separately.

- We have

$$\begin{aligned}
\texttt{score}_{31k_{11}} &= \frac{1}{L}\sum_{w=1}^{V}(\frac{n_w}{N} - \frac{1}{V})\frac{n_w}{N}\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\Big(\boldsymbol{e}_w\boldsymbol{e}_w^\top + \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{w}_k \\
&\quad + \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\frac{1}{NL}\sum_{w=1}^{V}(\frac{n_w}{N} - \frac{1}{V})\sum_{i\in\{i_1,\cdots,i_{n_w}\}}\Big(\boldsymbol{N}_i^\top \boldsymbol{N}_i - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{w}_k
\end{aligned}$$

  We have

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\frac{1}{NL}\sum_{w=1}^{V}(\frac{n_w}{N} - \frac{1}{V})\sum_{i\in\{i_1,\cdots,i_{n_w}\}}\Big(\boldsymbol{N}_i^\top \boldsymbol{N}_i - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{w}_k\Big)^2|\mathsf{Z}_{\mathrm{in}}\Big] \\
&= \mathbb{E}\Big[\Big\|\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\frac{1}{NL}\sum_{w=1}^{V}(\frac{n_w}{N} - \frac{1}{V})\sum_{i\in\{i_1,\cdots,i_{n_w}\}}\Big(\boldsymbol{N}_i^\top \boldsymbol{N}_i - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\mathrm{in}}^\top\Big\|_2^2|\mathsf{Z}_{\mathrm{in}}\Big] \\
&= \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\frac{1}{N^2L^2}\sum_{w=1}^{V}\mathbb{E}\Big[(\frac{n_w}{N} - \frac{1}{V})^2 n_w\Big]\,\mathbb{E}\Big[\Big(\boldsymbol{N}_1^\top \boldsymbol{N}_1 - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{Z}_{\mathrm{in}}\Big(\boldsymbol{N}_1^\top \boldsymbol{N}_1 - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)|\mathsf{Z}_{\mathrm{in}}\Big]\boldsymbol{Z}_{\mathrm{in}}^\top \mathsf{z}_{\nu,\delta} \\
&\le \frac{C}{N^2L^2V}\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\,\mathbb{E}\Big[\Big(\boldsymbol{N}_1^\top \boldsymbol{N}_1 - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{Z}_{\mathrm{in}}\Big(\boldsymbol{N}_1^\top \boldsymbol{N}_1 - \tfrac{L-1}{V}\boldsymbol{I}_V\Big)|\mathsf{Z}_{\mathrm{in}}\Big]\boldsymbol{Z}_{\mathrm{in}}^\top \mathsf{z}_{\nu,\delta} = \frac{C}{N^2Vd(L\wedge d)}
\end{aligned}$$

  Moreover, by using `Event`, we write

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\frac{1}{L}\sum_{w=1}^{V}(\frac{n_w}{N} - \frac{1}{V})\frac{n_w}{N}\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\Big(\boldsymbol{e}_w\boldsymbol{e}_w^\top + \tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{w}_k\Big)^2|\mathsf{Z}_{\mathrm{in}}\Big] \\
&= \frac{1}{L^2}\mathbb{E}\Big[\Big\|\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\Big(\sum_{w=1}^{V}(\frac{n_w}{N} - \frac{1}{V})\frac{n_w}{N}\boldsymbol{e}_w\boldsymbol{e}_w^\top + (\frac{n_w}{N} - \frac{1}{V})^2\tfrac{L-1}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\mathrm{in}}^\top\Big\|_2^2|\mathsf{Z}_{\mathrm{in}}\Big]
\end{aligned}$$

$$\leq \frac{V^2}{L^2d^2}\,\mathbb{E}\Big[\Big\|\sum_{w=1}^{V}(\frac{n_w}{N}-\frac{1}{V})\frac{n_w}{N}\boldsymbol{e}_w\boldsymbol{e}_w^\top+(\frac{n_w}{N}-\frac{1}{V})^2\frac{L-1}{V}\boldsymbol{I}_V\Big\|_2^2|\mathsf{Z}_{\text{in}}\Big]$$

$$\leq \frac{CV^2}{L^2d^2}\,\mathbb{E}\Big[\sup_{w\in[N]}\Big|(\frac{n_w}{N}-\frac{1}{V})\frac{n_w}{N}\Big|^2\Big]+\frac{C}{d^2}\,\mathbb{E}\Big[\Big(\sum_{w=1}^{V}(\frac{n_w}{N}-\frac{1}{V})^2\Big)^2\Big]\leq\frac{C}{d^2N^2}.$$

Therefore,

$$\mathbb{E}\left[\texttt{score}^2_{31k_{11}}|\mathsf{Z}_{\text{in}}\right]\leq\frac{C}{d^2N^2}.$$

- Moreover,

$$\texttt{score}^2_{31k_{12}}\leq\frac{C}{N}\Big\|\frac{1}{NL^2}\sum_{i=1}^{N}\boldsymbol{x}_i\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{w}_k\boldsymbol{w}_k^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\psi_1\big(\frac{1}{L}\boldsymbol{w}_k^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\big)\Big\|_2^2.$$

We have for any $i\in[N]$, by $\texttt{Event}$,

$$\Big|\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{w}_k\boldsymbol{w}_k^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\psi_1\big(\frac{1}{L}\boldsymbol{w}_k^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\big)\Big|\leq C(\log^4V)\sqrt{L}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{e}_\nu}+\frac{1}{\sqrt{d}})$$

Then,

$$\Big\|\frac{1}{NL^2}\sum_{i=1}^{N}\boldsymbol{x}_i\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{w}_k\boldsymbol{w}_k^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\psi_1\big(\frac{1}{L}\boldsymbol{w}_k^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\big)\Big\|_2^2$$

$$\leq\frac{C(\log^8V)}{L^3}\Big\|\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{e}_\nu}+\frac{1}{\sqrt{d}})\Big\|^2\leq\frac{C(\log^8V)}{VdL^3}$$

Then,

$$\mathbb{E}[\texttt{score}^2_{31k_{12}}|\mathsf{Z}_{\text{in}}]\leq\frac{C\log^8V}{NVdL^3}.$$

- Moreover,

$$\texttt{score}^2_{31k_{13}}\leq\frac{C}{N}\Big\|\frac{1}{NL^2}\sum_{i=1}^{N}\boldsymbol{x}_i\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{N}_i^\top\boldsymbol{N}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{w}_k\boldsymbol{w}_k^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\psi_1\big(\frac{1}{L}\boldsymbol{w}_k^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\big)\Big\|_2^2.$$

We have for any $i\in[N]$, by $\texttt{Events}$

$$\Big|\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{N}_i^\top\boldsymbol{N}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{w}_k\boldsymbol{w}_k^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\psi_1\big(\frac{1}{L}\boldsymbol{w}_k^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\big)\Big|$$

$$\leq C(\log^8V)\sqrt{L}\|\boldsymbol{Z}_{\text{in}}\boldsymbol{N}_i^\top\boldsymbol{N}_i\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}\|_2\leq C(\log^8V)\sqrt{L}\big(\boldsymbol{e}_\nu^\top\boldsymbol{N}_i^\top\mathbb{1}_{L-1}+\frac{L}{d}+\log^6V\sqrt{\frac{L}{d}}\big)$$

Then, by $\texttt{Events}$

$$\Big\|\frac{1}{NL^2}\sum_{i=1}^{N}\boldsymbol{x}_i\mathsf{z}_{\kappa*}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{N}_i^\top\boldsymbol{N}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{w}_k\boldsymbol{w}_k^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\psi_1\big(\frac{1}{L}\boldsymbol{w}_k^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\big)\Big\|_2^2$$

$$\leq\frac{C(\log V)^2}{N^2L^3}\Big\|\sum_{i=1}^{N}\boldsymbol{x}_i\Big(\mathbb{1}_{L-1}^\top\boldsymbol{N}_i\boldsymbol{e}_\nu+\frac{L}{d}+\log^6V\sqrt{\frac{L}{d}}\Big)\Big\|_2^2$$

$$\leq\frac{C\log^{14}V}{VLd(L\wedge d)}+\frac{C(\log V)^2}{N^2L^3}\Big\|\sum_{i=1}^{N}\boldsymbol{x}_i\mathbb{1}_{L-1}^\top\boldsymbol{N}_i\boldsymbol{e}_\nu\Big\|_2^2$$

We have

$$\frac{C(\log V)^2}{N^2L^3}\,\mathbb{E}[\Big\|\sum_{i=1}^{N}\boldsymbol{x}_i\mathbb{1}_{L-1}^\top\boldsymbol{N}_i\boldsymbol{e}_\nu\Big\|_2^2]\leq\frac{C(\log V)^2}{N^2L^3}\Big(\frac{N^2}{V}\frac{L^2}{V^2}+N\frac{L}{V}\Big)$$

$$= \frac{C(\log V)^2}{V^3 L} + \frac{C(\log V)^2}{NVL^2}.$$

Then,

$$\mathbb{E}[\texttt{score}_{31k_{13}}^2 | \mathsf{Z}_{\text{in}}] \leq \frac{C\log^{20} V}{NVLd(L \wedge d)}.$$

- Lastly, we have

$$|\texttt{score}_{31k_{14}}| \leq \Big\| \frac{1}{NL} \sum_{i=1}^N (\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V) \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{w}_k \phi'\big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L\big) \Big\|_2$$
$$\times \Big\| \frac{1}{N} \sum_{j=1}^N \psi\Big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L\Big) \frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L (\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top \Big\|_2.$$

By using the derivations in the two previous items, we have

$$\Big\| \frac{1}{NL} \sum_{i=1}^N (\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)^\top \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \boldsymbol{X}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{w}_k \phi'\big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L\big) \Big\|_2$$
$$\leq \Big\| \frac{1}{NL} \sum_{i=1}^N \boldsymbol{x}_i \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{w}_k \phi'\big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L\big) \Big\|_2$$
$$+ \Big\| \frac{1}{NL^2} \sum_{i=1}^N \boldsymbol{x}_i \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{N}_i^\top \boldsymbol{N}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{w}_k \boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L \psi_1\big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L\big) \Big\|_2$$
$$+ |\phi'(0)| \Big\| \frac{1}{NL} \sum_{i=1}^N (\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V) \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{N}_i^\top \boldsymbol{N}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{w}_k \Big\|_2$$
$$\leq \frac{C\log^7 V}{\sqrt{VLd(L\wedge d)}} + \phi'(0) \Big\| \frac{1}{NL} \sum_{i=1}^N (\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V) \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{N}_i^\top \boldsymbol{N}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{w}_k \Big\|_2$$

We have

$$\mathbb{E}\Big[\Big\| \frac{1}{NL} \sum_{i=1}^N (\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V) \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{N}_i^\top \boldsymbol{N}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{w}_k \Big\|_2^2 \Big| \mathsf{Z}_{\text{in}} \Big]$$
$$\leq \frac{1}{N^2L^2} \mathbb{E}\Big[\sum_{i,j=1}^N (\mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} - \frac{1}{V}) \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{N}_i^\top \boldsymbol{N}_i \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{N}_i^\top \boldsymbol{N}_i \boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta} \Big| \mathsf{Z}_{\text{in}} \Big]$$
$$\leq \frac{(1-\frac{1}{V})}{NL^2}\frac{L}{V} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \mathrm{Diag}(\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}) \boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta} + \frac{(1-\frac{1}{V})}{NL^2}\frac{L^2}{V^2} \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}$$
$$= \frac{(1-\frac{1}{V})}{Nd(L\wedge d)} \tag{38}$$

Moreover,

$$\Big\| \frac{1}{NL} \sum_{j=1}^N \psi\Big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L\Big) \boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L (\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top \Big\|_2$$
$$= |\psi(0)| \Big\| \frac{1}{NL} \sum_{j=1}^N \boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L (\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top \Big\|_2$$
$$+ \Big\| \frac{1}{NL^2} \sum_{j=1}^N \boldsymbol{x}_j \psi_2\Big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L\Big) (\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L)^2 \Big\|_2.$$

By `Event`, for all $j \in [N]$,

$$|\psi_2\Big(\frac{1}{L}\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L\Big) (\boldsymbol{w}_k^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_j^\top \mathbb{1}_L)^2| \leq C(\log^6 V) L.$$

Therefore,

$$\Big\|\frac{1}{NL^2}\sum_{j=1}^{N}\boldsymbol{x}_j\psi_2\Big(\frac{1}{L}\boldsymbol{w}_k^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_j^\top\mathbb{1}_L\Big)(\boldsymbol{w}_k^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_j^\top\mathbb{1}_L)^2\Big\|_2\leq\frac{C(\log^6 V)}{\sqrt{V}L}.$$

Moreover,

$$\begin{aligned}&\Big\|\frac{1}{NL}\sum_{j=1}^{N}\boldsymbol{w}_k^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_j^\top\mathbb{1}_L(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)^\top\Big\|_2\\&\leq\Big\|\frac{1}{NL}\sum_{j=1}^{N}\boldsymbol{w}_k^\top\boldsymbol{Z}_{\rm in}(\boldsymbol{X}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)^\top\Big\|_2\\&+\frac{1}{V}|\boldsymbol{w}_k^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V|\Big\|\frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)^\top\Big\|_2.\end{aligned}$$

By using `Event`, we have

- $\frac{1}{V}|\boldsymbol{w}_k^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V|\Big\|\frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)^\top\Big\|_2\leq\frac{C\log^2 V}{\sqrt{V}N}$
- Moreover,

$$\begin{aligned}&\Big\|\frac{1}{NL}\sum_{j=1}^{N}\boldsymbol{w}_k^\top\boldsymbol{Z}_{\rm in}(\boldsymbol{X}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)^\top\Big\|_2^2\\&=\boldsymbol{w}_k^\top\boldsymbol{Z}_{\rm in}\Big(\frac{1}{NL}\sum_{j=1}^{N}(\boldsymbol{X}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)^\top\Big)\\&\qquad\Big(\frac{1}{NL}\sum_{j=1}^{N}(\boldsymbol{X}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)^\top\Big)^\top\boldsymbol{Z}_{\rm in}^\top\boldsymbol{w}_k\\&\leq\frac{C\log^2 V}{NL}.\end{aligned}\tag{39}$$

Then, for $N\ll VL$

$$\mathbb{E}[\texttt{score}_{31k_{14}}^2|\mathsf{Z}_{\rm in}]\leq\frac{C\log^{16}V}{N^2Ld(L\wedge d)}$$

- On the other hand, we have

$$\begin{aligned}|\texttt{score}_{31k_2}|\leq\frac{1}{\sqrt{d}}&\Big\|\frac{1}{NL}\sum_{i=1}^{N}(\boldsymbol{x}_i-\frac{1}{V}\mathbb{1}_V)\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\rm in}^\top\boldsymbol{w}_k\phi'\big(\frac{1}{L}\boldsymbol{w}_k^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\mathbb{1}_L\big)\Big\|_2\\&\times\Big\|\frac{1}{N}\sum_{j=1}^{N}\phi\Big(\frac{1}{L}\boldsymbol{w}_k^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_j^\top\mathbb{1}_L\Big)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big\|_2\end{aligned}$$

Note that by `Event` and (39), we have

$$\begin{aligned}&\Big\|\frac{1}{N}\sum_{j=1}^{N}\phi\Big(\frac{1}{L}\boldsymbol{w}_k^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_j^\top\mathbb{1}_L\Big)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big\|_2=|\phi(0)|\Big\|\frac{1}{N}\sum_{j=1}^{N}(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big\|_2\\&+\Big\|\frac{1}{N}\sum_{j=1}^{N}\psi\Big(\frac{1}{L}\boldsymbol{w}_k^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_j^\top\mathbb{1}_L\Big)\frac{1}{L}\boldsymbol{w}_k^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_j^\top\mathbb{1}_L(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big\|_2\\&\leq\frac{1}{\sqrt{N}}+\frac{C\log^2 V}{\sqrt{NL}}\leq\frac{C}{\sqrt{N}}\end{aligned}$$

Therefore by (38), we have

$$\mathbb{E}\Big[|\texttt{score}_{31k_2}|^2|\mathsf{Z}_{\rm in}\Big]\leq\frac{C}{N^2d^2(L\wedge d)}$$

Therefore, we have

$$\mathbb{E}[\mathtt{score}_{31}|\mathsf{Z}_{\text{in}}] = 0 \text{ and } \text{Variance}(\mathtt{score}_{31}|\mathsf{Z}_{\text{in}}) \le \frac{C}{N^2 d^2 m}.$$

D LOWER BOUND

To prove a lower bound, we construct a Bayesian setting with the same likelihood distribution in our setting. In particular, the ground truth permutation is chosen from the set of permutation matrices:

$$\mathcal{H} \coloneqq \{\boldsymbol{P} \in \{0,1\}^{V\times V} \mid \boldsymbol{\Pi} \text{ is a permutation matrix}\}.$$

We describe our Bayesian setting as a game between `Environment` and `Learner` as follows:

- At the beginning, `Environment` samples $\boldsymbol{P}_* \sim \text{Unif}(\mathcal{H})$, probability vectors without revealing them to the learner.
- `Learner` observes $L+1$ channel that generates words from the set $\mathcal{V} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_V\}$ sequentially for $t = 1, 2, \cdots, N$ with distributions:
  - At every round, `Environment` randomly picks a channel ℓ_t
  - *Label:* Channel 0 generates $\boldsymbol{p}_t \sim_{iid} \text{Unif}(\mathcal{V})$
  - *Input:* Given ℓ_t and $\boldsymbol{p}_t$, Channel ℓ_t generates $\boldsymbol{X}_{\ell_t,t} = \boldsymbol{P}_*\boldsymbol{p}_t$
  - *Noise distribution:* Channel $j \in [L]\setminus\{\ell_t\}$ generate $\boldsymbol{X}_{j,t} \sim \text{Unif}(\mathcal{V})$ independent of Channel 0.
- Let $\mathcal{D} \coloneqq \{(\boldsymbol{X}_t, \boldsymbol{p}_t)\}_{t\le N}$ be the dataset. We study the Bayes estimator with $0-1$ loss given the representation of the past: $S = f(\mathcal{D}, \ell_{1:N})$:

$$\hat{\boldsymbol{P}} = \underset{\boldsymbol{P}\in\mathcal{H}}{\arg\max}\, \mathbb{P}[\boldsymbol{P} = \boldsymbol{P}_*|S, \mathsf{Z}_{\text{in}}]. \tag{40}$$

  In the following we consider the empirical mean and covariance of embedded words as the given data, i.e., $S \coloneqq \{(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t, \boldsymbol{p}_t)\}_{t\le N}$, where

$$\boldsymbol{\mu}_t \coloneqq \frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_t^\top \mathbb{1}_L + \frac{\sigma_{\boldsymbol{\mu}}}{\sqrt{L}}\boldsymbol{g}_t \text{ and } \boldsymbol{\Sigma}_t \coloneqq \frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_t^\top\boldsymbol{X}_t\boldsymbol{Z}_{\text{in}}^\top + \frac{\sigma_{\boldsymbol{\Sigma}}}{\sqrt{dL}}\boldsymbol{G}_t.$$

  where $\{(\boldsymbol{g}_t, \boldsymbol{G}_t)\}_{t\le N}$ are i.i.d. measurement noise with distributions $\boldsymbol{g}_t \sim \mathcal{N}(0, \frac{1}{d}\boldsymbol{I}_d)$ and $\boldsymbol{G}_{t,ij} = \boldsymbol{G}_{t,ji}$ with $\boldsymbol{G}_{t,ij} \sim \mathcal{N}\big(0, \frac{(1+\delta_{ij})}{d}\big)$ i.i.d. for $i<j$.

Theorem 4. *The following lower bound holds:*

$$\mathbb{P}[\hat{\boldsymbol{P}} \ne \boldsymbol{P}_*|\mathsf{Z}_{\text{in}}] \ge 1 - o_V(1) - \frac{\Omega(N)}{V}\left(1 \wedge \left(\frac{1}{\sigma_{\boldsymbol{\mu}}^2}\frac{d}{L\log V} + \frac{C}{\sigma_{\boldsymbol{\Sigma}}^2}\frac{d^2}{L\log V}\right)\right)$$

We use an information-theoretic argument to prove Theorem 4. For the proof, let $H(A)$ and $H(A|C)$ denote the entropy and conditional entropy of A given C; let $I(A;B) = H(A) - H(A|B)$ and $I(A;B|C) = H(A|C) - H(A|B,C)$ denote the mutual information between random variables A and B and the conditional mutual given C, respectively. We let D_{KL} denote the Kullback-Leibler (KL) divergence. We start with an auxiliary statement for the proof.

Lemma 1. *Let A, B, C, D be discrete random variables defined on the same probability space. The following statements hold:*

- *In general, $H(A|B,C) \le H(A|B)$. The equality is satisfied if and only if $A \perp\!\!\!\perp C|B$.*
- *If $B \perp\!\!\!\perp D \mid (A,C)$, we have $I(A,B|C,D) \le I(A,B|C)$.*
- *Let $S = g(A,C)$ be a measurable function of (A,C). If $B \perp\!\!\!\perp A|(S,C,D)$, then $I(A;B|C,D) = I(S;B|C,D)$.*
- *Given, $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \mathbb{R}^d$, positive definite $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ and $\text{supp}(A) \subseteq \mathbb{R}^d$, we have*

$$D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu} + A, \boldsymbol{\Sigma})||\mathcal{N}(\boldsymbol{\mu}' + A, \boldsymbol{\Sigma})) \le \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}')^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}').$$

Proof. We have

$$H(A|B) - H(A|B,C) = \mathbb{E}\left[\log\frac{\mathbb{P}(A|B,C)}{\mathbb{P}(A|B)}\right] = \mathbb{E}\left[\log\frac{\mathbb{P}(A,C|B)}{\mathbb{P}(A|B)\mathbb{P}(C|B)}\right] = I(A,C|B).$$

Since the mutual information is non-negative, the first item follows. Moreover, since $I(A,C|B) = 0$ if and only if $A \perp\!\!\!\perp C|B$. For the second item, by using the first item,

$$I(A,B|C,D) = H(B|C,D) - H(B|A,C,D) \leq H(B|C) - H(B|A,C) = I(A,B|C).$$

For the third item, since S is a function of (A,C), we have

$$\begin{aligned} I(A;B|C,D) = I((A,S);B|C,D) &= H(B|C,D) - H(B|A,S,C,D) \\ &= H(B|C,D) - H(B|S,C,D) = I(S;B|C,D). \end{aligned}$$

Let f denotes the Gaussian pdf with 0 and covariance $\mathbf{\Sigma}$. For any $\boldsymbol{x} \in \mathbb{R}^d$, since $t \to t\log t$ is convex

$$\begin{aligned} \Big(\sum_{\boldsymbol{a}\in\text{supp}(A)} p(\boldsymbol{a}) f(\boldsymbol{x}-\boldsymbol{\mu}-\boldsymbol{a})\Big) \log\frac{\Big(\sum_{\boldsymbol{a}\in\text{supp}(A)} p(\boldsymbol{a}) f(\boldsymbol{x}-\boldsymbol{\mu}-\boldsymbol{a})\Big)}{\Big(\sum_{\boldsymbol{a}\in\text{supp}(A)} p(a) f(\boldsymbol{x}-\boldsymbol{\mu}'-\boldsymbol{a})\Big)} \\ \leq \sum_{\boldsymbol{a}\in\text{supp}(A)} p(\boldsymbol{a}) f(\boldsymbol{x}-\boldsymbol{\mu}-\boldsymbol{a}) \log\frac{f(\boldsymbol{x}-\boldsymbol{\mu}-\boldsymbol{a})}{f(\boldsymbol{x}-\boldsymbol{\mu}'-\boldsymbol{a})}. \end{aligned}$$

Therefore, we have

$$\begin{aligned} D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}+A,\mathbf{\Sigma})||\mathcal{N}(\boldsymbol{\mu}'+A,\mathbf{\Sigma})) &\leq \sum_{\boldsymbol{a}\in\text{supp}(A)} p(\boldsymbol{a}) D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu}+\boldsymbol{a},\mathbf{\Sigma})||\mathcal{N}(\boldsymbol{\mu}'+\boldsymbol{a},\mathbf{\Sigma})) \\ &= D_{\text{KL}}(\mathcal{N}(\boldsymbol{\mu},\mathbf{\Sigma})||\mathcal{N}(\boldsymbol{\mu}',\mathbf{\Sigma})), \end{aligned}$$

where the last inequality follows the invariance of KL divergence in the second line to constant shifts. The final bound follows the known formula for the KL divergence between Gaussian distributions. □

The proof of Theorem 4 is given in the following:

Proof of Theorem 4. Since we assume Z_{in} is known by the learner, we will fix it in the following without explicitly conditioning thte terms on it. Note that we consider the Bayes decision rule in (40) and use Fano's inequality (Scarlett & Cevher, 2019) to lower bound its error probability:

$$\mathbb{P}[\hat{\boldsymbol{P}} \neq \boldsymbol{P}_*|\mathsf{Z}_{\text{in}}] \geq 1 - \frac{I(\boldsymbol{P}_*;S) + \log 2}{\log|\mathcal{H}|}. \tag{41}$$

We have

$$\begin{aligned} I(\boldsymbol{P}_*;S) = I(\boldsymbol{P}_*;\{(\boldsymbol{\mu}_t,\mathbf{\Sigma}_t,\boldsymbol{p}_t)\}_{t\leq N}) &= I(\boldsymbol{P}_*;\{\boldsymbol{p}_t\}_{t\leq N}) + I(\boldsymbol{P}_*;\{(\boldsymbol{\mu}_t,\mathbf{\Sigma}_t)\}_{t\leq N}|\{\boldsymbol{p}_t\}_{t\leq N},) \\ &\overset{(a)}{=} I(\boldsymbol{P}_*;\{(\boldsymbol{\mu}_t,\mathbf{\Sigma}_t)\}_{t\leq N}|\{\boldsymbol{p}_t\}_{t\leq N}) \\ &= \sum_{t=1}^{N} I(\boldsymbol{P}_*;(\boldsymbol{\mu}_t,\mathbf{\Sigma}_t)|\{(\boldsymbol{\mu}_u,\mathbf{\Sigma}_u)\}_{u<t},\{\boldsymbol{p}_t\}_{t\leq N}) \end{aligned}$$

Given fixed Z_{in}, we observe that $(\boldsymbol{\mu}_t,\mathbf{\Sigma}_t) \perp\!\!\!\perp \{(\boldsymbol{\mu}_u,\mathbf{\Sigma}_u)\}_{u<t} \mid \boldsymbol{P}_*,\{\boldsymbol{p}_t\}_{t\leq N}$ and $(\boldsymbol{\mu}_t,\mathbf{\Sigma}_t) \perp\!\!\!\perp \{\boldsymbol{p}_u\}_{u\neq t}|\boldsymbol{P}_*$,.Therefore, by Lemma 1,

$$I(\boldsymbol{P}_*;S) \leq \sum_{t=1}^{N} I(\boldsymbol{P}_*;(\boldsymbol{\mu}_t,\mathbf{\Sigma}_t)|\{\boldsymbol{p}_t\}_{t\leq N}) \leq \sum_{t=1}^{N} I(\boldsymbol{P}_*;(\boldsymbol{\mu}_t,\mathbf{\Sigma}_t)|\boldsymbol{p}_t).$$

Moreover, we have $\boldsymbol{P}_* \perp\!\!\!\perp (\boldsymbol{\mu}_t,\mathbf{\Sigma}_t) \mid \boldsymbol{X}_{\ell_t,t},\boldsymbol{p}_t$, where $\boldsymbol{X}_{\ell_t,t}$ is a function of $(\boldsymbol{P}_*,\boldsymbol{p}_t)$. Therefore, by Lemma 1,

$$I(\boldsymbol{P}_*;S) \leq \sum_{t=1}^{N} I(\boldsymbol{X}_{\ell_t,t};(\boldsymbol{\mu}_t,\mathbf{\Sigma}_t)|\boldsymbol{p}_t).$$

We have

$$I(\boldsymbol{X}_{\ell_t,t};(\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)|\boldsymbol{p}_t,\mathsf{Z}_{\text{in}}) = \frac{1}{V}\sum_{k=1}^{V} D_{\text{KL}}(\mathbb{P}^k_{(\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)}||\mathbb{P}_0) \overset{(a)}{\leq} \frac{1}{V^2}\sum_{j,k=1}^{V} D_{\text{KL}}(\mathbb{P}^k_{(\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)}||\mathbb{P}^j_{(\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)})$$

where $\mathbb{P}^k_{(\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)}$ denotes the distribution of $(\boldsymbol{s}_t,\boldsymbol{\Sigma}_t)|\boldsymbol{X}_{\ell_t,t}=\boldsymbol{e}_k$, $\mathbb{P}_0$ denotes $\mathbb{P}_0 = \frac{1}{V}\sum_{k=1}^{V}\mathbb{P}^k_{(\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)}$, and (a) follows the convexity of KL divergence in its second argument. For $k \neq j$, by the last item of Lemma 1, we have

$$D_{\text{KL}}(\mathbb{P}^k_{(\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)}||\mathbb{P}^j_{(\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)}) \leq \frac{C}{\sigma^2_{\boldsymbol{\mu}}}\frac{d}{L}\|\mathsf{z}_k-\mathsf{z}_j\|_2^2 + \frac{C}{\sigma^2_{\boldsymbol{\Sigma}}}\frac{d^2}{L}\|\mathsf{z}_k\mathsf{z}_k^\top - \mathsf{z}_j\mathsf{z}_j^\top\|_F^2 \leq \frac{C}{\sigma^2_{\boldsymbol{\mu}}}\frac{d}{L} + \frac{C}{\sigma^2_{\boldsymbol{\Sigma}}}\frac{d^2}{L}.$$

Therefore, we have

$$I(\boldsymbol{P}_*;S) \leq N\Big(\frac{C}{\sigma^2_{\boldsymbol{\mu}}}\frac{d}{L} + \frac{C}{\sigma^2_{\boldsymbol{\Sigma}}}\frac{d^2}{L}\Big).$$

Moreover, we can write

$$\begin{aligned} I(\boldsymbol{P}_*;S) \leq I(\boldsymbol{P}_*;\mathcal{D},\ell_{1:N}) &= I(\boldsymbol{P}_*;\{\boldsymbol{X}_t\}_{t\leq N}|\{\boldsymbol{p}_t\}_{t\leq N},\ell_{1:N}) \\ &\leq \sum_{t=1}^{N} I(\boldsymbol{P}_*;\boldsymbol{X}_{\ell_t,t}|\{\boldsymbol{p}_t,\ell_t\}_{t\leq N}) \\ &\leq \sum_{t=1}^{N} I(\boldsymbol{P}_*;\boldsymbol{X}_{\ell_t,t}|\boldsymbol{p}_t,\ell_t) \end{aligned}$$

where the first inequality follows data processing inequality, third and fourth inequalities follow the first and second items in Lemma 1. We have

$$I(\boldsymbol{P}_*;\boldsymbol{X}_{\ell_t,t}|\boldsymbol{p}_t,\ell_t) = \underbrace{H(\boldsymbol{X}_{\ell_t,t}|\boldsymbol{p}_t,\ell_t)}_{\log V} - \underbrace{H(\boldsymbol{X}_{\ell_t,t}|\boldsymbol{p}_t,\ell_t,\boldsymbol{P}_*)}_{=0} = \log V.$$

Therefore, we have $I(\boldsymbol{P}_*;S) \leq N\log V$. Finally, we have

$$I(\boldsymbol{P}_*;S) \leq N\left(\log V \wedge \Big(\frac{C}{\sigma^2_{\boldsymbol{\mu}}}\frac{d}{L} + \frac{C}{\sigma^2_{\boldsymbol{\Sigma}}}\frac{d^2}{L}\Big)\right).$$

The result follows from (41). □

E Auxiliary Statements

E.1 A nice event characterization

We characterize a "nice event" under which we use in the proof of Theorem 1 holds.

Lemma 2. *We assume $V^3 \gg N \gg V \gg L$ and $L \asymp V^{\epsilon_1}$ and $d \asymp V^{\epsilon_2}$ for some $\epsilon_1,\epsilon_2 \in (0,1)$. For the following we define, $m_{ij} \coloneqq (1-1/V)\delta_{ij} + \frac{L}{V}$. We define the following events:*

(E.1) Let $\mathsf{z}_\nu \coloneqq \boldsymbol{Z}_{\text{in}}\boldsymbol{e}_\nu$ and $\mathsf{z}_{\nu,\delta} \coloneqq (\mathsf{z}_k + \mathbb{1}_{l=1}\delta\mathsf{z}_{\text{trig}})$. We have

- *(E1.1) $\frac{1}{V}\|\boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}^\top\|_2 \leq \frac{2}{d}$ and $\max_{k\leq V}\|\mathsf{z}_k\|_2 \vee \|\mathsf{z}_{\text{trig}}\|_2 \leq 2$*
- *(E1.2) $\frac{1}{\sqrt{V}}\|\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\|_2 \leq 2$ and $\frac{1}{\sqrt{V}}\|\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\|_\infty \leq \frac{\log V}{\sqrt{d}}$*
- *(E1.3) $\left|\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\right| \leq 2\log V\sqrt{\frac{V}{d}}$ and $\left|\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\right| \leq C_K\big(\frac{V}{d}\big)^{\frac{3}{2}}$ and $\left|\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\text{diag}\big(\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\big)\right| \leq C_K\log V\sqrt{\frac{V}{d}}$*
- *(E1.4) For all $i\in[N]$, $|\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L| \leq \boldsymbol{e}_\nu^\top\boldsymbol{X}_i^\top\mathbb{1}_L + C_K\log V\frac{\|\boldsymbol{X}_i^\top\mathbb{1}_L\|_2}{\sqrt{d}}$*
- *(E1.5) For all $i\in[N]$, $|\mathbb{1}_V^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L| \leq L + C_K\log V\|\boldsymbol{X}_i^\top\mathbb{1}_L\|_2\sqrt{\frac{V}{d}}$.*

(E1.6) *For all* $i \in [N]$, $\left|\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L\right| \le \frac{V}{d}\left(\boldsymbol{e}_\nu^\top \boldsymbol{X}_i^\top \mathbb{1}_L + C_K \log V \frac{\|\boldsymbol{X}_i^\top \mathbb{1}_L\|_2}{\sqrt{d}}\right)$.

(E1.7) *For all* $i, j \in [N]$, $|\frac{1}{L}\mathbb{1}_L^\top \boldsymbol{X}_j \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L - m_{ij}| \le |\frac{1}{L}\mathbb{1}_L \boldsymbol{X}_j^\top \boldsymbol{X}_i^\top \mathbb{1}_L - m_{ij}| + C_K \frac{\|\boldsymbol{X}_i^\top \mathbb{1}_L\|_2 \|\boldsymbol{X}_j^\top \mathbb{1}_L\|_2}{L} \frac{\log V}{\sqrt{d}}$,

(E1.8) *For all* $i \in [N]$, $\|\boldsymbol{Z}_{\text{in}} \boldsymbol{N}_i^\top \boldsymbol{N}_i \boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\|_2 \le C_K\left(\boldsymbol{e}_\nu^\top \boldsymbol{N}_i^\top \mathbb{1}_{L-1} + \frac{L}{d} + \log^6 V \frac{\|\boldsymbol{N}_i^\top \mathbb{1}_{L-1}\|_2}{\sqrt{d}}\right)$.

(E.2) *We have*

(E2.1) *For all* $i, j \in [N]$, $|\frac{1}{L}\mathbb{1}_L \boldsymbol{X}_j^\top \boldsymbol{X}_i^\top \mathbb{1}_L - m_{ij}| \le C_K \frac{\log^2 V}{\sqrt{V \wedge L}}$,

(E2.2) *For all* $i \in [N]$, $\|\boldsymbol{X}_i^\top \mathbb{1}_L\|_\infty \le \log L$ *and* $\|\boldsymbol{X}_i^\top \mathbb{1}_L\|_0 \ge \frac{L}{2}$

(E2.3) $\left|\|\frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i\|_2 - \frac{1}{N} - \frac{1}{V}\right| \le C_K \frac{\log^2 N}{N\sqrt{V}}$ *and* $\left|\|\frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V\|_2 - \frac{1}{N}\right| \le C_K \frac{\log^2 N}{N\sqrt{V}}$ *and* $\|\frac{1}{N}\sum_{i=1}^N \boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V\|_\infty \le \frac{(e+1)L}{V}$

(E2.4) $\sum_{i,j=1}^N |\mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} - \frac{1}{V}| \le \frac{4N^2}{V}$ *and* $\sum_{i,j=1}^N (\mathbb{1}_{\boldsymbol{x}_i = \boldsymbol{x}_j} - \frac{1}{V}) \le \frac{4N^2}{V}$

(E2.5) $\|\boldsymbol{S}_1\|_2 \le \frac{e}{L^2 V^2}$ *and* $|\text{tr}(\boldsymbol{S}_1) - \frac{1}{L^2}\left(\frac{1}{N} + (1 - \frac{1}{V})\frac{1}{V}\right)| \le \frac{C_K \log^2 V}{L^2 N\sqrt{V}}$

(E2.6) $\|\boldsymbol{S}_2\|_2 \le \frac{C_K \log^2 V}{NLV}$ *and* $|\text{tr}(\boldsymbol{S}_2) - (1 - \frac{1}{V})^2 \frac{L-1}{L^2 N}| \le \frac{K C_K \log^3 V}{N\sqrt{LV}}$

(E2.7) $\frac{-C_K \log^2 V}{N\sqrt{V}} \frac{1}{V^2 L^2}\mathbb{1}_V \mathbb{1}_V^\top \preceq \boldsymbol{S}_3 - \frac{1}{N}\frac{1}{V^2 L^2}\mathbb{1}_V \mathbb{1}_V^\top \preceq \frac{C_K \log^2 V}{N\sqrt{V}} \frac{1}{V^2 L^2}\mathbb{1}_V \mathbb{1}_V^\top$

For any $K > 0$, *there exists a universal constant* $C_K > 0$ *depending only on* K *such that*

$$\mathbb{P}[(E.1)] \ge 1 - \frac{1}{V^K} \text{ and } \mathbb{P}[(E.2)] \ge 1 - \frac{1}{V^K}.$$

Proof. For (E.1):

- By Proposition 3, we have $\|\frac{1}{V}\boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}^\top - \frac{1}{d}\boldsymbol{I}_d\|_2 \le \frac{2\log V}{\sqrt{Vd}}$ and by Proposition 4, we have $\max_{k \le V}\|\mathsf{z}_k\|_2 \vee \|\mathsf{z}_{\text{trig}}\|_2 \le 2$ with probability at least $1 - CVd\exp(-c\log^2 V)$.

- By Proposition 4, $\frac{1}{\sqrt{V}}\|\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\|_2 \le 2$ and $\frac{1}{\sqrt{V}}\|\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}\mathbb{1}_V\|_\infty \le \frac{2\log V}{\sqrt{d}}$ with probability at least $1 - CVd\exp(-c\log^2 V)$.

- By Propositions 4 and 5, we have $\frac{1}{\sqrt{V}}\left|\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\mathbb{1}_V\right| \le \frac{2\log V}{\sqrt{d}}$ and $\left|\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}\mathbb{1}_V\right| \le C_K\left(\frac{V}{d}\right)^{\frac{3}{2}}$ with probability at least $1 - CVd\exp(-c\log^2 V)$. Moreover

$$\frac{1}{V}\left|\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\text{diag}\left(\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}\right)\right| = \frac{1}{V}\sum_{\substack{i=1\\ i\neq\nu}}^V \|\mathsf{z}_i\|_2^2 \langle \mathsf{z}_i, \mathsf{z}_\nu\rangle + \frac{\mathbb{1}_{l=1}\delta}{V}\sum_{\substack{i=1\\ i\neq\nu}}^V \|\mathsf{z}_i\|_2^2 \langle \mathsf{z}_i, \mathsf{z}_{\text{trig}}\rangle + \underbrace{\frac{1}{V}\mathsf{z}_{\nu,\delta}^\top \mathsf{z}_\nu}_{\in \frac{1}{V}[-C_K, C_K]},$$

where we used previous items to bound the last term. For $i \neq k$, by using Lemma 3, we have for $p \le \frac{d}{6}$,

$$\begin{aligned}\mathbb{E}[\|\mathsf{z}_i\|_2^{4p}\,|\langle \mathsf{z}_i, \mathsf{z}_k\rangle|^{2p}] &\le d^{-p}\,\mathbb{E}[\|\mathsf{z}_i\|_2^{6p}](2p)^p \\ &\le d^{-p}2^p p^p \frac{d(d+2)\cdots(d+6p-2)}{d^{3p}} \le d^{-p}2^{4p}p^p\end{aligned}$$

Therefore,

$$\mathbb{E}[\|\mathsf{z}_i\|_2^{4p}\,|\langle \mathsf{z}_i, \mathsf{z}_\nu\rangle|^{2p}]^{\frac{1}{2p}} \le 4d^{-1/2}\sqrt{p}.$$

By Proposition 13, we have for $2 \le p \le \frac{d}{6}$,

$$\mathbb{E}\left[\left|\frac{1}{V}\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}}\text{diag}\left(\boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}}\right)\right|^{2p}\right]^{\frac{1}{2p}} \le Cd^{-1/2}\left[\sqrt{\frac{p}{V}} + V^{\frac{1}{p}}\frac{p^{3/2}}{V}\right]$$

By using $p = \log V$, we have the bound in the statement with probability $1 - \frac{1}{V^K}$.

- By Proposition 4 with probability at least $1 - \frac{1}{V^K}$

$$|\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L| \leq |\boldsymbol{e}_\nu^\top \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L| + \mathbb{1}_{l=1}\delta |\mathsf{z}_{\text{trig}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L|$$
$$\leq \boldsymbol{e}_\nu^\top \boldsymbol{X}_i^\top \mathbb{1}_L + C_K \log V \frac{\|\boldsymbol{X}_i^\top \mathbb{1}_L\|_2}{\sqrt{d}}.$$

  By union bound, the item follows.

- By the same argument,

$$|\mathbb{1}_V^\top \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L| \leq \mathbb{1}_V^\top \boldsymbol{X}_i^\top \mathbb{1}_L + C_K \log V \|\boldsymbol{X}_i^\top \mathbb{1}_L\|_2 \sqrt{\frac{V}{d}}$$
$$= L + C_K \log V \|\boldsymbol{X}_i^\top \mathbb{1}_L\|_2 \sqrt{\frac{V}{d}}$$

- By Proposition 5, with probability at least $1 - CN\exp(-c\log^2 V)$, we have $\left|\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L\right| \leq \frac{V}{d}\big(\boldsymbol{e}_\nu^\top \boldsymbol{X}_i^\top \mathbb{1}_L + C_K \log V \frac{\|\boldsymbol{X}_i^\top \mathbb{1}_L\|_2}{\sqrt{d}}\big)$ for all $i \in [N]$.

- By Proposition 4, with probability at least $1 - \frac{1}{V^K}$

$$\frac{1}{L}\mathbb{1}_L^\top \boldsymbol{X}_j \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L - m_{ij} = \frac{1}{L}\mathbb{1}_L^\top \boldsymbol{X}_j \boldsymbol{X}_i^\top \mathbb{1}_L - m_{ij}$$
$$\pm C \log V \frac{\|\boldsymbol{X}_j^\top \mathbb{1}_L\|_2 \|\boldsymbol{X}_i^\top \mathbb{1}_L\|_2}{L} \frac{\log V}{\sqrt{d}}.$$

- For the last item, let $n_k := \boldsymbol{e}_k^\top \boldsymbol{Z}_{\text{in}}$. We have

$$\boldsymbol{Z}_{\text{in}} \boldsymbol{N}_i^\top \boldsymbol{N}_i \boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta} = n_\nu(\|\mathsf{z}_\nu\|_2^2 + \delta \mathbb{1}_{l=1} \mathsf{z}_\nu^\top \mathsf{z}_\delta - \frac{1}{d})\mathsf{z}_\nu + \frac{L}{d}\mathsf{z}_\nu + \sum_{\substack{k=1\\k\neq\nu}}^{V} n_k\big(\mathsf{z}_k \mathsf{z}_k^\top - \frac{1}{d}\boldsymbol{I}_d\big)\mathsf{z}_{\nu,\delta}.$$

  By Proposition 11, we have

$$\mathbb{E}\left[\Big\|\sum_{\substack{k=1\\k\neq\nu}}^{V} n_k\big(\mathsf{z}_k \mathsf{z}_k^\top - \frac{1}{d}\boldsymbol{I}_d\big)\mathsf{z}_{\nu,\delta}\Big\|_2^{2p}\right]^{\frac{1}{p}} \leq C(p-1)^6\, \mathbb{E}\left[\Big\|\sum_{\substack{k=1\\k\neq\nu}}^{V} n_k\big(\mathsf{z}_k \mathsf{z}_k^\top - \frac{1}{d}\boldsymbol{I}_d\big)\mathsf{z}_{\nu,\delta}\Big\|_2^{2}\right]$$
$$\leq \frac{C}{d}(p-1)^6 \|\boldsymbol{N}_i^\top \mathbb{1}_{L-1}\|_2^2$$

  Therefore, with probability $1 - \frac{1}{V^K}$, we have

$$\|\boldsymbol{Z}_{\text{in}} \boldsymbol{N}_i^\top \boldsymbol{N}_i \boldsymbol{Z}_{\text{in}}^\top \mathsf{z}_{\nu,\delta}\|_2 \leq C_K\Big(n_\nu + \frac{L}{d} + \log^6 V \frac{\|\boldsymbol{N}_i^\top \mathbb{1}_{L-1}\|_2}{\sqrt{d}}\Big).$$

For (E.2):

- By Proposition 7, we have the first item with probability $1 - \frac{N^2}{V^K}$.

- By Corollary 3, we have $\|\boldsymbol{X}_i^\top \mathbb{1}_L\|_\infty \leq \log L$. For the second part, we define $n_k := \boldsymbol{e}_k \boldsymbol{X}_i^\top \mathbb{1}_L$. We observe that

$$\mathbb{E}[\|\boldsymbol{X}_i^\top \mathbb{1}_L\|_0] = \sum_{k=1}^{V} \mathbb{P}[n_k > 0] = V\big(1 - (1 - \frac{1}{V})^L\big) = L(1 - \frac{L}{2V} + o(L/V)).$$

  By McDiarmid inequality, we have

$$\mathbb{P}\Big[\big|\|\boldsymbol{X}_i^\top \mathbb{1}_L\|_0 - L(1 - \frac{L}{2V} + o(L/V))\big| > \sqrt{L}\log V\Big] \leq 2\exp(-2\log^2 V),$$

  which gives the result.

- Let $\boldsymbol{n} = \sum_{i=1}^{N} \boldsymbol{x}_i$. We have $\mathbb{E}[\boldsymbol{n}] = \frac{N}{V}\mathbb{1}_V$ and by Proposition 7 with probability $1 - \frac{1}{V^K}$, we have

$$\Big|\|\frac{1}{N}\boldsymbol{n} - \frac{1}{V}\mathbb{1}_V\|_2^2 - (1-\frac{1}{V})\frac{1}{N}\Big| = \Big|\|\frac{1}{N}\boldsymbol{n}\|_2^2 - (1-\frac{1}{V})\frac{1}{N} - \frac{1}{V}\Big| \leq C_K \frac{\log^2 V}{N\sqrt{V}}.$$

Lastly, by Corollary 3, we have $\|\frac{1}{N}\boldsymbol{n} - \frac{1}{V}\mathbb{1}_V\|_\infty \leq \frac{(e+1)L}{V}$.

- We have

$$\begin{aligned}\sum_{i,j=1}^{N} |\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \frac{1}{V}| &= \Big(\sum_{i,j=1}^{N} |\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \frac{1}{V}| - \frac{2}{V}(1-\frac{1}{V})\Big) + \frac{2N^2}{V}(1-\frac{1}{V}) \\ &= (1-\frac{2}{V})\sum_{i,j=1}^{N}(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_j} - \frac{1}{V}) + \frac{2N^2}{V}(1-\frac{1}{V}) \\ &= (1-\frac{2}{V})\Big\|\sum_{i=1}^{N}(\boldsymbol{x}_i - \frac{1}{V}\mathbb{1}_V)\Big\|_2^2 + \frac{2N^2}{V}(1-\frac{1}{V})\end{aligned}$$

By the previous item, the statement follows

- The events for $\boldsymbol{S}_1$, $\boldsymbol{S}_2$ and $\boldsymbol{S}_3$ follows Proposition 8.

□

Proposition 2. *We consider the parameter regime in Lemma 2. Let* $\bar{\phi} := \sup_{k_1,k_2\geq 1}|\phi^{(k_1)}(0)\phi^{(k_2)}(0)|$. *The intersection of (E.1) and (E.2) implies the following events:*

(C.1) For all $i,j \in [N]$, $|\frac{1}{L}\mathbb{1}_L \boldsymbol{X}_j^\top \boldsymbol{Z}_{\text{in}}^\top \boldsymbol{Z}_{\text{in}} \boldsymbol{X}_i^\top \mathbb{1}_L - m_{ij}| \leq C_K\big(\frac{\log V}{\sqrt{d}} + \frac{\log^2 V}{L}\big)$,

(C.2) $\sup_{i,j}|\alpha_{ij} - \phi'(0)^2| \vee |\beta_{ij} - \phi''(0)\phi(0)| \leq \frac{\bar{\phi}}{L}\big(m_{ij} + C_K\frac{\log V}{\sqrt{d}} + C_K\frac{\log^2 V}{L}\big)$

(C.3) Let $\Delta_{u,ir} := \boldsymbol{A}_{u,ir} - \phi'(0)^4 \boldsymbol{Z}_{\text{in}}\boldsymbol{S}_u\boldsymbol{Z}_{\text{in}}^\top$ *for* $u \in \{1,2,3\}$. *We have*

- $\sup_{i,r\in[N]}\|\Delta_{1,ir}\|_2 \leq C_K\phi'(0)^2\Big(\frac{1}{NdL^3} + \frac{1}{VdL^2}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big)$.
- $\sup_{i,r\in[N]}\|\Delta_{2,ir}\|_2 \leq \frac{C_K\sqrt{V}}{d\sqrt{NL}}\Big(\frac{1}{NL^{\frac{3}{2}}} + \frac{1}{V\sqrt{L}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big)$.
- *We have* $\Delta_{3,ir} = \frac{\bar{\Delta}_{3,ir}}{V^2L^2}\mathsf{Z}_{\text{in}}\mathbb{1}_V\mathbb{1}_V^\top\mathsf{Z}_{\text{in}}^\top$ *such that*

$$\sup_{i,r\in[N]}|\bar{\Delta}_{3,ir}| \leq \frac{C_K\phi'(0)^2}{N}\Big(\frac{1}{NL} + \frac{1}{\sqrt{N}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big) + \Big(\frac{1}{NL} + \frac{1}{\sqrt{N}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big)^2.$$

(C.4) For all $i,r \in [N]$,

- *We have*

$$\begin{aligned}\Big\|\boldsymbol{A}_{1,ir} - \frac{\phi'(0)^4}{d}(\frac{1}{N} + (1-\frac{1}{V})\frac{1}{V})\boldsymbol{I}_d\Big\|_2 &\leq C_K\phi'(0)^2\Big(\frac{1}{NdL^3} + \frac{1}{VdL^2}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big) \\ &+ C_K\phi'(0)^4\Big(\frac{\log V}{L^2V^{3/2}\sqrt{d}} + \frac{\log^2 V}{L^2N\sqrt{V}d}\Big).\end{aligned}$$

- *We have*

$$\begin{aligned}\Big\|\boldsymbol{A}_{2,ir} - \frac{\phi'(0)^4}{d}(1-\frac{1}{V})^2\frac{L-1}{L^2N}\boldsymbol{I}_d\Big\|_2 &\leq \frac{C_K\sqrt{V}}{d\sqrt{NL}}\Big(\frac{1}{NL^{\frac{3}{2}}} + \frac{1}{V\sqrt{L}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big) \\ &+ C_K\phi'(0)^4\Big(\frac{\log V}{NL\sqrt{V}d} + \frac{\log^3 V}{N\sqrt{LV}d}\Big).\end{aligned}$$

- *We have* $\boldsymbol{A}_{3,ir} - \frac{\phi'(0)^4}{N}\frac{1}{V^2L^2}\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\text{in}}^\top =: \frac{\tilde{\Delta}_{3,ir}}{V^2L^2}\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\text{in}}^\top$ *such that*

$$|\tilde{\Delta}_{3,ir}| \le \frac{C_K\phi'(0)^4\log^2 V}{N\sqrt{V}} + \frac{C_K\phi'(0)^2}{N}\Big(\frac{1}{NL} + \frac{1}{\sqrt{N}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big) + \Big(\frac{1}{NL} + \frac{1}{\sqrt{N}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big)^2.$$

Proof. We have the following arguments.

- By (E1.7) and (E2.1), we have (C.1).
- For (C.2), we assume (E2.1) and (C.1) hold. Let $\|\frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\|_2 w_i \leftarrow \frac{1}{L}\boldsymbol{w}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L$. We write

$$\begin{aligned}
&\big|\,\mathbb{E}\left[\phi'\big(\|\frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\|_2 w_i\big)\phi'\big(\|\frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\|_2 w_j\big)\right] - \phi'(0)^2\big| \\
&= \Big|\sum_{u,v=1}^{p_\star}\|\frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\|_2^u\|\frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L\|_2^v\frac{\mathbb{E}\left[w_i^u w_j^v\right]}{u!v!}\phi^{(u+1)}(0)\phi^{(v+1)}(0)\Big| \\
&= \Big|\frac{1}{L^2}\mathbb{1}_L^\top\boldsymbol{X}_j\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\phi^{(2)}(0)\phi^{(2)}(0) \\
&\quad + \sum_{\substack{u,v=1\\ u+v \text{ is even}\\ u+v>2}}^{p_\star}\|\frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\|_2^u\|\frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L\|_2^v\frac{\mathbb{E}\left[w_i^u w_j^v\right]}{u!v!}\phi^{(u+1)}(0)\phi^{(v+1)}(0)\Big| \\
&\le \frac{\bar{\phi}}{L}\big(m_{ij} + C_K\frac{\log V}{\sqrt{d}} + C_K\frac{\log^2 V}{\sqrt{V}\wedge L}\big) + O(\frac{1}{L^2})
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\big|\,\mathbb{E}\left[\phi''\big(\|\frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\|_2 w_i\big)\phi\big(\|\frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\|_2 w_j\big)\right] - \phi'(0)^2\big| \\
&= \Big|\sum_{u,v=1}^{k_2}\|\frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\|_2^u\|\frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L\|_2^v\frac{\mathbb{E}\left[w_i^u w_j^v\right]}{u!v!}\phi^{(u+2)}(0)\phi^{(v)}(0)\Big| \\
&= \Big|\frac{1}{L^2}\mathbb{1}_L^\top\boldsymbol{X}_j\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\phi^{(2)}(0)\phi^{(2)}(0) \\
&\quad + \sum_{\substack{u,v=1\\ u+v \text{ is even}\\ u+v>2}}^{k_2}\|\frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\mathbb{1}_L\|_2^u\|\frac{1}{L}\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_j^\top\mathbb{1}_L\|_2^v\frac{\mathbb{E}\left[w_i^u w_j^v\right]}{u!v!}\phi^{(u+2)}(0)\phi^{(v)}(0)\Big| \\
&\le \frac{\bar{\phi}}{L}\big(m_{ij} + CK\frac{\log V}{\sqrt{d}} + CK\frac{\log^2 V}{\sqrt{V}\wedge L}\big) + O(\frac{1}{L^2})
\end{aligned}$$

- For (C.3), we assume (E1.1), (E1.2) and (E2.5)-**??**. We define

$$\begin{aligned}
\bar{\Delta}_{1,ir} &:= \Big(\frac{1}{LN}\sum_{j=1}^{N}(\alpha_{ij} - \phi'(0)^2)(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top\Big) \\
&\qquad \times \Big(\frac{1}{LN}\sum_{j=1}^{N}\phi'(0)^2(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top\Big) \\
&+ \Big(\frac{1}{LN}\sum_{j=1}^{N}\phi'(0)^2(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top\Big) \\
&\qquad \times \Big(\frac{1}{LN}\sum_{j=1}^{N}(\alpha_{uj} - \phi'(0)^2)(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)^\top\Big)
\end{aligned}$$

$$+\Big(\frac{1}{LN}\sum_{j=1}^{N}(\alpha_{ij}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)^\top\Big)$$
$$\times\Big(\frac{1}{LN}\sum_{j=1}^{N}(\alpha_{uj}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)^\top\Big)$$

We have

$$\begin{aligned}\|\bar{\Delta}_{1,ir}\|_2 &\le \frac{C\phi'(0)^2\sup_i|\alpha_{ii}-\phi'(0)^2|}{LN}\|\boldsymbol{S}_1\|_2^{\frac{1}{2}}+\phi'(0)^2\sup_{i\neq j}|\alpha_{ij}-\phi'(0)^2|\|\boldsymbol{S}_1\|_2\\ &\le C\phi'(0)^2\Big(\frac{1}{NVL^3}+\frac{1}{V^2L^2}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big).\end{aligned}$$

Therefore,

$$\|\Delta_{1,ir}\|_2=\|\mathsf{Z}_{\mathrm{in}}\bar{\Delta}_{1,ir}\mathsf{Z}_{\mathrm{in}}^\top\|_2\le C\phi'(0)^2\Big(\frac{1}{NdL^3}+\frac{1}{VdL^2}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big).$$

Moreover, we define

$$\begin{aligned}\bar{\Delta}_{2,ir}\coloneqq&\Big(\frac{1}{NL}\sum_{j=1}^{N}(\alpha_{ij}-\phi'(0)^2)(\boldsymbol{N}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)^\top\Big)\\ &\Big(\frac{1}{NL}\sum_{j=1}^{N}\phi'(0)^2(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big)\\ &+\Big(\frac{1}{NL}\sum_{j=1}^{N}\phi'(0)^2(\boldsymbol{N}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)^\top\Big)\\ &\Big(\frac{1}{NL}\sum_{j=1}^{N}(\alpha_{rj}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big)\\ &+\Big(\frac{1}{NL}\sum_{j=1}^{N}(\alpha_{ij}-\phi'(0)^2)(\boldsymbol{N}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)^\top\Big)\\ &\Big(\frac{1}{NL}\sum_{j=1}^{N}(\alpha_{rj}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big).\end{aligned}$$

We have

$$\begin{aligned}&\|\bar{\Delta}_{2,ir}\|_2\\ &\le\phi'(0)^2\|\boldsymbol{S}_2\|_2^{\frac{1}{2}}\Big\|\Big(\frac{1}{NL}\sum_{j=1}^{N}(\alpha_{rj}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big)\Big\|_2\\ &+\phi'(0)^2\|\boldsymbol{S}_2\|_2^{\frac{1}{2}}\Big\|\Big(\frac{1}{NL}\sum_{j=1}^{N}(\alpha_{ij}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big)\Big\|_2\\ &+\Big\|\Big(\frac{1}{NL}\sum_{j=1}^{N}(\alpha_{rj}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big)\Big\|_2\\ &\quad\times\underbrace{\Big\|\Big(\frac{1}{NL}\sum_{j=1}^{N}(\alpha_{ij}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big)\Big\|_2}_{\le\frac{C}{NL\sqrt{L}}+\frac{C}{V\sqrt{L}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}}\\ &\le\frac{C}{\sqrt{NVL}}\Big(\frac{1}{NL^{\frac{3}{2}}}+\frac{1}{V\sqrt{L}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big)+C^2\Big(\frac{1}{NL^{\frac{3}{2}}}+\frac{1}{V\sqrt{L}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big)^2.\end{aligned}$$

Therefore,

$$\|\Delta_{2,ir}\|_2 = \|\mathsf{Z}_{\text{in}}\bar{\Delta}_{2,ir}\mathsf{Z}_{\text{in}}^\top\|_2 \leq \frac{C\sqrt{V}}{d\sqrt{NL}}\Big(\frac{1}{NL^{\frac{3}{2}}} + \frac{1}{V\sqrt{L}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big).$$

Lastly, we define

$$\bar{\Delta}_{3,ir} \coloneqq \Big(\frac{1}{N}\sum_{j=1}^N(\alpha_{ij}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big)^\top\Big(\frac{1}{N}\sum_{j=1}^N\phi'(0)^2(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big)$$
$$+\Big(\frac{1}{N}\sum_{j=1}^N\phi'(0)^2(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big)^\top\Big(\frac{1}{N}\sum_{j=1}^N(\alpha_{rj}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big)$$
$$+\Big(\frac{1}{N}\sum_{j=1}^N(\alpha_{ij}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big)^\top\Big(\frac{1}{N}\sum_{j=1}^N(\alpha_{rj}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big)$$

We have

$$|\bar{\Delta}_{3,ir}| \leq \phi'(0)^2\Big\|\frac{1}{N}\sum_{j=1}^N(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big\|_2\Big\|\frac{1}{N}\sum_{j=1}^N(\alpha_{ij}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big\|_2$$
$$+\phi'(0)^2\Big\|\frac{1}{N}\sum_{j=1}^N(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big\|_2\Big\|\frac{1}{N}\sum_{j=1}^N(\alpha_{rj}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big\|_2$$
$$+\Big\|\frac{1}{N}\sum_{j=1}^N(\alpha_{ij}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big\|_2\underbrace{\Big\|\frac{1}{N}\sum_{j=1}^N(\alpha_{rj}-\phi'(0)^2)(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big\|_2}_{\leq\frac{C}{NL}+\frac{C}{\sqrt{N}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}}$$
$$\leq\frac{C\phi'(0)^2}{N}\Big(\frac{1}{NL}+\frac{1}{\sqrt{N}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big)+\Big(\frac{1}{NL}+\frac{1}{\sqrt{N}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big)^2.$$

Therefore,

$$\Delta_{3,ir} = \frac{\bar{\Delta}_{3,ir}}{V^2L^2}\mathsf{Z}_{\text{in}}\mathbb{1}_V\mathbb{1}_V^\top\mathsf{Z}_{\text{in}}^\top,$$

from which the last result follows.

- For (C.4), we assume (E1.1), (E1.2), (E2.3), and (E2.5)-(E2.7). We write

$$\boldsymbol{A}_{1,ir}-\frac{\phi'(0)^4}{d}\frac{1}{L^2}\big(\frac{1}{N}+(1-\frac{1}{V})\frac{1}{V}\big)\boldsymbol{I}_d$$
$$=\Delta_{1,ir}+\phi'(0)^4\Big(\boldsymbol{Z}_{\text{in}}\boldsymbol{S}_1\boldsymbol{Z}_{\text{in}}\pm\frac{\text{tr}(\boldsymbol{S}_1)}{d}\boldsymbol{I}_d-\frac{1}{d}\big((1-\frac{1}{V})\frac{1}{N}+(1-\frac{1}{V})\frac{1}{V}\big)\boldsymbol{I}_d\Big).$$

We have

$$\|\boldsymbol{A}_{1,ir}-\frac{\phi'(0)^4}{d}\big((1-\frac{1}{V})^2\frac{1}{N}+(1-\frac{1}{V})\frac{1}{V}\big)\boldsymbol{I}_d\|_2$$
$$\leq\|\Delta_{1,ir}\|_2+2\phi'(0)^4\log V\frac{\|\boldsymbol{S}_1\|_F}{\sqrt{d}}+\frac{|\text{tr}(\boldsymbol{S}_1)-\frac{1}{L^2}\big(\frac{1}{N}+(1-\frac{1}{V})\frac{1}{V}\big)|}{d}$$
$$\leq C\phi'(0)^2\Big(\frac{1}{NdL^3}+\frac{1}{VdL^2}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big)+C\phi'(0)^4\Big(\frac{\log V}{L^2V^{3/2}\sqrt{d}}+\frac{CK^2\log^2 V}{L^2N\sqrt{V}d}\Big).$$

Moreover,

$$\|\boldsymbol{A}_{2,ir}-\frac{\phi'(0)^4}{d}(1-\frac{1}{V})^2\frac{L-1}{L^2N}\boldsymbol{I}_d\|_2$$
$$\leq\|\Delta_{2,ir}\|_2+2\phi'(0)^4\log V\frac{\|\boldsymbol{S}_2\|_F}{\sqrt{d}}+\frac{|\text{tr}(\boldsymbol{S}_2)-(1-\frac{1}{V})^2\frac{L-1}{L^2N}|}{d}$$

$$\leq \frac{C\sqrt{V}}{d\sqrt{NL}}\Big(\frac{1}{NL^{\frac{3}{2}}}+\frac{1}{V\sqrt{L}}\frac{1}{V\wedge L^2\wedge L\sqrt{d}}\Big)+C\phi'(0)^4\Big(\frac{\log V}{NL\sqrt{Vd}}+\frac{K^{\frac{3}{2}}\log^3 V}{N\sqrt{LVd}}\Big).$$

Lastly,

$$\begin{aligned}&\boldsymbol{A}_{3,ir}-\frac{\phi'(0)^4}{N}\frac{1}{V^2L^2}\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\text{in}}^\top\\&=\Delta_{3,ir}+\phi'(0)^4\Big(\Big\|\frac{1}{N}\sum_{j=1}^N(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big\|_2^2-\frac{1}{N}\Big)\frac{1}{V^2L^2}\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\text{in}}^\top.\end{aligned}$$

By (E2.3), we have

$$\begin{aligned}\frac{CK^2\log^2 V}{N\sqrt{V}}\frac{1}{V^2L^2}\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\text{in}}^\top &\preceq \Big(\Big\|\frac{1}{N}\sum_{j=1}^N(\boldsymbol{x}_j-\frac{1}{V}\mathbb{1}_V)\Big\|_2^2-\frac{1}{N}\Big)\frac{1}{V^2L^2}\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\text{in}}^\top\\&\preceq\frac{CK^2\log^2 V}{N\sqrt{V}}\frac{1}{V^2L^2}\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\text{in}}^\top.\end{aligned}$$

By (C.3), the result follows.

□

E.2 Gaussian matrices and related statements

Lemma 3. *Let $\boldsymbol{z}\sim\mathcal{N}(0,\boldsymbol{I}_d)$. We have $\mathbb{E}[\|\boldsymbol{z}\|_2^{2k}]=d(d+2)\cdots(d+2k-2)$.*

Proof. We observe that $\|\boldsymbol{z}\|_2\sim\chi_d^2$. By using the moment formula for chi-squared distribution, we have the result. □

Lemma 4. *Let $\boldsymbol{z}\sim\mathcal{N}(0,\boldsymbol{I}_d)$ and $\boldsymbol{S}\in\mathbb{R}^{d\times d}$ be a symmetric matrix. For $u>0$,*

$$\mathbb{P}\left[|\boldsymbol{z}^\top\boldsymbol{S}\boldsymbol{z}-\text{tr}(\boldsymbol{S})|\geq 2\|\boldsymbol{S}\|_F u+2\|\boldsymbol{S}\|_2u^2\right]\leq 2e^{-u^2}.$$

Proof. We note that $\boldsymbol{z}^\top\boldsymbol{S}\boldsymbol{z}-\text{tr}(\boldsymbol{S})$ has the same distribution with $\sum_{i=1}^d\lambda_i(\boldsymbol{S})(Z_i^2-1)$, where $Z_i\sim_{iid}\mathcal{N}(0,1)$. By using the Laurent-Massart lemma, we have the result. □

Proposition 3. *Let $\boldsymbol{S}\in\mathbb{R}^{V\times V}$ be a symmetric positive semidefinite matrix. Let*

$$\boldsymbol{M}=\boldsymbol{Z}_{\text{in}}\boldsymbol{S}\boldsymbol{Z}_{\text{in}}^\top.$$

For $\text{poly}(d)\gg V\gg d$*, We have*

$$\mathbb{P}\left[\Big\|\boldsymbol{M}-\frac{\text{tr}(\boldsymbol{S})}{d}\boldsymbol{I}_d\Big\|_2\geq\max\Big\{\frac{\|\boldsymbol{S}\|_F}{\sqrt{d}}\log V,\|\boldsymbol{S}\|_2\log^2 V\Big\}\right]\leq\exp(-c\log^2 V).$$

Proof. Without loss of generality, we can assume that $\boldsymbol{S}$ is diagonal, i.e., $\boldsymbol{S}=\text{diag}(s_1,\cdots,s_V)$. We have

$$\boldsymbol{M}-\frac{\text{tr}(\boldsymbol{S})}{d}=\sum_{i=1}^V s_i\big(\boldsymbol{z}_i\boldsymbol{z}_i^\top-\frac{1}{d}\boldsymbol{I}_d\big).$$

We have

$$\mathbb{E}\Big[\Big(\sum_{i=1}^V s_i\big(\boldsymbol{z}_i\boldsymbol{z}_i^\top-\frac{1}{d}\boldsymbol{I}_d\big)\Big)^2\Big]=\frac{1}{d}(1+\frac{1}{d})\|\boldsymbol{S}\|_F^2\boldsymbol{I}_d$$

Moreover, for $p\leq\frac{d}{2}$

$$\mathbb{E}\left[\|\boldsymbol{z}_i\boldsymbol{z}_i^\top-\frac{1}{d}\boldsymbol{I}_d\|_2^p\right]\leq\mathbb{E}[\|\boldsymbol{z}_i\|_2^{2p}]\leq 2^p.$$

By Proposition 13, we have $2 \le p \le \frac{d}{2}$

$$\mathbb{E}\left[\|\boldsymbol{M} - \frac{\mathrm{tr}(\boldsymbol{S})}{d}\|_2^p\right] \le C\Big(\sqrt{p \vee \log d}\frac{\|\boldsymbol{S}\|_F}{\sqrt{d}} + (p \vee \log d)V^{\frac{1}{p}}\|\boldsymbol{S}\|_2\Big).$$

For $p = \frac{1}{e^2C^2}\log^2 V$, we have the result. □

Proposition 4. *Let $\boldsymbol{S} \in \mathbb{R}^{V\times V}$ be a square matrix. For $\boldsymbol{u}, \boldsymbol{v} \in S^{d-1}$ and $\boldsymbol{M} = \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{S}\boldsymbol{Z}_{\mathrm{in}}^\top$, we have*

$$\mathbb{P}\Big[\Big|\Big(\boldsymbol{v}^\top\boldsymbol{M}\boldsymbol{u} - \frac{\mathrm{tr}(\boldsymbol{S})}{d}\boldsymbol{v}^\top\boldsymbol{u}\Big)\Big| \ge \frac{\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2}{d}\max\Big\{\|\mathrm{sym}(\boldsymbol{S})\|_F t, \|\mathrm{sym}(\boldsymbol{S})\|_2 t^2\Big\}\Big] \le 2\exp(-ct^2).$$

Proof. Consider $\boldsymbol{g} = \sqrt{d}\mathrm{vec}(\boldsymbol{Z})$, where $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_{dV})$. We have

$$\boldsymbol{v}^\top\boldsymbol{M}\boldsymbol{u} = \frac{1}{d}\boldsymbol{g}^\top(\boldsymbol{u}\boldsymbol{v}^\top) \otimes \boldsymbol{S}\boldsymbol{g} = \frac{1}{d}\boldsymbol{g}^\top\mathrm{sym}(\boldsymbol{u}\boldsymbol{v}^\top) \otimes \mathrm{sym}(\boldsymbol{S})\boldsymbol{g}$$

By using Proposition 10, we have

$$\mathbb{E}[\boldsymbol{g}^\top\mathrm{sym}(\boldsymbol{u}\boldsymbol{v}^\top) \otimes \mathrm{sym}(\boldsymbol{S})\boldsymbol{g}] = \mathrm{tr}(\boldsymbol{S})\boldsymbol{u}^\top\boldsymbol{v}.$$

Moreover,

$$\Big(\boldsymbol{g}^\top\mathrm{sym}(\boldsymbol{u}\boldsymbol{v}^\top) \otimes \mathrm{sym}(\boldsymbol{S})\boldsymbol{g} - \mathrm{tr}(\boldsymbol{S})\boldsymbol{u}^\top\boldsymbol{v}\Big) =_d \sum_{i=1}^{dV}\lambda_i(g_i^2 - 1)$$

where $g_i \sim N(0,1)$. By using the subexponential concentration, we have the result. □

Proposition 5. *For $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^V$, we have*

$$\mathbb{P}\Big[\Big|\boldsymbol{v}^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{u} - \boldsymbol{u}^\top\boldsymbol{v}\Big(1 + \frac{V-1}{d}\Big)\Big| \ge C\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2\log V\Big(\frac{\sqrt{V}}{d} + \frac{V}{d^{3/2}}\Big)\Big] \le 10\exp(-c\log^2 V).$$

Proof. Without loss of generality, we assume that $\boldsymbol{u}$ and $\boldsymbol{v}$ have a unit norm. Let

$$\boldsymbol{v}_\perp := \frac{1}{\sqrt{1 - (\boldsymbol{u}^\top\boldsymbol{v})^2}}(\boldsymbol{I}_V - \boldsymbol{v}\boldsymbol{v}^\top)\boldsymbol{u}.$$

We have

$$\boldsymbol{v}^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{u} = (\boldsymbol{u}^\top\boldsymbol{v})\boldsymbol{v}^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{v} + \sqrt{1 - (\boldsymbol{u}^\top\boldsymbol{v})^2}\boldsymbol{v}^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{v}_\perp.$$

Without loss of generality, we consider $\boldsymbol{v} = \boldsymbol{e}_1$ and $\boldsymbol{v}_\perp = \boldsymbol{e}_2$. For the second term, we write $\boldsymbol{z}_i := \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{e}_i$ and let $\tilde{\boldsymbol{Z}} := \{\boldsymbol{z}_i\}_{i=3}^V$ and $\boldsymbol{g} = \sqrt{d}\mathrm{vec}(\tilde{\boldsymbol{Z}})$.

$$\begin{aligned}\boldsymbol{e}_1^\top\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{e}_2 &= (\|\boldsymbol{z}_1\|_2^2 + \|\boldsymbol{z}_2\|_2^2)\boldsymbol{z}_1^\top\boldsymbol{z}_2 + \boldsymbol{z}_1^\top\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^\top\boldsymbol{z}_2 \\ &= (\|\boldsymbol{z}_1\|_2^2 + \|\boldsymbol{z}_2\|_2^2)\boldsymbol{z}_1^\top\boldsymbol{z}_2 + \frac{1}{d}\boldsymbol{g}^\top\mathrm{sym}(\boldsymbol{z}_1\boldsymbol{z}_2^\top) \otimes \boldsymbol{I}_{V-2}\boldsymbol{g}.\end{aligned}$$

We have

- By Lemma 4, and Proposition 4

$$\mathbb{P}\Big[\big|\|\boldsymbol{z}_1\|_2^2 - 1\big| \le \frac{5\log V}{\sqrt{d}} \text{ and } \big|\|\boldsymbol{z}_2\|_2^2 - 1\big| \le \frac{5\log V}{\sqrt{d}} \text{ and } |\boldsymbol{z}_1^\top\boldsymbol{z}_2| \le \frac{\log V}{\sqrt{d}}\Big] \le 1 - 6\exp(-c\log^2 V).$$

- By Proposition 10, we have
  - $\|\mathrm{sym}(\boldsymbol{z}_1\boldsymbol{z}_2^\top) \otimes \boldsymbol{I}_{V-2}\|_2 \le \|\boldsymbol{z}_1\|_2\|\boldsymbol{z}_2\|_2$

- $\|\mathrm{sym}(\boldsymbol{z}_1\boldsymbol{z}_2^\top)\otimes \boldsymbol{I}_{V-2}\|_F \le \sqrt{V}\|\boldsymbol{z}_1\|_2\|\boldsymbol{z}_2\|_2$
- $\mathrm{tr}\big(\mathrm{sym}(\boldsymbol{z}_1\boldsymbol{z}_2^\top)\otimes \boldsymbol{I}_{V-2}\big) = (V-2)\boldsymbol{z}_1^\top\boldsymbol{z}_2$.

Therefore, by Lemma 4, we have

$$\mathbb{P}\Big[\Big|\frac{1}{d}\boldsymbol{g}^\top \mathrm{sym}(\boldsymbol{z}_1\boldsymbol{z}_2^\top)\otimes \boldsymbol{I}_{V-2}\boldsymbol{g} - \frac{(V-2)}{d}\boldsymbol{z}_1^\top\boldsymbol{z}_2\Big| \le 2\|\boldsymbol{z}_1\|_2\|\boldsymbol{z}_2\|_2\Big(\frac{\log V}{d}\sqrt{V} + \frac{\log^2 V}{d}\Big)\Big] \le 1 - 2\exp(-c\log^2 V).$$

By union bound of the precious two items, we have

$$\mathbb{P}\Big[\Big|\boldsymbol{e}_1^\top \boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{Z}_{\mathrm{in}}^\top\boldsymbol{Z}_{\mathrm{in}}\boldsymbol{e}_2\Big| \le 2\log V\Big(\frac{V}{d^{3/2}} + \frac{\sqrt{V}}{d}\Big)\Big] \ge 1 - 8\exp(-c\log^2 V). \tag{42}$$

Next, we redefine the notation: $\tilde{\boldsymbol{Z}} \coloneqq \{\boldsymbol{z}_i\}_{i=2}^V$. We write

$$\boldsymbol{z}_1^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{z}_1 - 1 - \frac{V-1}{d} = \|\boldsymbol{z}_1\|_2^4 - 1 + \boldsymbol{z}_1^\top\Big(\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^\top - \frac{V-1}{d}\boldsymbol{I}_d\Big)\boldsymbol{z}_1 - \frac{V-1}{d}(\|\boldsymbol{z}_1\|_2^2 - 1)$$

By Proposition bla, we have

$$\mathbb{P}\Big[\boldsymbol{z}_1^\top\Big(\tilde{\boldsymbol{Z}}\tilde{\boldsymbol{Z}}^\top - \frac{V-1}{d}\boldsymbol{I}_d\Big)\boldsymbol{z}_1 \le \log V\|\boldsymbol{z}_1\|_2^2\frac{\sqrt{V}}{d}\Big] \le 1 - 2\exp(-c\log^2 V)$$

By using the first item above, we have

$$\mathbb{P}\Big[\Big|\boldsymbol{z}_1^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{z}_1 - 1 - \frac{V-1}{d}\Big| \ge 6\log V\Big(\frac{\sqrt{V}}{d} + \frac{V}{d^{3/2}}\Big)\Big] \le 1 - 2\exp(-c\log^2 V). \tag{43}$$

The result follows (42) and (43). ◻

E.3 Multinomial distribution and related statements

Lemma 5. *Let* $(n_1,\cdots,n_V) \in \mathrm{Mult}\big(N;(p_1,\cdots,p_V)\big)$. *For* $\boldsymbol{t}\in\mathbb{R}^V$,

$$\mathbb{E}\left[\exp\Big(\sum_{w=1}^V t_w n_w\Big)\right] = \Big(\sum_{w=1}^N p_w e^{t_w}\Big)^N.$$

Then, if $p_w = \frac{1}{V}$, $w\in[V]$, *we have*

- $\mathbb{E}\left[\prod_{w=1}^V (n_w)_{j_w}\right] = N(N-1)\cdots(N-J+1)\prod_{w=1}^V p_w^{j_w}$, *where* $J \coloneqq \sum_{w=1}^V j_w$.
- *We have*
  - $\mathbb{E}\left[n_w^2\right] = \frac{N}{V} + \frac{N(N-1)}{V^2}$
  - $\mathbb{E}\left[\big(\frac{n_w}{N} - \frac{1}{V}\big)^2 n_w\right] = \frac{(V-1)(N+V-2)}{NV^3}$.
  - $\mathbb{E}\left[n_w^3\right] = \frac{N}{V} + \frac{3N(N-1)}{V^2} + \frac{N(N-1)(N-2)}{V^3}$
  - $\mathbb{E}\left[n_w^4\right] = \frac{N}{V} + \frac{7N(N-1)}{V^2} + \frac{6N(N-1)(N-2)}{V^3} + \frac{N(N-1)(N-2)(N-3)}{V^4}$
  - $\mathbb{E}\left[n_w^2 n_{w'}^2\right] = \frac{N(N-1)}{V^2} + \frac{2N(N-1)(N-2)}{V^3} + \frac{N(N-1)(N-2)(N-3)}{V^4}$.
  - $\mathbb{E}\left[\big(\sum_{w=1}^V n_w^2\big)^2\right] = N^2 + \frac{(N+4)N(N-1)}{V} + \frac{(N+2)N(N-1)(N-2)}{V^2}$

Proof. Let $\boldsymbol{x}_i$ sampled from $\{\boldsymbol{e}_1,\cdots,\boldsymbol{e}_V\}$ with $(p_1,\cdots,p_V)$. We have $n_w = \sum_{i=1}^N \boldsymbol{e}_w^\top \boldsymbol{x}_i$. We have

$$\mathbb{E}\left[\exp\Big(\sum_{w=1}^V t_w n_w\Big)\right] = \mathbb{E}\left[\exp\Big(\sum_{i=1}^N \langle \boldsymbol{t},\boldsymbol{x}_i\rangle\Big)\right] = \Big(\mathbb{E}\left[\exp\big(\langle \boldsymbol{t},\boldsymbol{x}_1\rangle\big)\right]\Big)^N = \Big(\sum_{w=1}^N p_w e^{t_w}\Big)^N.$$

The later statement can be derived by using $z_w = e^{t_w}$ and taking derivatives of both sides with respect $(z_1,\cdots,z_V)$. ◻

Proposition 6. *Let* $\boldsymbol{n} = (n_1, \cdots, n_V) \in \mathrm{Mult}\big(L, \frac{1}{V}\mathbb{1}_V\big)$ *and* $\boldsymbol{S} \in \mathbb{R}^{V\times V}$ *be a symmetric matrix.. The following statements hold:*

- *We have*

$$\mathbb{E}[\mathrm{Diag}(\boldsymbol{n})\boldsymbol{S}\mathrm{diag}(\boldsymbol{n})] = L\,\mathbb{E}[\boldsymbol{x}_1^\top \boldsymbol{S}\boldsymbol{x}_1\boldsymbol{x}_1\boldsymbol{x}_1^\top] + \frac{L(L-1)}{V^2}\boldsymbol{S},$$

$$\mathbb{E}[\mathrm{Diag}(\boldsymbol{n} - \tfrac{L}{V}\mathbb{1}_V)\boldsymbol{S}\mathrm{Diag}(\boldsymbol{n} - \tfrac{L}{V}\mathbb{1}_V)] = L\,\mathbb{E}[\boldsymbol{x}_1^\top \boldsymbol{S}\boldsymbol{x}_1\boldsymbol{x}_1\boldsymbol{x}_1^\top] - \frac{L}{V^2}\boldsymbol{S}.$$

- *We have*

$$\mathbb{E}[\boldsymbol{n}\boldsymbol{n}^\top \boldsymbol{S}\boldsymbol{n}] = \frac{2L(L-1)}{V^2}\boldsymbol{S}\mathbb{1}_V + L\,\mathbb{E}\left[\boldsymbol{x}_1\boldsymbol{x}_1^\top \boldsymbol{S}\boldsymbol{x}_1\right] + \Big(\frac{L(L-1)}{V^2}\mathrm{tr}(\boldsymbol{S}) + \frac{L(L-1)(L-2)}{V^3}\mathbb{1}_V^\top \boldsymbol{S}\mathbb{1}_V\Big)\mathbb{1}_V$$

- *We have*

$$\mathbb{E}\left[\Big(\big(\boldsymbol{n} - \tfrac{L}{V}\mathbb{1}_V\big)^\top \boldsymbol{S}\big(\boldsymbol{n} - \tfrac{L}{V}\mathbb{1}_V\big)\Big)^2\right] = \frac{L}{V}\Big\|\mathrm{diag}(\boldsymbol{S}) - \frac{2}{V}\boldsymbol{S}\mathbb{1}_V + \frac{1}{V^2}(\mathbb{1}_V^\top \boldsymbol{S}\mathbb{1}_V)\mathbb{1}_V\Big\|_2^2 + \frac{L(L-1)}{V^2}\mathrm{tr}\Big(\big(\boldsymbol{I}_V - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top\big)\boldsymbol{S}\Big)^2 + \frac{2L(L-1)}{V^2}\mathrm{tr}\Big(\big(\boldsymbol{I}_V - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top\big)\boldsymbol{S}\big(\boldsymbol{I}_V - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top\big)\boldsymbol{S}\Big)$$

Proof. For the first item, we observe that

$$\boldsymbol{e}_j^\top\,\mathbb{E}[\mathrm{Diag}(\boldsymbol{n})\boldsymbol{S}\mathrm{diag}(\boldsymbol{n})]\boldsymbol{e}_i = \mathbb{E}[n_j n_i]\boldsymbol{S}_{ij} = \Big(\frac{L}{V}\delta_{ij} + \frac{L(L-1)}{V^2}\Big)\boldsymbol{S}_{ij},$$

from which the first equation follows. For the second equation,

$$\begin{aligned}\boldsymbol{e}_j^\top\,\mathbb{E}[\mathrm{diag}(\boldsymbol{n} - \tfrac{L}{V}\mathbb{1}_V)\boldsymbol{S}\mathrm{diag}(\boldsymbol{n} - \frac{L}{V}\mathbb{1}_V)]\boldsymbol{e}_i &= \mathbb{E}[(n_j - \frac{L}{V})(n_i - \frac{L}{V})]\boldsymbol{S}_{ij}\\ &= \Big(\frac{L}{V}\delta_{ij} - \frac{L}{V^2}\Big)\boldsymbol{S}_{ij}.\end{aligned}$$

For the second item, we have

$$\begin{aligned}(\mathbb{E}[\boldsymbol{n}\boldsymbol{n}^\top \boldsymbol{S}\boldsymbol{n}])_i &= \sum_{jk}\boldsymbol{S}_{jk}\,\mathbb{E}[n_i n_j n_k]\\ &= \frac{L(L-1)(L-2)}{V^3}(\sum_{i\neq j\neq k}\boldsymbol{S}_{jk}) + \Big(\frac{L(L-1)(L-2)}{V^3} + \frac{L(L-1)}{V^2}\Big)(2\sum_{i\neq k}\boldsymbol{S}_{ik} + \sum_{i\neq k}\boldsymbol{S}_{kk})\\ &\quad + \Big(\frac{L}{V} + \frac{3L(L-1)}{V^2} + \frac{L(L-1)(L-2)}{V^3}\Big)\boldsymbol{S}_{ii}\\ &= \frac{L}{V}\boldsymbol{S}_{ii} + \frac{L(L-1)}{V}\mathrm{tr}(\boldsymbol{S}) + \frac{2L(L-1)}{V}\sum_k \boldsymbol{S}_{ik} + \frac{L(L-1)(L-2)}{V^3}(\sum_{jk}\boldsymbol{S}_{jk}).\end{aligned}$$

For the third item, we have $\big(\boldsymbol{n} - \frac{L}{V}\mathbb{1}_V\big) =_d \sum_{\ell=1}^L(\xi_{1,\ell} - \frac{1}{V}\mathbb{1}_V)$ where the equality holds in distribution. For notational convenience, let

$$\gamma_{\ell u} \coloneqq (\xi_{1,\ell} - \frac{1}{V}\mathbb{1}_V)^\top \boldsymbol{S}(\xi_{1,u} - \frac{1}{V}\mathbb{1}_V).$$

Then,

$$\mathbb{E}\left[\Big(\big(\boldsymbol{n} - \frac{L}{V}\mathbb{1}_V\big)^\top \boldsymbol{S}\big(\boldsymbol{n} - \frac{L}{V}\mathbb{1}_V\big)\Big)^2\right] =_d \sum_{\ell,u}\sum_{\ell',u'}\mathbb{E}[\gamma_{\ell u}\gamma_{\ell' u'}]$$

By independence, only (ℓ, u, ℓ', u') where each index occur even times contribute. The possible cases are as follows:

- All four indices equal ($\ell = u = \ell' = u'$): There are L many terms here with contribution

$$\begin{aligned}\mathbb{E}[\gamma_{\ell\ell}\gamma_{\ell\ell}] &= \frac{1}{V}\Big\|\mathrm{diag}\Big((\boldsymbol{I}_V - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\boldsymbol{S}(\boldsymbol{I}_V - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\Big)\Big\|_2^2 \\ &= \frac{1}{V}\Big\|\mathrm{diag}(\boldsymbol{S}) - \frac{2}{V}\boldsymbol{S}\mathbb{1}_V + \frac{1}{V^2}(\mathbb{1}_V^\top\boldsymbol{S}\mathbb{1}_V)\mathbb{1}_V\Big\|_2^2.\end{aligned}$$

- Two distinct indices, both pairs diagonal ($\ell = u$ and $\ell' = u'$ and $\ell \neq \ell'$): There are $L(L-1)$ many terms here with contribution

$$\mathbb{E}[\gamma_{\ell\ell}\gamma_{\ell'\ell'}] = \mathbb{E}[\gamma_{\ell\ell}]\,\mathbb{E}[\gamma_{\ell'\ell'}] = \mathrm{tr}\Big((\boldsymbol{I}_V - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\boldsymbol{S}\Big)^2$$

- Two distinct indices, paired off-diagonal: ($\ell = \ell'$ and $u = u'$ and $\ell \neq u$): There are $2L(L-1)$ many terms here with contribution

$$\begin{aligned}\mathbb{E}[\gamma_{\ell u}\gamma_{\ell u}] &= \mathrm{tr}\Big(\,\mathbb{E}[(\xi_{1,1} - \frac{1}{V}\mathbb{1}_V)(\xi_{1,1} - \frac{1}{V}\mathbb{1}_V)^\top\boldsymbol{S}(\xi_{1,2} - \frac{1}{V}\mathbb{1}_V)(\xi_{1,2} - \frac{1}{V}\mathbb{1}_V)^\top S]\Big) \\ &= \mathrm{tr}\Big((\boldsymbol{I}_V - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\boldsymbol{S}(\boldsymbol{I}_V - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\boldsymbol{S}\Big)\end{aligned}$$

□

Proposition 7. *Let $V^3 \gg L$. There exists a universal $C > 0$ such that the following holds:*

- *Let $m_{ij} \coloneqq (1 - \frac{1}{V})\mathbb{1}_{i=j} + \frac{L}{V}$. For $K > 0$ and $p \geq \log V$,*

$$\mathbb{E}\Big[\Big|\frac{1}{L}\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{X}_j^\top\mathbb{1}_L - m_{ij}\Big|^p\Big]^{\frac{1}{p}} \leq C\Big(\frac{p^{\frac{3}{2}}}{\sqrt{V}} + \frac{p^2}{L}\Big)$$

$$\mathbb{P}\Big[\Big|\frac{1}{L}\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{X}_j^\top\mathbb{1}_L - m_{ij}\Big| \geq CK^2\frac{\log^2 V}{\sqrt{V}\wedge L}\Big] \leq \frac{1}{V^K}.$$

- *For $K > 0$ and $p \geq \log V$,*

$$\begin{aligned}\mathbb{E}\Big[\Big\|\frac{1}{NL}\sum_{i=1}^N\big(\boldsymbol{X}_i^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top\big)\mathbb{1}_L\mathbb{1}_L^\top\big(\boldsymbol{X}_i^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top\big)^\top - \frac{1}{V}(\boldsymbol{I} - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\Big\|_2^p\Big]^{\frac{1}{p}} \\ \leq C\left(\sqrt{\frac{p}{NV}} + \frac{p}{N}\Big(1 + \frac{p^2}{\sqrt{V}\wedge L}\Big)\right)\end{aligned}$$

$$\begin{aligned}\mathbb{P}\Big[\Big\|\frac{1}{NL}\sum_{i=1}^N\big(\boldsymbol{X}_i^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top\big)\mathbb{1}_L\mathbb{1}_L^\top\big(\boldsymbol{X}_i^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top\big)^\top - \frac{1}{V}(\boldsymbol{I} - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\Big\|_2 \\ > CK\log^2 V\Big(\frac{1}{\sqrt{NV}} + \frac{1}{N}\Big(1 + \frac{\log^2 V}{\sqrt{V}\wedge L}\Big)\Big)\Big] \leq \frac{1}{V^K}.\end{aligned}$$

Proof. For $i = j$ in the first item, we have

$$\frac{1}{L}\mathbb{1}_L^\top\boldsymbol{X}_i\boldsymbol{X}_i^\top\mathbb{1}_L = 1 + \frac{2}{L}\sum_{1\leq j<k\leq L}\mathbb{1}_{\xi_{i,j}=\xi_{i,k}} = 1 + \frac{(L-1)}{V} + \frac{2}{L}\sum_{k=2}^L\sum_{j=1}^{k-1}\mathbb{1}_{\xi_{i,j}=\xi_{i,k}} - \frac{1}{V}$$

Define

$$Y_k \coloneqq \sum_{j=1}^{k-1}\big(\mathbb{1}_{\xi_{i,j}=\xi_{i,k}} - \frac{1}{V}\big) \text{ and } \mathcal{F}_k \coloneqq \sigma(Y_1, \cdots, Y_k).$$

Given that

$$\sum_{j=1}^{k-1}\mathbb{1}_{\xi_{i,j}=\xi_{i,k}}|\xi_{i,k} \sim \mathrm{Binomial}(k-1, \frac{1}{V}) \Rightarrow \mathbb{E}[|Y_k|^p]^{\frac{1}{p}} \leq C(\sqrt{p}\sqrt{\frac{L}{V}} + p), \;\; p \geq \log V.$$

where we used Corollary 3. As for the quadratic variation

$$Q_L \coloneqq \sum_{k=1}^{L} \mathbb{E}[Y_k^2|\mathcal{F}_{k-1}] = \sum_{k=1}^{L} \frac{1}{V}\Big(\|\boldsymbol{X}_i^\top \mathbb{1}_{k-1}\|_2^2 - \frac{(k-1)^2}{V})\Big)$$
$$= \frac{1}{V}\sum_{k=1}^{L}\|(\boldsymbol{X}_i^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{k-1}^\top)\mathbb{1}_{k-1}\|_2^2$$

For $p \geq \log V$, by using triangle inequality,

$$\begin{aligned}
\mathbb{E}[|Q_L|^{\frac{p}{2}}]^{\frac{2}{p}} &\leq \frac{1}{V}\sum_{k=1}^{L}\mathbb{E}\left[\|(\boldsymbol{X}_i^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{k-1}^\top)\mathbb{1}_{k-1}\|_2^p\right]^{\frac{2}{p}} \\
&\leq \frac{1}{V}\sum_{k=1}^{V}(k-1)\,\mathbb{E}\left[\|\boldsymbol{X}_i^\top \mathbb{1}_{k-1}\|_p^p\right]^{\frac{2}{p}} + \sum_{k=V+1}^{L}\mathbb{E}\left[\|(\boldsymbol{X}_i^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{k-1}^\top)\mathbb{1}_{k-1}\|_p^p\right]^{\frac{2}{p}} \\
&\leq Cp^2\frac{1}{V}\sum_{k=1}^{L} k = Cp^2\frac{L^2}{V}
\end{aligned}$$

where we used Corollary 3. By Proposition 13, for $p \geq \log V$, we have

$$\mathbb{E}\left[\Big|\sum_{k=1}^{L} Y_k\Big|^p\right]^{\frac{1}{p}} \leq C\Big(p\sqrt{p}\frac{L}{\sqrt{V}} + p^2\Big).$$

By using $p = \log V$, we have

$$\mathbb{P}\left[\Big|\frac{1}{L}\sum_{k=1}^{L} Y_k\Big| > CeK^2\frac{\log^2 V}{\sqrt{V}\wedge L}\right] \leq \frac{1}{V^K}.$$

Hence, we have the $i = j$ case. For the $i \neq j$ case, we have

$$\frac{1}{L}\mathbb{1}_L^\top \boldsymbol{X}_j\boldsymbol{X}_i^\top \mathbb{1}_L = \frac{L}{V} + \frac{1}{L}\sum_{\ell=1}^{L}\sum_{k=1}^{L}\mathbb{1}_{\xi_{i,k}=\xi_{j,\ell}} - \frac{1}{V}$$

We redefine the martingale difference sequence as

$$Y_k \coloneqq \sum_{\ell=1}^{L}\mathbb{1}_{\xi_{i,k}=\xi_{j,\ell}} - \frac{1}{V}.$$

Conditioned on $\boldsymbol{X}_j$, we have $\{Y_1, \cdots, Y_L\}$ are i.i.d. and

$$\mathbb{E}[Y_k|\boldsymbol{X}_j] = 0 \;\text{ and }\; \mathbb{E}[Y_k^p|\boldsymbol{X}_j] = \frac{1}{V}\|(\boldsymbol{X}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\|_p^p$$

By Proposition 13, for $p \geq \log V$, we have

$$\mathbb{E}[|\frac{1}{L}\sum_{k=1}^{L} Y_k|^p]^{\frac{1}{p}} \leq C(\frac{\sqrt{p}}{\sqrt{V}} + \frac{p^{\frac{3}{2}}}{\sqrt{LV}} + \frac{p^2}{L}).$$

By using $p = \log V$, we have

$$\mathbb{P}\Big[\Big|\frac{1}{L}\mathbb{1}_L^\top \boldsymbol{X}_j\boldsymbol{X}_i^\top \mathbb{1}_L - \frac{L}{V}\Big| \geq \frac{CK^2\log^2 V}{\sqrt{V}\vee L}\Big] \leq \frac{1}{V^K}.$$

For the second item, we define

$$\boldsymbol{Y}_k \coloneqq \frac{1}{L}\big(\boldsymbol{X}_i^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top\big)\mathbb{1}_L\mathbb{1}_L^\top\big(\boldsymbol{X}_i^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top\big)^\top - \frac{1}{V}(\boldsymbol{I}_V - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)$$

and $\boldsymbol{Q}_N \coloneqq N\,\mathbb{E}[\boldsymbol{Y}_1^2]$. We have

$$\boldsymbol{Q}_N \preceq N\,\mathbb{E}\left[\Big\|\frac{1}{\sqrt{L}}\big(\boldsymbol{X}_1^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top\big)\mathbb{1}_L\Big\|_2^2\frac{1}{L}\big(\boldsymbol{X}_1^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top\big)\mathbb{1}_L\mathbb{1}_L^\top\big(\boldsymbol{X}_1^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top\big)^\top\right]$$

$$
\begin{aligned}
&= N\,\mathbb{E}\Big[(1-\frac{1}{V})\frac{1}{L}(\boldsymbol{X}_1^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\mathbb{1}_L^\top(\boldsymbol{X}_1^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)^\top\Big] \\
&+ N\,\mathbb{E}\Big[\Big(\Big\|\frac{1}{\sqrt{L}}(\boldsymbol{X}_1^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\Big\|_2^2 - (1-\frac{1}{V})\Big) \\
&\quad \times \Big(\frac{1}{L}(\boldsymbol{X}_1^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\mathbb{1}_L^\top(\boldsymbol{X}_1^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)^\top - \frac{1}{V}(\boldsymbol{I}_V - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top\Big)\Big] \\
&\preceq \frac{CN}{V}\boldsymbol{I}_V + \frac{1}{2}\boldsymbol{Q}_N
\end{aligned}
$$

Therefore, we have $\|\boldsymbol{Q}_N\|_2 \le \frac{CN}{V}$. Moreover, by using the first item,

$$
\mathbb{E}[\|\boldsymbol{Y}_k\|_2^p]^{\frac{1}{p}} \le \mathbb{E}\Big[\Big\|\frac{1}{\sqrt{L}}(\boldsymbol{X}_i^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L\|_2^{2p}\Big]^{\frac{1}{p}} \le 1 + C\Big(\frac{p^{\frac{3}{2}}}{\sqrt{V}} + \frac{p^2}{L}\Big)
$$

Therefore, by using Proposition 13, we have

$$
\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{i=1}^N \boldsymbol{Y}_i - \frac{1}{V}(\boldsymbol{I} - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\Big\|_2^p\Big] \\
&\quad \le C\Bigg(\sqrt{p\vee\log V}\,\sqrt{\frac{1}{NV}} + (p\vee\log V)N^{\frac{1}{p}-1}\Big(1 + \frac{p^{\frac{3}{2}}}{\sqrt{V}} + \frac{p^2}{L}\Big)\Bigg)
\end{aligned}
$$

By using $p = \log V$, we have

$$
\mathbb{P}\Big[\Big\|\frac{1}{N}\sum_{i=1}^N \boldsymbol{Y}_i - \frac{1}{V}(\boldsymbol{I} - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\Big\|_2 > CK\log^2 V\Big(\frac{1}{\sqrt{NV}} + \frac{1}{N}\Big(1 + \frac{\log^2 V}{\sqrt{V}\wedge L}\Big)\Big)\Big] \le \frac{1}{V^K}.
$$

□

Proposition 8. *We consider $\boldsymbol{S}_1$, $\boldsymbol{S}_2$ and $\boldsymbol{S}_3$ defined in (16), (17) and (18) in the regime $V^3 \gg N \gg V$ and $L \asymp V^\varepsilon$, $\varepsilon \in (0,1)$. For any $K > 0$ and $V \ge \Omega_{K,\varepsilon}(1)$, the following holds:*

1. *We have*

$$
\mathbb{P}\Bigg[\Big|\mathrm{tr}(\boldsymbol{S}_1) - \frac{1}{L^2}\Big(\frac{1}{N} + (1-\frac{1}{V})\frac{1}{V}\Big)\Big| > CK^2\frac{\log^2 V}{L^2N\sqrt{V}} \;\text{ or }\; \|\boldsymbol{S}_1\|_2 > \frac{e^2}{L^2V^2}\Bigg] \le \frac{2}{V^K}.
$$

2. *We have*

$$
\mathbb{P}\Bigg[\Big|\mathrm{tr}(\boldsymbol{S}_2) - (1-\frac{1}{V})^2\frac{L-1}{L^2N}\Big| > C\frac{K^{\frac{3}{2}}\log^3 V}{N\sqrt{LV}} \;\text{ or }\; \|\boldsymbol{S}_2\|_2 > C\frac{K^{\frac{3}{2}}\log^2 V}{NLV}\Bigg] \le \frac{4}{V^K}.
$$

3. *We have*

$$
\mathbb{P}\Big[\tfrac{-CK^2\log^2 V}{N\sqrt{V}}\tfrac{1}{V^2L^2}\mathbb{1}_V\mathbb{1}_V^\top \preceq \boldsymbol{S}_3 - \tfrac{1}{N}\tfrac{1}{V^2L^2}\mathbb{1}_V\mathbb{1}_V^\top \preceq \tfrac{CK^2\log^2 V}{N\sqrt{V}}\tfrac{1}{V^2L^2}\mathbb{1}_V\mathbb{1}_V^\top\Big] \le \frac{1}{V^K}.
$$

Proof. We define $n_i \coloneqq |\{j \le N \mid \boldsymbol{x}_j = \boldsymbol{e}_i\}|$. We have

$$
\mathrm{tr}(\boldsymbol{S}_1) = (1-\frac{1}{V})\frac{1}{L^2N^2}\sum_{i=1}^V n_i^2 \quad\text{and}\quad \|\boldsymbol{S}_1\|_2 \le \sup_{i\le N}\frac{n_i^2}{L^2N^2}
$$

By using Proposition 7 and Corollary 3, we have the first item. For the second item, we write

$$
\begin{aligned}
\boldsymbol{S}_2 &= \frac{(1-\frac{1}{V})}{L^2N^2}\sum_{j=1}^N (\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top \\
&+ \frac{2(1-\frac{1}{V})}{L^2N^2}\sum_{j<k}\big(\mathbb{1}_{\boldsymbol{x}_j=\boldsymbol{x}_k} - \tfrac{1}{V}\big)\mathrm{sym}\Big((\boldsymbol{N}_j^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\mathbb{1}_{L-1}^\top(\boldsymbol{N}_k^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big)
\end{aligned}
$$

$$=: \boldsymbol{S}_{21} + \boldsymbol{S}_{22}$$

We will analyze $\boldsymbol{S}_{21}$ and $\boldsymbol{S}_{22}$ separately. We start with $\boldsymbol{S}_{21}$. We have

$$\operatorname{tr}(\boldsymbol{S}_{21}) - (1-\frac{1}{V})^2\frac{L-1}{L^2N} = (1-\frac{1}{V})\frac{L-1}{L^2N^2}\sum_{j=1}^{N}\underbrace{\|\frac{1}{\sqrt{L-1}}(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\|_2^2 - (1-\frac{1}{V})}_{:=Y_i}.$$

We have $\mathbb{E}[Y_i^2] \le \frac{2}{V}$ and by the first item in Proposition 7

$$\mathbb{E}[|Y_i|^p]^{\frac{1}{p}} \le \frac{Cp^2}{\sqrt{V}\vee L}.$$

Therefore, by Proposition 13,

$$\mathbb{E}\left[\left|\operatorname{tr}(\boldsymbol{S}_{21}) - (1-\frac{1}{V})^2\frac{L-1}{L^2N}\right|^p\right]^{\frac{1}{p}} \le \frac{C}{LN^2}\Big(\sqrt{\frac{pN}{V}} + pN^{\frac{1}{p}}\frac{p^2}{\sqrt{V}\vee L}\Big)$$

By using $p = \log V$, we have

$$\mathbb{P}\Big[\Big|\operatorname{tr}(\boldsymbol{S}_{21}) - (1-\frac{1}{V})^2\frac{L-1}{L^2N}\Big| > C\frac{K\log^3 V}{LN\sqrt{NV}}\Big] \le \frac{1}{V^K}. \tag{44}$$

Moreover, by Proposition 7, we have

$$\mathbb{P}\Big[\Big\|\boldsymbol{S}_{21} - \big(1-\frac{1}{V}\big)\frac{L-1}{L^2N}\frac{1}{V}(\boldsymbol{I}_V - \mathbb{1}_V\mathbb{1}_V^\top)\Big\|_2 > C\frac{K\log^2 V}{LN}\Big(\frac{1}{\sqrt{NV}} + \frac{1}{N}(1+\frac{\log^2 V}{\sqrt{V}\wedge L})\Big)\Big] \le \frac{1}{V^K}. \tag{45}$$

As for $\boldsymbol{S}_{22}$, we have

$$\begin{aligned}
&\operatorname{tr}(\boldsymbol{S}_{22})\\
&= \frac{2(1-\frac{1}{V})}{L^2N^2}\sum_{k=2}^{N}\sum_{j=1}^{k-1}\big(\mathbb{1}_{\boldsymbol{x}_j=\boldsymbol{x}_k} - \tfrac{1}{V}\big)\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top(\boldsymbol{N}_k^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\\
&= \frac{2(1-\frac{1}{V})}{L^2N^2}\sum_{k=2}^{N}\sum_{j=1}^{k-1}\big(\mathbb{1}_{\boldsymbol{x}_j=\boldsymbol{x}_k} - \frac{1}{V}\big)\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\boldsymbol{N}_k^\top\mathbb{1}_{L-1}
\end{aligned}$$

We define $\mathcal{F}_k \coloneqq \sigma(\boldsymbol{N}_{1:k})$ and

$$Y_k \coloneqq (1-\frac{1}{V})\frac{2}{L^2N^2}\sum_{j=1}^{k-1}\big(\mathbb{1}_{\boldsymbol{x}_j=\boldsymbol{x}_k} - \frac{1}{V}\big)\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\boldsymbol{N}_k^\top\mathbb{1}_{L-1}$$

We have

$$\mathbb{E}[Y_k^2|\mathcal{F}_{k-1}] = (1-\frac{1}{V})^3\frac{4(L-1)}{L^4N^4}\frac{1}{V}\sum_{j=1}^{k-1}\|(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\|_2^2$$

Then,

$$\begin{aligned}
Q_N = \sum_{k=1}^{N}\mathbb{E}[Y_k^2|\mathcal{F}_{k-1}] &= (1-\frac{1}{V})^3\frac{4(L-1)}{L^4N^4}\frac{1}{V}\sum_{k=1}^{N}\sum_{j=1}^{k-1}\|(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\|_2^2\\
&= (1-\frac{1}{V})^3\frac{4(L-1)}{L^4N^4}\frac{1}{V}\sum_{k=1}^{N}(N-k)\|(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\|_2^2
\end{aligned}$$

Then, for $p \ge \log V$,

$$\mathbb{E}\left[|Q_N|^{\frac{p}{2}}\right]^{\frac{2}{p}} \le \frac{5}{L^3N^3V}\sum_{k=1}^{N}\mathbb{E}\left[\|\boldsymbol{N}_k^\top\mathbb{1}_{L-1}\|_2^p\right]^{\frac{2}{p}} \le \frac{5}{LN^3V}\sum_{k=1}^{N}\mathbb{E}\left[\|\boldsymbol{N}_k^\top\mathbb{1}_{L-1}\|_p^p\right]^{\frac{2}{p}} \le \frac{5p^2}{LN^2V} \tag{46}$$

By using Proposition 13, we show the following:

- To bound $\mathbb{E}[|Y_k|^p]^{\frac{1}{p}}$ for $p \geq \log V$, we first write

$$\begin{aligned}
&\mathbb{E}[|Y_k|^p|\boldsymbol{N}_{1:k}, \boldsymbol{x}_k]^{\frac{1}{p}} \\
&\leq \frac{C}{LN^2}\frac{\sqrt{p}}{\sqrt{V}}\Big(\sum_{j=1}^{k-1}\Big|\frac{1}{L-1}\Big\langle(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}, (\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\Big\rangle\Big|^2\Big)^{\frac{1}{2}} \\
&+ \frac{Cp}{LN^2}\Big(\sum_{j=1}^{k-1}\Big|\frac{1}{L-1}\Big\langle(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}, (\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\Big\rangle\Big|^p\Big)^{\frac{1}{p}}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}[|Y_k|^p]^{\frac{1}{p}} &\leq \frac{C}{LN^2}\Big(\frac{\sqrt{p}\sqrt{k}}{\sqrt{V}} + pk^{\frac{1}{p}}\Big) \\
&\quad\times \mathbb{E}\Big[\Big|\frac{1}{L-1}\Big\langle(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}, (\boldsymbol{N}_1^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\Big\rangle\Big|^p\Big]^{\frac{1}{p}} \\
&\leq \frac{C}{LN^2}\Big(\frac{\sqrt{p}\sqrt{k}}{\sqrt{V}} + pk^{\frac{1}{p}}\Big)\mathbb{E}\Big[\Big|\frac{1}{L-1}\Big\langle\boldsymbol{N}_k^\top\mathbb{1}_{L-1}, \boldsymbol{N}_1^\top\mathbb{1}_{L-1}\Big\rangle - \tfrac{L-1}{V}\Big|^p\Big]^{\frac{1}{p}} \\
&\leq \frac{Cp}{LN^2}\Big(\frac{\sqrt{p}\sqrt{k}}{\sqrt{V}} + pk^{\frac{1}{p}}\Big)\Big(\frac{1}{\sqrt{V}} \vee \frac{1}{L}\Big). \qquad (47)
\end{aligned}$$

- Then by using (46) and (47), we have for $p \geq \log V$

$$\mathbb{E}[|\mathrm{tr}(\boldsymbol{S}_{22})|^p]^{\frac{1}{p}} \leq C\Big(\frac{p^{\frac{3}{2}}}{N\sqrt{LV}} + \frac{p^3N^{\frac{2}{p}}}{LN^{\frac{3}{2}}\sqrt{V}}\Big(\frac{1}{\sqrt{V}} \vee \frac{1}{L}\Big)\Big) \leq \frac{Cp^{\frac{3}{2}}}{N\sqrt{LV}}.$$

Therefore, by using $p = \log V$,

$$\mathbb{P}\Big[|\mathrm{tr}(\boldsymbol{S}_{22})| > \frac{CK^{\frac{3}{2}}\log^{\frac{3}{2}} V}{N\sqrt{LV}}\Big] \leq \frac{1}{V^K} \qquad (48)$$

To bound $\|\boldsymbol{S}_{22}\|_2$, we define

$$\boldsymbol{Y}_k \coloneqq \sum_{j=1}^{k-1}(\mathbb{1}_{\boldsymbol{x}_j=\boldsymbol{x}_k} - \frac{1}{V})\mathrm{sym}\Big((\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\mathbb{1}_{L-1}^\top(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big).$$

We have

$$\begin{aligned}
&\mathbb{E}[\boldsymbol{Y}_k^2|\mathcal{F}_{k-1}] \\
&\preceq \frac{2}{V}\sum_{j=1}^{k-1}\mathbb{E}\Big[(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\mathbb{1}_{L-1}^\top(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big\|(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\Big\|_2^2\Big|\mathcal{F}_{k-1}\Big] \\
&+ \frac{2}{V}\sum_{j=1}^{k-1}\mathbb{E}\Big[\Big\|(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\Big\|_2^2(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big|\mathcal{F}_{k-1}\Big] \\
&\preceq \frac{2L}{V^2}\sum_{j=1}^{k-1}\Big\|(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\Big\|_2^2\boldsymbol{I}_V + \frac{2L}{V}\sum_{j=1}^{k-1}(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top
\end{aligned}$$

Therefore, we have

$$\boldsymbol{Q}_N \coloneqq \sum_{k=1}^{N}\mathbb{E}[\boldsymbol{Y}_k^2|\mathcal{F}_{k-1}]$$

$$\preceq \frac{2NL}{V^2}\sum_{j=1}^{N-1}\Big\|(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\Big\|_2^2 \boldsymbol{I}_V$$
$$+ \frac{2L^2N^2}{V}\frac{1}{LN}\sum_{j=1}^{N-1}(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top.$$

Then,

$$\mathbb{E}[\|\boldsymbol{Q}_N\|_2^{\frac{p}{2}}]^{\frac{2}{p}} \le \frac{2NL}{V^2}\,\mathbb{E}\Big[\Big(\sum_{j=1}^{N-1}\Big\|(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\Big\|_2^2\Big)^{\frac{p}{2}}\Big]^{\frac{2}{p}}$$
$$+ \frac{2L^2N^2}{V}\,\mathbb{E}\Big[\Big\|\frac{1}{N(L-1)}\sum_{j=1}^{N-1}(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big\|_2^{\frac{p}{2}}\Big]^{\frac{2}{p}}$$
$$\le \frac{2N^2L^2}{V^2}\,\mathbb{E}\Big[\Big\|\frac{1}{\sqrt{L}}(\boldsymbol{N}_1^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\Big\|_2^{p}\Big]^{\frac{2}{p}}$$
$$+ \frac{2L^2N^2}{V}\,\mathbb{E}\Big[\Big\|\frac{1}{N(L-1)}\sum_{j=1}^{N-1}(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\mathbb{1}_{L-1}^\top(\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big\|_2^{\frac{p}{2}}\Big]^{\frac{2}{p}}$$
$$\le \frac{CN^2L^2}{V^2}\Big(1+\frac{p^2}{\sqrt{V}\vee L}\Big) + \frac{CL^2N^2}{V}\Big(\frac{1}{V} + \sqrt{\frac{p}{NV}} + \frac{p}{N}\Big(1+\frac{p^2}{\sqrt{V}\wedge L}\Big)\Big)$$
$$\le \frac{CN^2L^2}{V^2}\Big(1+\frac{p^2}{\sqrt{V}\vee L}\Big).$$

To bound $\mathbb{E}[\|\boldsymbol{Y}_k\|_2^p]$, we observe that

- We have

$$\mathbb{E}\Big[\Big((\mathbb{1}_{\boldsymbol{x}_j=\boldsymbol{x}_k} - \tfrac{1}{V})\mathrm{sym}\Big((\boldsymbol{N}_j^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\mathbb{1}_{L-1}^\top(\boldsymbol{N}_k^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big)\Big)^2\Big|\boldsymbol{x}_k, \boldsymbol{N}_k\Big]$$
$$\preceq \tfrac{L}{V^2}\|(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\|_2^2\boldsymbol{I}_V$$
$$+ \frac{L}{V}(\boldsymbol{N}_k^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\mathbb{1}_{L-1}^\top(\boldsymbol{N}_k^\top - \tfrac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top$$

  Moreover,

$$\mathbb{E}\Big[\|(\mathbb{1}_{\boldsymbol{x}_j=\boldsymbol{x}_k} - \tfrac{1}{V})\mathrm{sym}\Big((\boldsymbol{N}_j^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\mathbb{1}_{L-1}^\top(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)^\top\Big)\|^p\Big|\boldsymbol{x}_k, \boldsymbol{N}_k\Big]^{\frac{1}{p}}$$

$$= \mathbb{E}\Big[|(\mathbb{1}_{\boldsymbol{x}_j=\boldsymbol{x}_k} - \frac{1}{V})|^p|\boldsymbol{x}_k\Big]^{\frac{1}{p}}\,\mathbb{E}\Big[\Big|\Big\langle \boldsymbol{N}_j^\top\mathbb{1}_{L-1}, (\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\Big\rangle\Big|^p|\boldsymbol{N}_k\Big]^{\frac{1}{p}}$$
$$\le C(\sqrt{\frac{p}{V}} + p)\Big(\sqrt{\frac{p}{V}}\|(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\|_2 + p\|(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\|_p\Big)$$

- By Proposition 13, we have

$$\mathbb{E}[\|\boldsymbol{Y}_k\|_2^p|\boldsymbol{x}_k, \boldsymbol{N}_k]^{\frac{1}{p}} \le C\sqrt{p\vee\log V}\sqrt{\frac{Lk}{V}}\|(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\|_2$$
$$+ C(p\vee\log V)k^{\frac{1}{p}}\frac{p^{3/2}}{V}\|(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\|_2$$
$$+ C(p\vee\log V)k^{\frac{1}{p}}p^2\|(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\|_p$$

  Therefore, for $p \ge \log V$

$$\mathbb{E}[\|\boldsymbol{Y}_k\|_2^p]^{\frac{1}{p}} \le C\sqrt{\frac{pLk}{V}}\,\mathbb{E}[\|(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\|_2^p]^{\frac{1}{p}}$$

$$+ Ck^{\frac{1}{p}} \frac{p^{5/2}}{\sqrt{V}} \mathbb{E}[\|(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V \mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\|_2^p]^{\frac{1}{p}}$$

$$+ Ck^{\frac{1}{p}} p^3 \mathbb{E}[\|(\boldsymbol{N}_k^\top - \frac{1}{V}\mathbb{1}_V \mathbb{1}_{L-1}^\top)\mathbb{1}_{L-1}\|_p^p]^{\frac{1}{p}}$$

$$\le C\Big(\sqrt{\frac{pLk}{V}} + \frac{p^{5/2}}{\sqrt{V}}\Big)(\sqrt{L} + p\sqrt{\frac{pL}{V}} + \frac{p^2}{\sqrt{L}}) + Cp^3(\sqrt{\frac{pL}{V}} + p)$$

Therefore, for $p = \log V$, we have

$$\mathbb{E}[\|\boldsymbol{S}_{22}\|_2^p] \le C\Big(\frac{\sqrt{p}}{NLV} + \frac{p^{3/2}}{LN\sqrt{NV}} + \frac{p^5}{L^2N^2}\Big)$$

Therefore, we have

$$\mathbb{P}\Big[\|\boldsymbol{S}_{22}\|_2 > C\frac{K^{3/2}\log^{3/2} V}{NLV}\Big] \le \frac{1}{V^K}. \tag{49}$$

By (44) , (45), (48), and (49), we have the second item. For the last item, we have

$$\boldsymbol{S}_3 - \frac{1}{N}\frac{1}{V^2L^2}\mathbb{1}_V\mathbb{1}_V^\top = \frac{1}{V^2L^2}\mathbb{1}_V\mathbb{1}_V^\top\left(\Big\|\frac{1}{N}\sum_{j=1}^N(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)\Big\|_2^2 - \frac{1}{N}\right)$$

By Proposition 7,

$$\mathbb{P}\Big[\ \Big|\Big\|\frac{1}{N}\sum_{j=1}^N(\boldsymbol{x}_j - \frac{1}{V}\mathbb{1}_V)\Big\|_2^2 - \frac{1}{N}\Big| > \frac{CK^2\log^2 V}{N\sqrt{V}}\ \Big] \le \frac{1}{V^K}.$$

The displayed equation implies the third item. □

Proposition 9. *Let* $\mathsf{z}_{\nu,\delta} = \mathsf{z}_\nu + \mathbb{1}_{k=1}\delta\mathsf{z}_{\text{trig}}$. *Given that (E.1) holds, the following statements hold:*

1. We have for $i \neq r$,

$$\Big|\mathbb{E}\left[(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\text{in}}\right]\Big| \le \frac{C}{Vd}.$$

2. We have

$$\mathbb{E}[\|\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}\|_2^2|\mathsf{Z}_{\text{in}}] \le C(\frac{L}{d} + \frac{L^2}{d^2}).$$

3. We have for $i \neq r$,

$$\Big|\mathbb{E}\left[(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r} - \frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\text{in}}\right]\Big| \le \frac{C\log^2 V}{d^2}.$$

4. We have

$$\mathbb{E}\left[\left(\mathbb{1}_V^\top\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\text{in}}^\top\mathsf{z}_{\nu,\delta}\right)^2|\mathsf{Z}_{\text{in}}\right] \le C\frac{V\log^2 V}{d}\Big(\frac{L}{d} + \frac{L^2}{d^2}\Big).$$

5. For notational convenience, let

$$\gamma := \mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\text{in}}\Big(\boldsymbol{X}_i^\top\boldsymbol{X}_i - \frac{L}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\text{in}}^\top\boldsymbol{Z}_{\text{in}}(\boldsymbol{X}_i^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top)\mathbb{1}_L.$$

We have

$$|\mathbb{E}[\gamma|\mathsf{Z}_{\text{in}}]| \le \frac{CL\log V}{\sqrt{Vd}} \ \textit{ and } \ \mathbb{E}[\gamma^2|\mathsf{Z}_{\text{in}}] \le C\log^2 V\Big(\frac{L}{d} + \frac{L^2}{d^2}\Big).$$

Proof. For the first item, we have

$$\mathbb{E}\left[(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r}-\frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\rm in}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\rm in}\right]$$
$$=\mathbb{E}\left[(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_u}-\frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{x}_r\boldsymbol{x}_r^\top\boldsymbol{Z}_{\rm in}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\rm in}\right]$$
$$=\frac{1}{V}\mathbb{E}\left[\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{Z}_{\rm in}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\rm in}\right]-\frac{1}{V^3}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{Z}_{\rm in}^\top\mathsf{z}_{\nu,\delta}$$
$$\le\frac{C}{Vd}$$

For the second item, we write

$$\mathbb{E}[\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\rm in}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\rm in}]$$
$$=\frac{L}{V}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\mathrm{diag}(\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in})\boldsymbol{Z}_{\rm in}^\top\mathsf{z}_{\nu,\delta}+\frac{L(L-1)}{V^2}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{Z}_{\rm in}^\top\mathsf{z}_{\nu,\delta}^\top\le C\Big(\frac{L}{d}+\frac{L^2}{d^2}\Big)$$

For the third item, we have

$$\mathbb{E}\left[(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r}-\frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_r^\top\boldsymbol{X}_r\boldsymbol{Z}_{\rm in}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\rm in}\right]$$
$$=\mathbb{E}\left[(\mathbb{1}_{\boldsymbol{x}_i=\boldsymbol{x}_r}-\frac{1}{V})\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{x}_r\boldsymbol{x}_r^\top\boldsymbol{Z}_{\rm in}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\rm in}\right]$$
$$=\frac{1}{V}\mathbb{E}\left[\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{Z}_{\rm in}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\rm in}\right]$$
$$-\frac{1}{V^3}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{Z}_{\rm in}^\top\mathsf{z}_{\nu,\delta}$$
$$\le C\frac{\log^2 V}{d^2}$$

For the fourth item, we have

$$\mathbb{E}\left[\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\rm in}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\rm in}\right]$$
$$=L\ \mathbb{E}\left[\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{Z}_{\rm in}^\top\mathsf{z}_{\nu,\delta}|\mathsf{Z}_{\rm in}\right]$$
$$+\frac{L(L-1)}{V^2}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V\mathbb{1}_V^\top\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{Z}_{\rm in}^\top\mathsf{z}_{\nu,\delta}$$
$$=C\Big(\frac{LV\log^2 V}{d^2}+\frac{L^2V\log^2 V}{d^3}\Big)$$

For the fifth item, we have

$$\mathbb{E}\left[\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\Big(\boldsymbol{X}_i^\top\boldsymbol{X}_i-\frac{L}{V}\boldsymbol{I}_V\Big)\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\mathbb{1}_L|\mathsf{Z}_{\rm in}\right]$$
$$=\mathbb{E}\left[\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\mathbb{1}_L|\mathsf{Z}_{\rm in}\right]-\frac{L^2}{V^2}(\mathsf{z}_k+\mathbb{1}_{k=1}\delta\mathsf{z}_{\rm trig})^\top\boldsymbol{Z}_{\rm in}\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V$$
$$=\mathbb{E}\left[\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\boldsymbol{X}_i\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{X}_i^\top\boldsymbol{X}_i|\mathsf{Z}_{\rm in}\right]\mathbb{1}_V-\frac{L^2}{V^2}(\mathsf{z}_k+\mathbb{1}_{k=1}\delta\mathsf{z}_{\rm trig})^\top\boldsymbol{Z}_{\rm in}\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V$$
$$=L\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{x}_i\boldsymbol{x}_i^\top|\mathsf{Z}_{\rm in}\right]\mathbb{1}_V-\frac{L}{V^2}\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\mathbb{1}_V$$
$$=\frac{CL\log V}{\sqrt{Vd}}.$$

For the second part, let $c_i := \boldsymbol{e}_i^\top\boldsymbol{X}_i\mathbb{1}_L$. We have

$$\mathsf{z}_{\nu,\delta}^\top\boldsymbol{Z}_{\rm in}\big(\boldsymbol{X}_i^\top\boldsymbol{X}_i-\frac{L}{V}\boldsymbol{I}_V\big)\boldsymbol{Z}_{\rm in}^\top\boldsymbol{Z}_{\rm in}\big(\boldsymbol{X}_i^\top-\frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top\big)\mathbb{1}_L$$
$$=\sum_{i=1}^V(c_i-\frac{L}{V})\mathsf{z}_{\nu,\delta}^\top\mathsf{z}_i\mathsf{z}_i^\top\Big(\sum_{j=1}^V(c_j-\frac{L}{V})\mathsf{z}_j)\Big)$$

$$= \sum_{i=1}^{V}\sum_{j=1}^{V}(c_i - \frac{L}{V})(c_j - \frac{L}{V})\mathsf{z}_{\nu,\delta}^\top \mathsf{z}_i \mathsf{z}_i^\top \mathsf{z}_j$$

Let $\boldsymbol{S} = (s_{ij})_{ij\in[V]}$ such that $s_{ij} \coloneqq \frac{1}{2}\big(\mathsf{z}_{\nu,\delta}^\top \mathsf{z}_i \mathsf{z}_i^\top \mathsf{z}_j + \mathsf{z}_{\nu,\delta}^\top \mathsf{z}_j \mathsf{z}_j^\top \mathsf{z}_i\big)$.

- We have

$$\begin{aligned}\Big|\mathrm{tr}\big((\boldsymbol{I}_V - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\boldsymbol{S}\big)\Big| &= \Big|\mathrm{tr}(\boldsymbol{S}) - \frac{1}{V}\mathbb{1}_V^\top \boldsymbol{S}\mathbb{1}_V\Big| \\ &= V\Big|\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\,\mathbb{E}[\boldsymbol{x}_1\boldsymbol{x}_1^\top \boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{x}_1] - \frac{1}{V}\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{Z}_{\mathrm{in}}\mathbb{1}_V\Big| \\ &\le \frac{C\log V\sqrt{V}}{\sqrt{d}}.\end{aligned}$$

- Moreover,

$$\mathrm{tr}\big((\boldsymbol{I}_V - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\boldsymbol{S}(\boldsymbol{I}_V - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\boldsymbol{S}\big) = \mathrm{tr}(\boldsymbol{S}^2) - \frac{2}{V}\|\boldsymbol{S}\mathbb{1}_V\|_2^2 + \frac{1}{V^2}\big(\mathbb{1}_V^\top \boldsymbol{S}\mathbb{1}_V\big)^2.$$

  We have $\mathrm{tr}(\boldsymbol{S}^2) \le \frac{CV^2\log^2 V}{d^2}$ and

$$\boldsymbol{e}_i^\top \boldsymbol{S}\mathbb{1}_V = \frac{1}{2}\sum_{j=1}^{V}\mathsf{z}_{\nu,\delta}^\top \mathsf{z}_i\mathsf{z}_i^\top \mathsf{z}_j + \frac{1}{2}\sum_{j=1}^{V}\mathsf{z}_{\nu,\delta}^\top \mathsf{z}_j\mathsf{z}_j^\top \mathsf{z}_i = \mathsf{z}_{\nu,\delta}^\top \mathsf{z}_i\mathsf{z}_i^\top \boldsymbol{Z}_{\mathrm{in}}\mathbb{1}_V + \mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\boldsymbol{Z}_{\mathrm{in}}^\top \mathsf{z}_i.$$

  Therefore,

$$\Big|\mathrm{tr}\big((\boldsymbol{I}_V - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\boldsymbol{S}(\boldsymbol{I}_V - \frac{1}{V}\mathbb{1}_V\mathbb{1}_V^\top)\boldsymbol{S}\big)\Big| \le \frac{CV^2\log^2 V}{d^2}.$$

- Moreover, $\|\mathrm{diag}(\boldsymbol{S})\|_2^2 \le \frac{CV\log^2 V}{d}$.

Therefore, by Proposition 6, we have

$$\mathbb{E}\left[\Big(\mathsf{z}_{\nu,\delta}^\top \boldsymbol{Z}_{\mathrm{in}}\big(\boldsymbol{X}_i^\top \boldsymbol{X}_i - \frac{L}{V}\boldsymbol{I}_V\big)\boldsymbol{Z}_{\mathrm{in}}^\top \boldsymbol{Z}_{\mathrm{in}}\big(\boldsymbol{X}_i^\top - \frac{1}{V}\mathbb{1}_V\mathbb{1}_L^\top\big)\mathbb{1}_L\Big)^2 \Big| \mathsf{Z}_{\mathrm{in}}\right] \le C\log^2 V\Big(\frac{L}{d} + \frac{L^2}{d^2}\Big).$$

□

F MISCELLANEOUS

Proposition 10. *Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{B} \in \mathbb{R}^{V\times V}$. Let $\boldsymbol{M} \coloneqq \boldsymbol{A}\otimes\boldsymbol{B}$. We have*

$$\|\boldsymbol{M}\|_2 = \|\boldsymbol{A}\|_2\|\boldsymbol{B}\|_2 \;\; \text{and} \;\; \|\boldsymbol{M}\|_F = \|\boldsymbol{A}\|_F\|\boldsymbol{B}\|_F \;\; \text{and} \;\; \mathrm{tr}(\boldsymbol{M}) = \mathrm{tr}(\boldsymbol{A})\mathrm{tr}(\boldsymbol{B}).$$

Proof. The Frobenius norm and trace are straightforward. For the ℓ_2 norm, let $\boldsymbol{A} =: \sum_{i=1}^{d}\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^\top$ and $\boldsymbol{B} =: \sum_{j=1}^{V}\tilde{\sigma}_j\tilde{\boldsymbol{u}}_j\tilde{\boldsymbol{v}}_j^\top$. We have

$$\boldsymbol{M} = \sum_{i=1}^{d}\sum_{j=1}^{V}\sigma_i\tilde{\sigma}_j(\boldsymbol{u}_i\boldsymbol{v}_i^\top)\otimes(\tilde{\boldsymbol{u}}_j\tilde{\boldsymbol{v}}_j^\top) = \sum_{i=1}^{d}\sum_{j=1}^{V}\sigma_i\tilde{\sigma}_j(\boldsymbol{u}_i\otimes\tilde{\boldsymbol{u}}_j)(\boldsymbol{v}_i\otimes\tilde{\boldsymbol{v}}_j)^\top.$$

For any $(i,j)\neq(i',j')$, we have

$$(\boldsymbol{u}_i\otimes\tilde{\boldsymbol{u}}_j)^\top(\boldsymbol{u}_{i'}\otimes\tilde{\boldsymbol{u}}_{j'}) = (\boldsymbol{v}_i\otimes\tilde{\boldsymbol{v}}_j)^\top(\boldsymbol{v}_{i'}\otimes\tilde{\boldsymbol{v}}_{j'}) = 0.$$

Therefore,

$$\|\boldsymbol{M}\|_2 = \max_{i,j}\sigma_i\tilde{\sigma}_j = \max_i\sigma_i\max_j\tilde{\sigma}_j.$$

□

Proposition 11. *Let $z \sim \mathcal{N}(0, I_d)$ and $P_k : \mathbb{R}^d \to [0, \infty)$ denotes a degree k polynomial which takes nonnegative values. For $p \geq 1$, we have*

$$\mathbb{E}[|P_k(\boldsymbol{z})|^p]^{\frac{1}{p}} \leq \big(8(p-1)\big)^{\frac{k}{2}}\, \mathbb{E}[P_k(\boldsymbol{z})].$$

Proof. By hypercontractivity, it is sufficient to prove that $\frac{\mathbb{E}[|P_k(\boldsymbol{z})^2]^{\frac{1}{2}}}{\mathbb{E}[P_k(\boldsymbol{z})]} \leq 8^{\frac{k}{2}}$. We have

$$\mathbb{E}[|P_k(\boldsymbol{z})^2]^2 \leq \mathbb{E}[|P_k(\boldsymbol{z})]\, \mathbb{E}[|P_k(\boldsymbol{z})^3] \leq 2^{\frac{3k}{2}}\, \mathbb{E}[|P_k(\boldsymbol{z})]\, \mathbb{E}[|P_k(\boldsymbol{z})^2]^{\frac{3}{2}}$$

which proves the result. □

Proposition 12. *Let $k \in \mathbb{N}$ and $\boldsymbol{w} \sim N(0, \boldsymbol{I}_d)$. For $L > 0$ and $\boldsymbol{u}, \boldsymbol{v} \in S^{d-1}$, we have*

$$\mathbb{E}\left[H_{e_k}\Big(\frac{1}{\sqrt{L}}\boldsymbol{w}^\top \boldsymbol{u}\Big) H_{e_k}\Big(\frac{1}{\sqrt{L}}\boldsymbol{w}^\top \boldsymbol{v}\Big)\right] = \frac{k!}{L^k}\sum_{i=0}^{\lfloor k/2 \rfloor} \frac{(2i-1)!!}{2i!!}\binom{k}{2i}(L-1)^{2i}\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-2i}$$

Proof. For $a \in \mathbb{R}$, we have

$$H_{e_k}(ax) = \sum_{i=0}^{\lfloor k/2 \rfloor} \frac{k!}{2^i i!(k-2i)!}(a^2-1)^i a^{k-2i} H_{e_{k-2i}}(x)$$

Therefore, for $a = 1/\sqrt{L}$, we have

$$\begin{aligned}
&\mathbb{E}\left[H_{e_k}\Big(\frac{1}{\sqrt{L}}\boldsymbol{w}^\top \boldsymbol{u}\Big) H_{e_k}\Big(\frac{1}{\sqrt{L}}\boldsymbol{w}^\top \boldsymbol{v}\Big)\right] \\
&= \mathbb{E}\Bigg[\Big(\sum_{i=0}^{\lfloor k/2 \rfloor} \frac{k!}{2^i i!(k-2i)!}(a^2-1)^i a^{k-2i} H_{e_{k-2i}}(\boldsymbol{w}^\top \boldsymbol{u})\Big) \\
&\quad \times \Big(\sum_{i=0}^{\lfloor k/2 \rfloor} \frac{k!}{2^i i!(k-2i)!}(a^2-1)^i a^{k-2i} H_{e_{k-2i}}(\boldsymbol{w}^\top \boldsymbol{v})\Big)\Bigg] \\
&= \sum_{i=0}^{\lfloor k/2 \rfloor}\Big(\frac{k!}{2^i i!(k-2i)!}\Big)^2 (a^2-1)^{2i} a^{2(k-2i)}(k-2i)!\,\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-2i} \\
&= \frac{k!}{L^k}\sum_{i=0}^{\lfloor k/2 \rfloor} \frac{(2i-1)!!}{2i!!}\binom{k}{2i}(L-1)^{2i}\langle \boldsymbol{u}, \boldsymbol{v}\rangle^{k-2i}
\end{aligned}$$

□

F.1 ROSENTHAL-BURKHOLDER INEQUALITY AND COROLLARIES

We will rely on the following inequality:

Proposition 13 ((Peng et al., 2025, Theorem 2.1))**.** *Let $\{\boldsymbol{M}_k\}_{k=1}^N$ be a d-dimensional symmetric matrix valued martingale adapted to the filtration $\{\mathcal{F}_k\}_{k=0}^N$. Let $\boldsymbol{Y}_k \coloneqq \boldsymbol{M}_k - \boldsymbol{M}_{k-1}$ be its corresponding difference sequence and the quadratic variation is defined as*

$$\boldsymbol{Q}_N \coloneqq \sum_{k=1}^N \mathbb{E}[\boldsymbol{Y}_k^2 | \mathcal{F}_{k-1}].$$

For any $p \geq 2$, suppose

$$\mathbb{E}\left[\|\boldsymbol{Q}_N\|_2^{\frac{p}{2}}\right]^{\frac{1}{p}} < \infty \ \ \textit{and} \ \ \sup_{k \in [N]} \mathbb{E}\left[\|\boldsymbol{Y}_k\|_2^p\right]^{\frac{1}{p}} < \infty.$$

Then it holds that

$$\mathbb{E}\left[\|\boldsymbol{M}_N\|_2^p\right]^{\frac{1}{p}} \leq C\Big(\sqrt{p \vee \log d}\ \mathbb{E}\left[\|\boldsymbol{Q}_N\|_2^{\frac{p}{2}}\right]^{\frac{1}{p}} + (p \vee \log d) N^{\frac{1}{p}} \sup_{k \in [N]} \mathbb{E}\left[\|\boldsymbol{Y}_k\|_2^p\right]^{\frac{1}{p}}\Big).$$

We have the following corollaries:

Corollary 3. *The following statements holds for general $L, V > 0$:*

1. *For $X \sim \mathrm{Binomial}(L, \frac{1}{V})$, we have*
$$\mathbb{E}[|X - kq|^p]^{\frac{1}{p}} \leq C\Big(\sqrt{p}\sqrt{\frac{L}{V}} + p\Big(\frac{L}{V}\Big)^{\frac{1}{p}}\Big).$$

2. *Let $\boldsymbol{c} = (c_1, \cdots, c_V) \sim \mathrm{Multinomial}(L, \frac{1}{V}\mathbb{1}_V)$. For $p \geq 1$, we have*
$$\mathbb{E}[\|\boldsymbol{c}\|_p^p] \leq C^p V\left(\Big(\frac{L}{V}\Big)^p + \Big(\frac{pL}{V}\Big)^{\frac{p}{2}} + p^p\frac{L}{V}\right).$$

3. *By following the notation in the second item,*
   - *If $V \gg L$, we have for $L \geq e^{2e} + 1$,*
$$\mathbb{P}\left[\|\boldsymbol{c}\|_\infty \geq \log L\right] \leq \Big(\frac{2e}{\log L - 1}\Big)^{\frac{\log L - 1}{2}}\Big(\frac{L}{V}\Big)^{\log L - 2}$$
   - *If $L \gg V$, we have*
$$\mathbb{P}\left[\|\boldsymbol{c}\|_\infty \geq \frac{eL}{V}\right] \leq 2Ve^{-L/V}.$$

Proof. The first two items are direct consequence of Proposition 13. For the third item, using $\mathbb{1}_{c_w \geq k} \leq \frac{c_w(c_w-1)\cdots(c_w-k+1)}{k!}$ and linearity of expectation
$$\begin{aligned}\mathbb{P}[\|\boldsymbol{c}\|_\infty \geq k] \leq \sum_{w=1}^{V}\mathbb{P}[c_w \geq k] &\leq \sum_{w=1}^{V}\frac{\mathbb{E}[c_w(c_w-1)\cdots(c_w-k+1)]}{k!}\\ &= \frac{L(L-1)\cdots(L-k+1)}{k!V^{k-1}}.\end{aligned}$$
For $V \gg L$, by choosing $k = \lfloor \log L \rfloor$, the result follows. For $L \gg V$, by choosing $k = \lfloor \frac{eL}{V} \rfloor$, the result follows. $\square$