Concentration of polynomial random matrices via Efron-Stein inequalities

Goutham Rajendran* Madhur Tulsiani †

Abstract

Analyzing concentration of large random matrices is a common task in a wide variety of fields. Given independent random variables, several tools are available to bound the norms of random matrices whose entries are linear in the variables, such as the matrix-Bernstein inequality. However, for many recent applications, we need to bound the norms of random matrices whose entries are polynomials in the variables. Such matrices arise naturally in the analysis of spectral algorithms (e.g., Hopkins et al. [STOC 2016], Moitra and Wein [STOC 2019]), and in lower bounds for semidefinite programs based on the Sum-of-Squares (SoS) hierarchy (e.g. Barak et al. [FOCS 2016], Jones et al. [FOCS 2021]).

In this work, we present a general framework to obtain such bounds, based on the beautiful matrix Efron-Stein inequalities developed by Paulin, Mackey and Tropp [Annals of Probability 2016]. The Efron-Stein inequality bounds the norm of a random matrix by the norm of another potentially simpler (but still random) matrix. We view the latter matrix as arising by "differentiating" the starting matrix. By recursively differentiating, our framework reduces the main task to bounding the norms of far simpler matrices. These simpler matrices are in fact deterministic matrices in the case of Rademacher random variables and hence, bounding their norm is a far easier task. In general for non-Rademacher random variables, the task reduces to the much easier task of scalar concentration. Moreover, in the setting of polynomial matrices, our main result also generalizes the work of Paulin, Mackey and Tropp.

As applications of our basic framework, we recover known bounds in the literature, especially for simple "tensor networks" and "dense graph matrices". As applications of our general framework, we derive bounds for "sparse graph matrices". The sparse graph matrix bounds were obtained only recently by Jones et al. [FOCS 2021] using a nontrivial application of the trace power method, and was a core component in their work. We expect this framework will also be helpful for other applications involving concentration phenomena for nonlinear random matrices.

1 Introduction

In optimization, statistics, and spectral algorithms, we often want to understand the concentration of various random matrices. To do this, we can appeal to the powerful theory of matrix-deviation inequalities [Tro15]. For example, the matrix-Bernstein inequality addresses random matrices of the form

$$\mathbf{M} \;=\; x_1 \cdot \mathbf{C_1} + \cdots + x_n \cdot \mathbf{C_n}$$

where $x_1, \ldots, x_n$ are independent scalar random variables, and $\mathbf{C_1}, \ldots, \mathbf{C_n}$ are fixed matrices. A large selection of such inequalities are available when the random matrix (say) $\mathbf{M}$ is a *linear* function of independent random variables. However, several recent works require us to understand random matrices which are *non-linear* functions, and in particular low-degree polynomial functions, of scalar random variables. This forms the focus of our work.

As a motivating example, consider the random matrix $\mathbf{M} \in \mathbb{R}^{[n]^2 \times [n]^2}$ obtained as

$$\mathbf{M} \;=\; \mathbf{A_1} \otimes \mathbf{A_1} + \cdots + \mathbf{A_m} \otimes \mathbf{A_m}\,,$$

where $\mathbf{A_1}, \ldots, \mathbf{A_m} \in \mathbb{R}^{[n] \times [n]}$ are independent random matrices, with i.i.d. entries uniformly distributed in $\{-1, 1\}$. It is easy to see that the entries of the matrix $\mathbf{M}$ are degree-2 polynomial functions of the independent random variables describing the entries of $\mathbf{A_1}, \ldots, \mathbf{A_m}$. The concentration of such a matrix was analyzed by Hopkins et al. [HSS15, Hop18], who use it to design spectral algorithms for a variant of the principal components analysis (PCA). This matrix is a special case of a more general setting that we study in this work.

*University of Chicago. Supported in part by NSF grants CCF-1816372 and CCF-2008920. goutham@uchicago.edu.
†Toyota Technological Institute at Chicago. Supported by NSF grant CCF-1816372. madhurt@ttic.edu

Matrix-valued polynomial functions. In the example above, the entries of the matrices are low-degree polynomials in independent (Rademacher) random variables. In this work, we consider a general setting where we take an n-tuple $Z = (Z_1, \ldots, Z_n)$ of independent and identically distributed random variables[1] distributed in Ω. We consider random matrices given by a matrix-valued function $\mathbf{F}(Z)$ taking values in $\mathbb{R}^{\mathcal{I}\times\mathcal{J}}$ for arbitrary index sets $\mathcal{I}, \mathcal{J}$, where each entry $\mathbf{F}[I, J](Z)$ is a polynomial in $Z_1, \ldots, Z_n$. We develop a general framework to analyze concentration of such matrices. Our matrix concentration results are simpler to state in the case when $Z_1, \ldots, Z_n$ are independent Rademacher variables uniformly distributed in $\{-1, 1\}$, but apply for the general case as well.

Special cases of such non-linear random matrices have been used in several applications in spectral algorithms and lower bounds. We now briefly discuss a few examples below. Note that while the previous methods used for these examples have been somewhat problem-specific, the goal of this work is to develop a general method. While our techniques *also* apply for these examples (providing a proof of concept), understanding these examples is not required to follow our results. A reader only interested in the techniques for obtaining concentration, may also choose to skip ahead to the next section directly.

1. **Tensor networks.** Random matrices such as the above were viewed as a special case of "flattened tensor networks" by Moitra and Wein [MW19], who also considered spectral algorithms obtained via somewhat larger tensor networks. A tensor network is a graph with nodes corresponding to tensors (see the figure below for an example). An edge between two nodes corresponds to shared indices for one of the dimensions and the degree of each node is equal to the order of the corresponding tensor (the number of dimensions). Such networks indicate how tensors of different orders can be multiplied to obtain larger ones. For example, the first network in the figure below illustrates the network corresponding to simple multiplication $\mathbf{A} \cdot \mathbf{B}$ of two matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times m}$, where the red and blue edges indicate the row and column indices respectively. Similarly, the second network in the figure below illustrates the network corresponding to the application by Hopkins et al. [HSSS16], where $\mathbf{T} \in \mathbb{R}^{n\times n\times m}$ is a random tensor with i.i.d. entries in $\{-1, 1\}$. While the latter network yields an order-4 tensor, they obtain a matrix in $\mathbb{R}^{n^2\times n^2}$ by "flattening" it, where the row is indicated by the indices in the red edges and the column is indicated by the indices in the blue edges. In the figure, we also indicate the index sets corresponding to each of the edges (though these are often supressed in the diagrams). Moitra and Wein [MW19] analyzed a larger tensor network, with a graph consisting of 10 nodes, in their algorithm for the continuous multi-reference alignment problem.

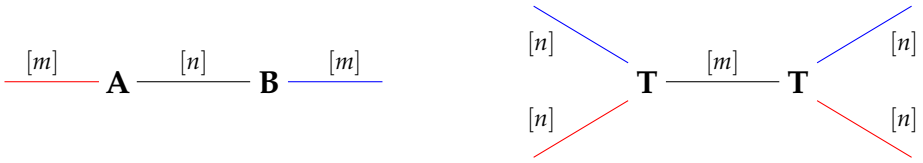

Figure 1: Tensor networks for matrix multiplication and the algorithm in [HSSS16]

2. **Graph matrices.** Another setting of nonlinear concentration arises from the analysis of the so-called "graph matrices" [MP16, AMP16]. Graph matrices play an important role in lower bounds for average-case problems, against algorithms based on the powerful Sum-of-Squares (SoS) SDP hierarchy running in polynomial time and even sub-exponential time [MPW15, DM15, HKP15, RS15, BHK+19, MRX20, GJJ+20, JPR+21, Raj22a, PR22, Jon22].

   Let $\mathbf{X}$ be the $\{\pm 1\}$-adjacency matrix of a random graph in $\mathcal{G}_{n,1/2}$ i.e., $\mathbf{X}[i, j]$ is uniform $\{-1, 1\}$ when $i \neq j$ and 0 when $i = j$. Graph matrices are random matrices corresponding to the occurences of a small graph pattern called a "shape". A shape τ is a small, fixed graph with two ordered subsets U_τ, V_τ of vertices. For simplicity, let τ be a shape of a fixed size, where the vertex set $V(\tau)$ is partitioned into two ordered sets $V(\tau) = U_\tau \sqcup V_\tau$. For such a shape τ, the corresponding *graph matrix* $\mathbf{M}_\tau$ has rows and columns indexed

[1] Our framework also applies when the variables are not necessarily identically distributed, as long as they are independent.

by $[n]^{|U_\tau|}$ and $[n]^{|V_\tau|}$ respectively, and we view the row and column indices I and J as defining a (unique in this case) map $\varphi : U_\tau \sqcup V_\tau \to [n]$. The corresponding entry is given by

$$\mathbf{M}_\tau[I, J] \;=\; \mathbf{M}_\tau[\varphi(U_\tau), \varphi(V_\tau)] \;=\; \begin{cases} \prod_{(u,v)\in E(\tau)} \mathbf{X}[\varphi(u), \varphi[v]] & \text{if } \varphi \text{ is injective} \\ 0 & \text{otherwise} \end{cases}$$

In the case of general graph matrices (defined formally in Section 4.2), U_τ, V_τ are arbitrary ordered subsets of the vertex set of τ, and we sum over all feasible injective maps φ. As an example, consider the case shown in Fig. 2, where τ is a triangle on three vertices $\{u_1, v_1, v_2\}$ with $U_\tau = (u_1)$ and $V_\tau = (v_1, v_2)$. Then, the corresponding matrix is given by

$$\mathbf{M}_\tau[i_1, (i_2, i_3)] \;=\; \mathbf{X}[i_1, i_2] \cdot \mathbf{X}[i_2, i_3] \cdot \mathbf{X}[i_3, i_1]\,,$$

where $\mathbf{X}$ automatically enforces injectivity.

Graph matrices are closely related to tensor networks (ignoring the injectivity constraint on φ). For instance, the above matrix can be viewed as the flattened tensor network below, where the tensor $\mathbf{I}$ denotes the "diagonal" tensor of order 3 with entries being 1 if all indices are equal and 0 otherwise.

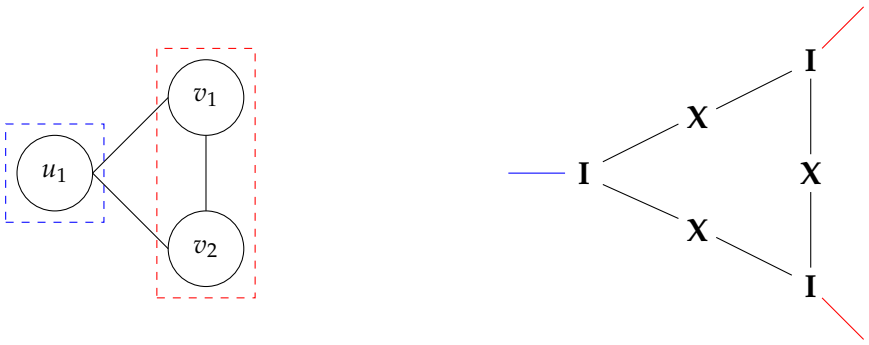

Figure 2: The graph τ and corresponding flattened tensor network

Analyzing concentration Recall that our objective is to analyze the concentration of polynomial random matrices. To motivate our approach, consider first the problem of obtaining concentration bounds on a *scalar* polynomial $f(Z)$ with mean zero. To obtain such bounds, because of Markov's inequality, it suffices to compute moment estimates

$$\mathbb{P}\left[|f(Z)| \geq \lambda\right] \;=\; \mathbb{P}\left[(f(Z))^{2t} \geq \lambda^{2t}\right] \;\leq\; \lambda^{-2t} \cdot \mathbb{E}\left[(f(Z))^{2t}\right]$$

While in some cases $\mathbb{E}[(f(Z))^{2t}]$ can be computed by direct expansion, it often involves an intricate analysis of the structure of terms with degrees growing with t, and therefore indirect methods may be more convenient. One such method is based on hypercontractive inequalities. In particular for Rademacher variables, the hypercontractive inequality [O'D08] gives that for a polynomial f of degree d_p, we have

$$\mathbb{E}\left[(f(Z))^{2t}\right] \;\leq\; (2t-1)^{d_p \cdot t} \cdot \left(\mathbb{E}\left[(f(Z))^{2}\right]\right)^{t}.$$

Thus, for (scalar) polynomial functions, the hypercontractive inequality gives moment estimates using $(f(Z))^2$, which is convenient because $(f(Z))^2$ is a polynomial of *fixed* degree and therefore is much easier to understand. In fact, it can often be conveniently analyzed using the Fourier coefficients of f.

The matrix analog of the above argument involves the Schatten-$2t$ norm $\|.\|_{2t}$, which is defined for a matrix $\mathbf{M}$ with non-zero singular values $\sigma_1, \ldots, \sigma_r$ as $\|\mathbf{M}\|_{2t}^{2t} \;:=\; \sum_{j\in[r]} \sigma_j^{2t}$. For a function $\mathbf{F}$ with $\mathbb{E}[\mathbf{F}(Z)] = 0$, we have the following bound using Schatten norms.

$$\mathbb{P}\left[\sigma_1(\mathbf{F}) \geq \lambda\right] \;\leq\; \lambda^{-2t} \cdot \mathbb{E}\,\|\mathbf{F}\|_{2t}^{2t} \;=\; \lambda^{-2t} \cdot \mathbb{E}\,\mathrm{tr}\left[(\mathbf{F}(Z)\mathbf{F}(Z)^\intercal)^t\right]$$

Known norm bounds for tensor networks [MW19] (which involves Gaussian variables) and graph matrices [AMP16, JPR+21], start with the above inequality, and rely on direct expansion of the trace. They analyze terms in the expansion as being formed by $2t$ copies of the network/shape, with possibly overlapping vertex sets. To analyze such graphs, they both rely on intricate combinatorics, as well as arguments relying crucially on the problem structure.

In terms of general techniques, while hypercontractive inequalities are also known for matrix-valued functions of Rademacher variables [BARDW08], their form involves Schatten-p norms for $p \in [1,2]$ and (to the best of our knowledge) are not known to imply matrix concentration. To get around this, we consider another indirect method based on Efron-Stein inequalities. In the scalar case, Efron-Stein inequalities give us a slight weakening of the above scalar bound. Interestingly, it turns out that this can indeed be generalized to the matrix case.

Efron-Stein inequalities. Efron-Stein inequalities bound the global variance of a function of independent random variables, in terms of local variance estimates obtained by changing one variable at a time. For $i \in [n]$ and tuple $Z = (Z_1, \ldots, Z_n)$, let $Z^{(i)}$ denote the tuple $(Z_1, \ldots, Z_{i-1}, \widetilde{Z}_i, Z_{i+1}, \ldots, Z_n)$, where $\widetilde{Z}_i$ is an independent copy of Z_i. For a scalar function $f(Z)$, the Efron-Stein inequality states that

$$\mathrm{Var}\left[f(Z)\right] \;=\; \mathbb{E}\left[(f(Z) - \mathbb{E}f)^2\right] \;\leq\; \frac{1}{2} \cdot \sum_{i \in [n]} \mathbb{E}\left[\left(f(Z) - f\left(Z^{(i)}\right)\right)^2\right] \;=\; \mathbb{E}\left[V(Z)\right],$$

where $V(Z) \;:=\; \sum_{i \in [n]} \mathbb{E}\left[\left(f(Z) - f\left(Z^{(i)}\right)\right)^2 | Z\right]$. For Rademacher variables, $\mathbb{E}[V(Z)]$ is equal to the total influence from boolean Fourier analysis and indeed, the above inequality can also be observed via Fourier analysis. In fact, when f is a polynomial of degree d_p, the two sides are within a factor d_p.

A moment version of the Efron-Stein inequality was developed by Boucheron et al. [BBLM05], who obtain bounds in terms of $V(Z)$ (in fact, in terms of more refined quantities $V_+(Z)$ and $V_-(Z)$) which serves as a proxy for the variance. Their results imply that for a function f,

$$\mathbb{E}\left[(f(Z) - \mathbb{E}f)^{2t}\right] \;\leq\; (C_0 \cdot t)^t \cdot \mathbb{E}\left[(V(Z))^t\right].$$

A beautiful matrix generalization of the above inequality (Theorem 1.1 below) was obtained by Paulin, Mackey and Tropp [PMT16], via the method of exchangeable pairs (see also [HT21a] for a different proof). Their inequality is stated for Hermitian matrix valued functions $\mathbf{H}$. But we can also use it for non-Hermitian functions $\mathbf{F}$, where we simply apply it to the Hermitian dilation $\mathbf{H} = \begin{bmatrix} 0 & \mathbf{F} \\ \mathbf{F}^\intercal & 0 \end{bmatrix}$ instead.

THEOREM 1.1. ([PMT16]) *Let $\mathbf{H}(Z)$ be a Hermitian matrix valued function of independent random variables $Z = (Z_1, \ldots, Z_n)$ with $\mathbb{E}\,\|\mathbf{H}\| < \infty$. Then, for each natural number $t \geq 1$,*

$$\mathbb{E}\,\mathrm{tr}\left[(\mathbf{H} - \mathbb{E}\mathbf{H})^{2t}\right] \;\leq\; (4t-2)^t \cdot \mathbb{E}\,\mathrm{tr}\left[\mathbf{V}^t\right],$$

where $\mathbf{V}(Z)$ is the variance proxy defined as

$$\mathbf{V}(Z) := \frac{1}{2} \cdot \sum_{i=1}^{n} \mathbb{E}\left[\left(\mathbf{H}(Z) - \mathbf{H}\left(Z^{(i)}\right)\right)^2 \mid Z\right].$$

A simple bound for Rademacher variables. The form of the variance proxy suggests a recursive approach for polynomial functions (say of degree d_p) of Rademacher variables. Consider the scalar case again, where we assume without loss of generality that f is multi-linear. In particular, consider the Efron-Stein inequality by

Boucheron et al. [BBLM05], where the variance proxy can be written as

$$V(Z) = \frac{1}{2}\cdot\sum_{i\in[n]}\mathbb{E}\left[\left(f(Z)-f\left(Z^{(i)}\right)\right)^2 \mid Z\right] = \frac{1}{2}\cdot\sum_{i\in[n]}\mathbb{E}\left[(Z_i-\widetilde{Z}_i)^2\cdot\left(\frac{\partial f(Z)}{\partial Z_i}\right)^2 \mid Z\right]$$
$$= \sum_{i\in[n]}\left(\frac{\partial f(Z)}{\partial Z_i}\right)^2 = \|\mathbf{f}_1(Z)\|_2^2,$$

where $\mathbf{f}_1(Z)$ is a vector-valued function given by $\mathbf{f}_1[i](Z) = \frac{\partial f(Z)}{\partial Z_i}$. Thus, to estimate $\mathbb{E}\,(f(Z))^{2t}$, we just need to estimate $\mathbb{E}\,\|\mathbf{f}_1(Z)\|_2^{2t}$, where $\mathbf{f}_1(Z)$ is now a vector valued function. The key observation is that $\mathbf{f}_1(Z)$ has entries of degree at most $d_p - 1$. This suggests that we can apply this inequality recursively until we end up with constant polynomials, which we fully understand. We can do a similar computation for matrix-valued functions $\mathbf{F}(Z)$ usingTheorem 1.1. This yields two matrices $\mathbf{F}_{0,1}$ and $\mathbf{F}_{1,0}$ of partial derivatives, where an extra index i is added either to the row or column indices. Iterating this yields the following result, which we state in terms of the partial derivative operators $\nabla_\alpha(f) = \left(\prod_{i:\alpha_i=1}\frac{\partial}{\partial Z_i}\right)(f)$ for $\alpha\in\{0,1\}^n$ (extended entry-wise to matrices).

THEOREM 1.2. (RADEMACHER RECURSION) *Let $\mathbf{F}:\{-1,1\}^n\to\mathbb{R}^{\mathcal{I}\times\mathcal{J}}$ be a matrix valued polynomial function of degree at most d_p. Then, for each natural number $t\geq 1$,*

$$\mathbb{E}\,\|\mathbf{F}-\mathbb{E}\mathbf{F}\|_{2t}^{2t} \leq \sum_{1\leq a+b\leq d_p}(16td_p)^{(a+b)\cdot t}\cdot\left\|\mathbb{E}\mathbf{F}_{a,b}\right\|_{2t}^{2t},$$

where $\mathbf{F}_{a,b}$ is a matrix of partial derivatives indexed by the sets $\mathcal{I}\times\binom{[n]}{a}$ and $\mathcal{J}\times\binom{[n]}{b}$ with

$$\mathbf{F}_{a,b}[(\cdot,\alpha),(\cdot,\beta)] = \begin{cases}\nabla_{\alpha+\beta}(\mathbf{F}) & \text{if } \alpha\cdot\beta = 0\\ 0 & \text{otherwise}\end{cases}$$

where $\alpha,\beta\in\{0,1\}^n$ are indicator vectors of sets in $\binom{[n]}{a}$ and $\binom{[n]}{b}$ respectively.

REMARK 1.1. *While we state our results in terms of polynomial moment bounds, it is also possible to obtain exponential tails using these results. This can be done either using an appropriate (known) variant of Theorem 1.1, or by using a sufficiently large value of t. These results can also recover (known) matrix Chernoff or Bernstein inequalities when the function $\mathbf{F}$ is linear, but of course the much more interesting case is when $\mathbf{F}$ is a polynomial function.*

We cover some applications of the above theorem in Section 4. Similar to the hypercontractive bound for the scalar case, the bound above is in terms of a small number ($O(d_p^2)$) of matrices that arise from polynomials of fixed degree (not growing with t), but importantly, they are *deterministic* matrices. Because they are deterministic, analyzing them is considerably easier. Note that bounds depending on norms of a fixed number of deterministic matrices, arise even in the study of concentration for *scalar* polynomial functions [AW15], and thus it is not surprising that they are needed to control the much more challenging case of matrix-valued functions.

When we apply this theorem to the case $\mathbf{F} = \mathbf{M}_\tau$, the graph matrix of a shape τ, we obtain bounds in terms of combinatorial objects known as "vertex separators" of the shape τ. This recovers the bounds by Ahn et al. [AMP16] and perhaps surprisingly (to the authors), this gives an alternative and direct derivation of these combinatorial structures such as vertex separators, compared to the ingenious observations made in Ahn et al. [AMP16]. The important takeaway is that these norm bounds can be recovered by our more general technique rather than relying on problem-specific methods.

Extending the framework to general product distributions. A key contribution of our work is to show how the above framework can be extended to arbitrary product distributions (with bounded moments). A motivating example of this is norm bounds for the so-called "sparse graph matrices". In sparse graph matrices, the variables Z_i can be thought of as (normalized) edges of a $\mathcal{G}_{n',p}$ graph, that is, $Z_i = -\sqrt{\frac{1-p}{p}}$ with probability

p and $Z_i = \sqrt{\frac{p}{1-p}}$ with probability $1-p$. These variables are standard in p-biased Fourier analysis [O'D14] and are chosen to satisfy $\mathbb{E}\, Z_i = 0$ and $\mathbb{E}\, Z_i^2 = 1$. Sparse graph matrices naturally arise when analyzing average case problems on $\mathcal{G}_{n,p}$ graphs for $p = o(1)$, as opposed to $\mathcal{G}_{n,1/2}$ graphs.

Until recently, little was known about norm bounds for sparse graph matrices. The difficulty stems partly from the fact that when $p = o(1)$, it is important that sparse graph matrix norm bounds have the right dependence on p and not just on n. Such norm bounds were obtained recently by Jones et al. [JPR+21], via the trace power method which involved a delicate combinatorial counting argument. On the other hand, we obtain similar norm bounds using our framework but in a more mechanical fashion. We can also readily apply our framework in the even more general case of sub-Gaussian random variables and our bounds will depend on the sub-Gaussian norm of the distributions.

To extend our framework to general product distributions, we could take inspiration from the Rademacher case and could attempt to simply recursively apply the Efron-Stein inequality. Unfortunately, this idea will fail. The issue can be observed by again considering the scalar case. Assume that $Z_1, \ldots, Z_n$ are i.i.d. with $\mathbb{E}\, Z_i = 0$ and $\mathbb{E}\, Z_i^2 = 1$ for all $i \in [n]$. Also assume for simplicity that $f(Z)$ is a multi-linear polynomial of degree d_p. Analyzing the variance proxy as before, we get

$$V(Z) \;=\; \frac{1}{2} \cdot \sum_{i \in [n]} \mathbb{E}\left[(Z_i - \widetilde{Z}_i)^2 \cdot \left(\frac{\partial f(Z)}{\partial Z_i} \right)^2 \mid Z \right] \;=\; \frac{1}{2} \sum_{i \in [n]} \mathbb{E}\left[(Z_i - \widetilde{Z}_i)^2 | Z \right] \cdot \left(\frac{\partial f(Z)}{\partial Z_i} \right)^2 .$$

In the Rademacher case, we had $\mathbb{E}[(Z_i - \widetilde{Z}_i)^2 | Z] = 2$. This left us with the polynomials corresponding to partial derivatives but which importantly had a strictly lower degree. However, for a general product distribution, we instead have $\mathbb{E}[(Z_i - \widetilde{Z}_i)^2 | Z] = 1 + Z_i^2$. This gives back a term $\left(Z_i \cdot \frac{\partial f}{\partial Z_i} \right)^2$ where the polynomial inside the square could have degree possibly still equal to d_p. This means that in the next step of the recursion, we may again have to consider a derivative with respect to Z_i and may again end up with the same polynomial f. Therefore, the recursion is stalled! A similar issue occurs for matrices, which is elaborated in Section 5. To get around this, we generalize the work of [PMT16].

Generalizing [PMT16] via explicit inner kernels. To resolve the above issue, we modify the proof of [PMT16] and our proof techniques may be of independent interest.

We first recall how the matrix Efron-Stein inequality,Theorem 1.1, was proved in [PMT16]. Their basic strategy is to utilize the theory of *exchangeable pairs* [Ste72, Ste86, Cha05, Cha06], in particular *kernel Stein pairs*. A kernel Stein pair is an exchangeable pair of random matrices that has a "kernel", a bivariate function that "reproduces" the matrices in the pair. More concretely, consider an exchangeable pair of random variables (Z, Z') (which means (Z', Z) has the same distribution). For this exchangeable pair, a bivariate matrix-valued function $\mathbf{K}(z, z')$ is said to be a kernel for a matrix-valued function $\mathbf{F}$ if it satisfies

- Anti-symmetry: $\mathbf{K}(z', z) = -\mathbf{K}(z, z')$ for all inputs (z, z').
- Reproducing property: $\mathbb{E}[\mathbf{K}(Z, Z') \mid Z] = \mathbf{F}(Z)$.

If such a kernel $\mathbf{K}$ exists, then the pair of random variables $(\mathbf{F}(Z), \mathbf{F}(Z'))$ is said to be a kernel Stein pair.

Building on ideas from [Ste86, Cha05], Paulin, Mackey and Tropp [PMT16] first show the existence of a kernel, by exhibiting it as a limit of coupled Markov Chains. By studying the evolution of this kernel coupling, they prove analytic properties of the kernel. Then, using this kernel, they employ the powerful method of exchangeable pairs to evaluate moments of the random matrix, which in turn will imply concentration.

For a Hermitian random matrix $\mathbf{X}$, they introduce two matrices - the *conditional variance* $\mathbf{V_X}$ which measures the squared fluctuations of $\mathbf{X}$ when resampling a coordinate of Z; and the *kernel conditional variance* $\mathbf{V^K}$ which measures the squared fluctation of the kernel when resampling a coordinate of Z. With these matrices in hand, they bound the Schatten $2t$-norm of $\mathbf{X}$ by the Schatten t-norm of $s\mathbf{V_X} + s^{-1}\mathbf{V^K}$ for any parameter $s > 0$. Finally, they choose s appropriately to make these two quantities approximately equal, in which case it simplifies to the variance proxy $\mathbf{V}$, provingTheorem 1.1.

In our setting, no such choice of s is feasible because for any choice of s, either the conditional variance term $s\mathbf{V}_{\mathbf{X}}$ will dominate $\mathbf{X}^2$ or the kernel conditional variance term $s^{-1}\mathbf{V}^{\mathbf{K}}$ will dominate $\mathbf{X}^2$. This will make the main inequalityTheorem 1.1 trivial.

To get around this, we will exploit the structure of the matrix we have, i.e. $\mathbf{F} = \mathbf{DGD}$ where $\mathbf{D}$ is a diagonal matrix that encodes all variables that have already been differentiated on and $\mathbf{G}$ is a polynomial matrix of the remaining variables. Since $\mathbf{D}$ is a simple diagonal matrix with low degrees, most of the deviations exhibited by $\mathbf{F}$ are in fact likely to be exhibited by $\mathbf{G}$. To capture this intuition, we consider a kernel for only the inner matrix $\mathbf{G}$ instead of $\mathbf{F}$ as a whole. We call this an *inner kernel*.

This helps us avoid the root cause of the issue, i.e. differentiating on variables we have already encountered (which correspond to entries in $\mathbf{D}$). Therefore, the recursion will not stall!

However, in general, this is not realizable since $\mathbf{D}$ and the kernel of $\mathbf{G}$ can interact in unexpected ways. To study this interaction, we construct explicit polynomial kernels (Theorem 7.1) (compared to [PMT16] who show the existence of the kernel but for all functions).

We study how this explicit inner kernel interacts with $\mathbf{D}$ (see Lemma 7.3) and use it to obtain a generalization of the inequalities by [PMT16] (generalized because setting $\mathbf{D} = \mathbf{I}$ will give back their result) stated in Lemma 7.5.

A subtle issue is that the conditional variance of $\mathbf{X}$ may still have additional deviations due to the diagonal matrices $\mathbf{D}$ (which still involve random variables). We control the additional deviations using Jensen's operator trace inequality (for non-commuting averages) [HP03] (stated in Lemma 2.1). Putting these ideas together lets us obtain a version of the Efron-Stein inequality where the variance proxy only corresponds to the conditional variance of the inner kernel. In the setting of polynomial functions, this inequality generalizes the work of [PMT16].

With the modified Efron-Stein inequality from above, we cannot guarantee that the matrices $\mathbf{F}$ at intermediate steps are of lower degree, but on the other hand, the degree of the inner matrix $\mathbf{G}$ reduces at each step. Therefore, we can recursively apply this inequality to obtain our final bounds. The final bounds are then stated in terms of norm bounds for the simplified matrices of the form $\mathbf{DGD}$ where $\mathbf{G}$ are deterministic matrices and $\mathbf{D}$ are diagonal matrices which are still functions of Z. While random, these matrices can be easily analyzed via simple scalar concentration tools.

The main theorem is stated in Section 6, in particularTheorem 6.1, with the proof following in Section 7. While our proof builds on the work by [PMT16], the argument here is self-contained.

Applications. Our framework is suitable for many nonlinear concentration results obtained in the literature [BBH+12, GM15, HSS15, MP16, AMP16, HSSS16, SS17, Hop18, HSS19, MW19, KP20, PR20, JPR+21, BHKX22, Raj22b, Jon22]. We show a few of these applications in Section 4 and Section 8. We expect similar future applications to benefit from our framework because the task is mechanically reduced to analyzing considerably simpler matrices.

In Section 4.2, we derive norm bounds on dense graph matrices. In earlier works, dense graph matrices have been used extensively in analysis of semidefinite programming hierarchies, especially the Sum-of-Squares (SoS) hierarchy [MPW15, DM15, HKP15, RS15, BHK+19, MRX20, GJJ+20, PR22, Raj22a]. For more applications and a detailed treatment of graph matrices, see [AMP16, Jon22].

In Section 8, we derive norm bounds for sparse graph matrices. Sparse graph matrices have been relatively less understood until recently, when [JPR+21] obtained norm bounds for such matrices via the trace power method. They use these bounds to prove SoS lower bounds for the maximum independent set problem on sparse graphs.

Other related work Nonlinear concentration for the case of scalar-valued functions has been the subject of an extensive body of work. In addition to the results of Schudy and Sviridenko [SS11] which we use, strong concentration results have also been obtained (for example) in the results of Latała [Lat06], Adamczak and Wolff [AW15], and Bobkov, Götze, and Sambale [BGS19]. In addition to the above results, hypercontractive inequalities can also be used to obtain concentration inequalities for low-degree (scalar) polynomial functions [O'D08] with possible sub-optimal exponents.

For the case of matrix-valued functions, while we rely here on the work of Paulin, Mackey, and Tropp [PMT16] for product distributions, later works have also extended these results to distributions satisfying weaker assumptions. In particular, Aoun, Banna, and Youssef [ABY20] obtained matrix concentration for distributions satisfying matrix Poincaré inequalities, building on earlier work of Cheng, Hsieh, and Tomamichel [CHT17, CH19]. It was later proved by Garg, Kathuria, and Srivastava [GKS21] that the matrix Poincaré inequalities are implied by scalar Poincaré inequalities. Independently, matrix concentration based on scalar Poincaré inequalities was also proved by Huang and Tropp [HT21a]. Another work of Huang and Tropp [HT21b] also establishes matrix concentration inequalities via semigroup methods. While some hypercontractive inequalities are also known for matrix-valued functions [BARDW08, AD21], to the best of our knowledge, they do not imply concentration bounds for matrices with low-degree polynomial entries.

Potential extensions In this work, we assumed that the input forms a product distribution. In other words, the variables $Z_1, \dots, Z_n$ are independent. A natural extension is the case when they are not independent. This has important applications for many problems such as when the input is a uniform d-regular graph, or when the input is sampled from a distribution with a global constraint, etc. In such cases, the input variables are not independent but it may be possible to use similar ideas to analyze concentration.

More concretely, to study concentration in the non-independent setting, one can use the recent work of Huang and Tropp [HT21a] on matrix concentration from Poincaré inequalities, together with our framework. For this, we just need to exhibit a Markov process that converges to our desired distribution.

Organization of the paper and bibliographic note. We start with preliminaries in Section 2. In Section 3, we state and prove the Rademacher recursion. We illustrate some applications of this framework in Section 4. In Section 5, we explain why similar ideas may not be enough in the general case. We then propose our general framework in Section 6 and prove it in Section 7. We end with an application of the general framework to sparse graph matrices in Section 8. An earlier version of this paper also appeared in the dissertation of the first author [Raj22b].

2 Preliminaries

Notation We use boldface letters such as $\mathbf{I}, \mathbf{M}, \mathbf{X} \dots$, to denote matrices. Entries of a matrix $\mathbf{X} \in \mathbb{R}^{\mathcal{I} \times \mathcal{J}}$ will be denoted by $\mathbf{X}[I, J]$ for $I \in \mathcal{I}, J \in \mathcal{J}$. Let $\mathbb{H}^n$ denote the set of $n \times n$ real symmetric matrices. The trace of a matrix $\mathbf{X} \in \mathbb{H}^n$ equals $\sum_{i \in [n]} \mathbf{X}[i, i]$ and is denoted by $\operatorname{tr} \mathbf{X}$.

Multi-index notation For any pair of vectors $\alpha, \beta \in \mathbb{N}^n$ and scalar $c \in \mathbb{N}$, we define $\alpha + \beta, \alpha \cdot \beta, c\alpha$ entrywise. We also define the orderings $\alpha \leq \beta$ and $\alpha \trianglelefteq \beta$ where we say $\alpha \leq \beta$ if for each i, $\alpha_i \leq \beta_i$, and $\alpha \trianglelefteq \beta$ if for each i, α_i is either 0 or β_i. We denote by $|\alpha|_0$ the number of nonzero entries of α and by $|\alpha|_1$, the sum of entries of α. For a boolean vector $\gamma \in \{0,1\}^n$, we define $1 - \gamma$ the vector with all its bits flipped.

Derivatives For variables $Z_1, \dots, Z_n$ and $\alpha \in \mathbb{N}^n$, define the monomial $Z^\alpha := \prod_{i=1}^n Z_i^{\alpha_i}$. This forms a standard basis for polynomials.

For $\alpha \in \mathbb{N}^n$, we define the linear operator ∇_α that acts on polynomials by defining its action on the elements Z^β as follows and then extend linearly to all polynomials.

$$\nabla_\alpha(Z^\beta) = \begin{cases} Z^{\beta - \alpha} & \text{if } \alpha \trianglelefteq \beta \\ 0 & \text{o.w.} \end{cases}$$

Informally, for a polynomial f written as a linear combination of the standard basis polynomials Z^β, $\nabla_\alpha(f)$ isolates the terms that precisely contain the powers $Z_i^{\alpha_i}$ for all i such that $\alpha_i \neq 0$ and then truncates these powers. In other words, it's the coefficient of Z^α in f. In particular, observe that $\nabla_\alpha(f)$ does not depend on Z_i for any i such that $\alpha_i \neq 0$.

Supose f is multilinear, as we can assume in the Rademacher case when we are working with $Z_i \in \{-1, 1\}$. For $\alpha \in \{0,1\}^n$ with nonzero indices $i_1, \dots, i_k \in [n]$, we have $\nabla_\alpha(f) = \frac{\partial}{\partial Z_{i_1}} \dots \frac{\partial}{\partial Z_{i_k}} f$. So this linear operator generalizes the partial derivative operator. But note that in general, ∇ is not simply the standard partial

derivative operator.

Matrix Analysis Linear operators that act on polynomials can also be naturally defined to act on matrices by acting on each entry.

We define $\mathbf{I}_m$ to be the $m \times m$ identity matrix. We drop the subscript when it's clear. For matrices $\mathbf{F}, \mathbf{G}$, define $\mathbf{F} \oplus \mathbf{G}$ to be the matrix $\begin{bmatrix} 0 & \mathbf{F} \\ \mathbf{G} & 0 \end{bmatrix}$. For a matrix $\mathbf{F}$, define its Hermitian dilation $\overline{\mathbf{F}}$ as $\mathbf{F} \oplus \mathbf{F}^T$. Denote by $\preceq$ the Loewner order, that is, $\mathbf{A} \preceq \mathbf{B}$ for $\mathbf{A}, \mathbf{B} \in \mathbb{H}^n$ if and only if $\mathbf{B} - \mathbf{A}$ is positive semi-definite.

DEFINITION 2.1. *For a matrix* $\mathbf{F}$ *and an integer* $t \geq 0$*, define the Schatten* $2t$*-norm as*

$$\|\mathbf{F}\|_{2t}^{2t} = \mathrm{tr}[(\mathbf{F}\mathbf{F}^T)^t]$$

FACT 2.1. *For real symmetric matrices* $\mathbf{X}_1, \ldots, \mathbf{X}_n$*, we have*

$$(\mathbf{X}_1 + \ldots + \mathbf{X}_n)^2 \preceq n(\mathbf{X}_1^2 + \ldots + \mathbf{X}_n^2)$$

FACT 2.2. *For positive semidefinite matrices* $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$ *such that* $\mathbf{X} \preceq \mathbf{X}_1 + \ldots + \mathbf{X}_n$ *and for any integer* $t \geq 1$*,*

$$\mathrm{tr}[\mathbf{X}^t] \leq n^{t-1}(\mathrm{tr}[\mathbf{X}_1^t] + \ldots + \mathrm{tr}[\mathbf{X}_n^t])$$

Proof. By Hölder's inequality, $n^{t-1}(\mathrm{tr}[\mathbf{X}_1^t] + \ldots + \mathrm{tr}[\mathbf{X}_n^t]) \geq (\|\mathbf{X}_1\|_t + \ldots + \|\mathbf{X}_n\|_t)^t$. By triangle inequality of Schatten norms, this is at least $\|\mathbf{X}_1 + \ldots + \mathbf{X}_n\|_t^t$. Finally, because $\mathbf{X}_1 + \ldots + \mathbf{X}_n \succeq \mathbf{X} \succeq 0$, we can use the monotonicity of trace functions (see [Pet94, Proposition 1]) where we use the increasing function $f(x) = x^t$ on $x \in [0, \infty)$. This proves the result. ∎

LEMMA 2.1. (JENSEN'S OPERATOR TRACE INEQUALITY) *[HP03, Corollary 2.5] Let* f *be a convex, continuous function defined on an interval* I *and suppose that* $0 \in I$ *and* $f(0) \leq 0$*. Then, for all integers* $m, n \geq 1$*, for every tuple* $\mathbf{B}_1, \ldots, \mathbf{B}_n$ *of real symmetric* $m \times m$ *matrices with spectra contained in* I *and every tuple* $\mathbf{A}_1, \ldots, \mathbf{A}_n$ *of* $m \times m$ *matrices with* $\sum_{i=1}^n \mathbf{A}_i^T \mathbf{A}_i \preceq \mathbf{I}$*, we have*

$$\mathrm{tr}[f(\sum_{i=1}^n \mathbf{A}_i^T \mathbf{B}_i \mathbf{A}_i)] \leq \mathrm{tr}[\sum_{i=1}^n \mathbf{A}_i^T f(\mathbf{B}_i) \mathbf{A}_i]$$

3 The basic framework for Rademacher random variables

Let $Z = (Z_1, \ldots, Z_n)$ be sampled uniformly from $\{-1, 1\}^n$. We will consider matrix-valued functions $\mathbf{F} : \{-1, 1\}^n \to \mathbb{R}^{\mathcal{I} \times \mathcal{J}}$, with rows and columns indexed by arbitrary sets $\mathcal{I}, \mathcal{J}$ respectively such that for all $I \in \mathcal{I}, J \in \mathcal{J}$,

$$\mathbf{F}[I, J] = f_{I,J}(Z)$$

where $f_{I,J}$ are polynomials of $Z_1, \ldots, Z_n$. Since $Z_i \in \{-1, 1\}$, we can assume without loss of generality that $f_{I,J}$ are multilinear. Let d_p be the maximum degree of any $f_{I,J}$ in $\mathbf{F}$. In this section, we will give a general framework using which we can obtain bounds on $\mathbb{E}\,\|\mathbf{F} - \mathbb{E}\mathbf{F}\|_{2t}^{2t}$ for any integer $t \geq 1$. We restate the theorem for convenience.

THEOREM 3.1. (RADEMACHER RECURSION) *Let* $\mathbf{F} : \{-1, 1\}^n \to \mathbb{R}^{\mathcal{I} \times \mathcal{J}}$ *be a matrix valued polynomial function of degree at most* d_p*. Then, for each natural number* $t \geq 1$*,*

$$\mathbb{E}\,\|\mathbf{F} - \mathbb{E}\mathbf{F}\|_{2t}^{2t} \leq \sum_{1 \leq a+b \leq d_p} (16td_p)^{(a+b)\cdot t} \cdot \left\|\mathbb{E}\mathbf{F}_{a,b}\right\|_{2t}^{2t},$$

where $\mathbf{F}_{a,b}$ *is a matrix of partial derivatives indexed by the sets* $\mathcal{I} \times \binom{[n]}{a}$ *and* $\mathcal{J} \times \binom{[n]}{b}$ *with*

$$\mathbf{F}_{a,b}[(\cdot, \alpha), (\cdot, \beta)] = \begin{cases} \nabla_{\alpha+\beta}(\mathbf{F}) & \text{if } \alpha \cdot \beta = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha, \beta \in \{0,1\}^n$ are indicator vectors of sets in $\binom{[n]}{a}$ and $\binom{[n]}{b}$ respectively.

REMARK 3.1. *To obtain high probability norm bounds from moment estimates, we can set $t = \mathrm{polylog}(n)$ and invoke Markov's inequality. Since we do not attempt to optimize the dependence on the logarithmic factors, we do not attempt to optimize the exponent of t in the main theorem.*

To prove this, we will prove Lemma 3.1 and then recursively apply it. For each $i \le n$, define the random vector

$$Z^{(i)} := (Z_1, \dots, Z_{i-1}, \widetilde{Z}_i, Z_{i+1}, \dots, Z_n)$$

where $\widetilde{Z}_i$ is an independent copy of Z_i, that is, is independently resampled from $\{-1, 1\}$.

Let $\mathbf{X} := \mathbf{F} - \mathbb{E}\mathbf{F}$. When the input is Z, we denote the matrices as $\mathbf{F}, \mathbf{X}$, etc and when the input is $Z^{(i)}$, denote the corresponding matrices as $\mathbf{F}^{(i)}, \mathbf{X}^{(i)}$, etc. That is, for $I \in \mathcal{I}, J \in \mathcal{J}$, we have $\mathbf{F}^{(i)}[I, J] = f_{I,J}(Z^{(i)})$. Define $\mathbf{X}_{a,b} = \mathbf{F}_{a,b} - \mathbb{E}\mathbf{F}_{a,b}$.

LEMMA 3.1. *For integers $a, b \ge 0$, we have*

$$\mathbb{E}\left\|\mathbf{X}_{a,b}\right\|_{2t}^{2t} \le (16td_p)^t(\mathbb{E}\left\|\mathbf{X}_{a,b+1}\right\|_{2t}^{2t} + \mathbb{E}\left\|\mathbf{X}_{a+1,b}\right\|_{2t}^{2t} + \left\|\mathbb{E}\mathbf{F}_{a,b+1}\right\|_{2t}^{2t} + \left\|\mathbb{E}\mathbf{F}_{a+1,b}\right\|_{2t}^{2t})$$

Using this lemma, we can complete the proof of the main theorem.

Proof. [Proof of Theorem 1.2] Observing that $\mathbf{X}$ is a principal submatrix of $\mathbf{X}_{0,0}$ with all other entries being 0, we can apply Lemma 3.1 repeatedly until $\mathbf{X}_{a,b} = 0$, which will be the case if $a + b > d_p$. ∎

In the rest of this section, we will prove Lemma 3.1. We start with a basic fact. Let $\mathbf{e}_i \in \{0,1\}^n$ be the vector with a unique nonzero entry $(\mathbf{e}_i)_i = 1$.

PROPOSITION 3.1. *For a multilinear polynomial $f(Z) = f(Z_1, \dots, Z_n)$, we have*

$$f(Z) - f(Z^{(i)}) = (Z_i - \widetilde{Z}_i) \cdot \nabla_{\mathbf{e}_i} f(Z)$$

Proof. [Proof of Lemma 3.1] Consider the Hermitian dilation $\overline{\mathbf{F}}_{a,b} = \mathbf{F}_{a,b} \oplus \mathbf{F}_{a,b}^T$. Define $\overline{\mathbf{X}}_{a,b} = \overline{\mathbf{F}}_{a,b} - \mathbb{E}\overline{\mathbf{F}}_{a,b} = \mathbf{X}_{a,b} \oplus \mathbf{X}_{a,b}^T$. ByTheorem 1.1 applied to $\overline{\mathbf{X}}_{a,b}$, $\mathbb{E}\,\mathrm{tr}\left[\overline{\mathbf{X}}_{a,b}^{2t}\right] \le (2(2t-1))^t \mathbb{E}\,\mathrm{tr}\left[\mathbf{V}_{a,b}^t\right]$ where $\mathbf{V}_{a,b}$ is the variance proxy $\mathbf{V}_{a,b} = \frac{1}{2}\sum_{i=1}^n \mathbb{E}[(\overline{\mathbf{X}}_{a,b} - \overline{\mathbf{X}}_{a,b}^{(i)})^2|Z]$. By a simple computation, $\mathbb{E}\,\mathrm{tr}\left[\overline{\mathbf{X}}_{a,b}^{2t}\right] = \mathbb{E}\,\mathrm{tr}\left[(\mathbf{X}_{a,b}\mathbf{X}_{a,b}^\intercal)^t\right] + \mathbb{E}\,\mathrm{tr}\left[(\mathbf{X}_{a,b}^\intercal\mathbf{X}_{a,b})^t\right] = 2\mathbb{E}\left\|\mathbf{X}_{a,b}\right\|_{2t}^{2t}$, therefore

$$\begin{aligned}\mathbf{V}_{a,b} &= \frac{1}{2}\sum_{i=1}^n \mathbb{E}\left[\begin{bmatrix}(\mathbf{X}_{a,b} - \mathbf{X}_{a,b}^{(i)})(\mathbf{X}_{a,b} - \mathbf{X}_{a,b}^{(i)})^\intercal & 0 \\ 0 & (\mathbf{X}_{a,b} - \mathbf{X}_{a,b}^{(i)})^\intercal(\mathbf{X}_{a,b} - \mathbf{X}_{a,b}^{(i)})\end{bmatrix}|Z\right] \\ &= \frac{1}{2}\begin{bmatrix}\sum_{i=1}^n \mathbb{E}[(\mathbf{F}_{a,b} - \mathbf{F}_{a,b}^{(i)})(\mathbf{F}_{a,b} - \mathbf{F}_{a,b}^{(i)})^\intercal|Z] & 0 \\ 0 & \sum_{i=1}^n \mathbb{E}[(\mathbf{F}_{a,b} - \mathbf{F}_{a,b}^{(i)})^\intercal(\mathbf{F}_{a,b} - \mathbf{F}_{a,b}^{(i)})|Z]\end{bmatrix}\end{aligned}$$

We will use the following claim that we will prove later.

CLAIM 3.1. *We have the following relations.*

$$\sum_{i=1}^n \mathbb{E}[(\mathbf{F}_{a,b} - \mathbf{F}_{a,b}^{(i)})(\mathbf{F}_{a,b} - \mathbf{F}_{a,b}^{(i)})^\intercal|Z] = 2(b+1)\mathbf{F}_{a,b+1}\mathbf{F}_{a,b+1}^\intercal$$

$$\sum_{i=1}^n \mathbb{E}[(\mathbf{F}_{a,b} - \mathbf{F}_{a,b}^{(i)})^\intercal(\mathbf{F}_{a,b} - \mathbf{F}_{a,b}^{(i)})|Z] = 2(a+1)\mathbf{F}_{a+1,b}^\intercal\mathbf{F}_{a+1,b}$$

This gives $\mathbb{E}\operatorname{tr}\left[\mathbf{V}_{a,b}^t\right] = (b+1)^t\mathbb{E}\left\|\mathbf{F}_{a,b+1}\right\|_{2t}^{2t} + (a+1)^t\mathbb{E}\left\|\mathbf{F}_{a+1,b}\right\|_{2t}^{2t}$. Therefore, we get

$$\begin{aligned}
2\mathbb{E}\left\|\mathbf{X}_{a,b}\right\|_{2t}^{2t} &= \mathbb{E}\operatorname{tr}\left[\overline{\mathbf{X}}_{a,b}^{2t}\right] \\
&\leq (2(2t-1))^t\mathbb{E}\operatorname{tr}\left[\mathbf{V}_{a,b}^t\right] \\
&= (2(2t-1))^t((b+1)^t\mathbb{E}\left\|\mathbf{F}_{a,b+1}\right\|_{2t}^{2t} + (a+1)^t\mathbb{E}\left\|\mathbf{F}_{a+1,b}\right\|_{2t}^{2t}) \\
&= (2(2t-1))^t((b+1)^t\mathbb{E}\left\|\mathbf{X}_{a,b+1} + \mathbb{E}\mathbf{F}_{a,b+1}\right\|_{2t}^{2t} + (a+1)^t\mathbb{E}\left\|\mathbf{X}_{a+1,b} + \mathbb{E}\mathbf{F}_{a+1,b}\right\|_{2t}^{2t}) \\
&\leq (16t)^t((b+1)^t(\mathbb{E}\left\|\mathbf{X}_{a,b+1}\right\|_{2t}^{2t} + \left\|\mathbb{E}\mathbf{F}_{a,b+1}\right\|_{2t}^{2t}) + (a+1)^t(\mathbb{E}\left\|\mathbf{X}_{a+1,b}\right\|_{2t}^{2t} + \left\|\mathbb{E}\mathbf{F}_{a+1,b}\right\|_{2t}^{2t}) \\
&\leq (16td_p)^t(\mathbb{E}\left\|\mathbf{X}_{a,b+1}\right\|_{2t}^{2t} + \left\|\mathbb{E}\mathbf{F}_{a,b+1}\right\|_{2t}^{2t} + \mathbb{E}\left\|\mathbf{X}_{a+1,b}\right\|_{2t}^{2t} + \left\|\mathbb{E}\mathbf{F}_{a+1,b}\right\|_{2t}^{2t})
\end{aligned}$$

□

It remains to prove the claim.

Proof. [Proof of Claim 3.1] We will prove the first equality. The second one is analogous. For $I \in \mathcal{I}, J \in \mathcal{J}, \alpha, \beta \in \{0,1\}^n$, we have

$$(\mathbf{F}_{a,b} - \mathbf{F}_{a,b}^{(i)})[(I,\alpha),(J,\beta)] = \begin{cases} \nabla_{\alpha+\beta}(f_{I,J}(Z) - f_{I,J}(Z^{(i)})) & \text{if } |\alpha|_0 = a, |\beta|_0 = b, \alpha\cdot\beta = 0 \\ 0 & \text{o.w.} \end{cases}$$

By Proposition 3.1, the first expression simplifies to $(Z_i - \widetilde{Z}_i)\nabla_{\mathbf{e}_i}\nabla_{\alpha+\beta}f_{I,J}(Z)$. Define the matrix $\mathbf{F}_{a,b,i}$ to be the matrix with the same set of rows and columns as $\mathbf{F}_{a,b}$ and whose only nonzero entries are given by

$$\mathbf{F}_{a,b,i}[(I,\alpha),(J,\beta+\mathbf{e}_i)] = \nabla_{\mathbf{e}_i}\nabla_{\alpha+\beta}f_{I,J}(Z) \text{ if } |\alpha|_0 = a, |\beta|_0 = b, \beta\cdot\mathbf{e}_i = 0, \alpha\cdot(\beta+\mathbf{e}_i) = 0$$

Then, it's easy to see that $\sum_{i=1}^n \mathbf{F}_{a,b,i}\mathbf{F}_{a,b,i}^\intercal = (b+1)\mathbf{F}_{a,b+1}\mathbf{F}_{a,b+1}^T$ and $(\mathbf{F}_{a,b} - \mathbf{F}_{a,b}^{(i)})(\mathbf{F}_{a,b} - \mathbf{F}_{a,b}^{(i)})^\intercal = (Z - \widetilde{Z}_i)^2\mathbf{F}_{a,b,i}\mathbf{F}_{a,b,i}^\intercal$. The latter equality implies

$$\mathbb{E}[(\mathbf{F}_{a,b} - \mathbf{F}_{a,b}^{(i)})(\mathbf{F}_{a,b} - \mathbf{F}_{a,b}^{(i)})^\intercal|Z] = \mathbb{E}[(Z_i - \widetilde{Z}_i)^2\mathbf{F}_{a,b,i}\mathbf{F}_{a,b,i}^\intercal|Z] = 2\mathbf{F}_{a,b,i}\mathbf{F}_{a,b,i}^\intercal$$

Therefore,

$$\sum_{i=1}^n \mathbb{E}[(\mathbf{F}_{a,b} - \mathbf{F}_{a,b}^{(i)})(\mathbf{F}_{a,b} - \mathbf{F}_{a,b}^{(i)})^\intercal|Z] = 2\sum_{i=1}^n \mathbf{F}_{a,b,i}\mathbf{F}_{a,b,i}^\intercal = 2(b+1)\mathbf{F}_{a,b+1}\mathbf{F}_{a,b+1}^\intercal$$

□

4 Applications

To illustrate our framework, we apply it to obtain concentration bounds for nonlinear random matrices that have been considered in the literature before. The first application is a simple tensor network that arose in the analysis of spectral algorithms for a variant of principal components analysis (PCA) [HSS15, Hop18]. The second application is to obtain norm bounds on dense graph matrices [MP16, AMP16]. In the second application, the norm bounds are governed by a combinatorial structure called *the minimum vertex separator of a shape*. We will show how this notion arises naturally under our framework, whereas prior works that derived such bounds used the trace power method and required nontrivial combinatorial insights.

4.1 A simple tensor network

We consider the following result from [HSS15, Hop18]. We remark that this result could also be obtained via other standard techniques, but we showcase it as it serves as a simple warm-up to familiarize the reader with our method.

LEMMA 4.1. ([HOP18], THEOREM 6.7.1) *Let $c \in \{1,2\}$ and let $d \geq 1$ be an integer. Let $\mathbf{A}_1, \ldots, \mathbf{A}_{n^c}$ be i.i.d. random matrices uniformly sampled from $\{-1,1\}^{n^d \times n^d}$. Then, with probability $1 - O(n^{-100})$,*

$$\left\| \sum_{k \leq n^c} \mathbf{A}_k \otimes \mathbf{A}_k - \mathbb{E} \sum_{k \leq n^c} \mathbf{A}_k \otimes \mathbf{A}_k \right\| \leq C\sqrt{d} n^{(2d+c)/2} (\log n)^{1/2}$$

for an absolute constant $C > 0$.

Using our framework, we will prove a slightly relaxed version of the inequality where $\sqrt{d}(\log n)^{1/2}$ is replaced by $\log n$, while not losing on the dominating term $n^{(2d+c)/2}$. We remark that we have not attempted to optimize these extra factors in front of the dominating term $n^{(2d+c)/2}$, so it's plausible that a more careful analysis can obtain a slightly better bound.

Proof. [Proof of the relaxed bound] Let the i,j-th entry of $\mathbf{A}_k$ be $a_{k,i,j}$. Let $\mathbf{F} = \sum_{i \leq n^c} \mathbf{A}_k \otimes \mathbf{A}_k - \mathbb{E} \sum_{i \leq n^c} \mathbf{A}_k \otimes \mathbf{A}_k$ be a random matrix on the variables $a_{k,i,j}$ for $k \leq n^c, i, j \leq n^d$. So $\mathbb{E}\mathbf{F} = 0$ and we are looking for bounds on $\|\mathbf{F}\|$. The entries are given by

$$\mathbf{F}[(i_1,i_2),(j_1,j_2)] = \begin{cases} \sum_{k \leq n^c} a_{k,i_1,j_1} a_{k,i_2,j_2} & \text{if } (i_1,j_1) \neq (i_2,j_2) \\ 0 & \text{if } (i_1,j_1) = (i_2,j_2) \end{cases}$$

The nonzero entries are homogeneous polynomials of degree 2. UsingTheorem 1.2,

$$\mathbb{E} \|\mathbf{F}\|_{2t}^{2t} \leq (32t)^{2t} (\|\mathbb{E}\mathbf{F}_{2,0}\|_{2t}^{2t} + \|\mathbb{E}\mathbf{F}_{1,1}\|_{2t}^{2t} + \|\mathbb{E}\mathbf{F}_{0,2}\|_{2t}^{2t})$$

We will consider each of these terms. In the following arguments, we restrict attention to indices i_1, i_2, j_1, j_2 such that $(i_1,j_1) \neq (i_2,j_2)$.

1. $\mathbb{E}\mathbf{F}_{2,0}$ has nonzero entries in row $((i_1,i_2), \{(k,i_1,j_1),(k,i_2,j_2)\})$ and column (j_1,j_2) and all these entries are 1. The Schatten norm does not change when we permute the rows and columns. So, we can group the rows on k, i_1, i_2 and within each group, we can sort j_1, j_2 in both rows and columns. We get a matrix having n^{2d+c} identity matrices, each of dimensions $n^{2d} \times n^{2d}$, stacked on top of each other. Using the definition, the Schatten-$2t$ norm of this matrix is easily computed to be $\|\mathbb{E}\mathbf{F}_{2,0}\|_{2t}^{2t} = n^{c+4d} n^{t(2d+c)}$.

2. $\mathbb{E}\mathbf{F}_{1,1}$ has nonzero entries in either row $((i_1,i_2), \{(k,i_1,j_1)\})$ and column $((j_1,j_2), \{(k,i_2,j_2)\})$; or row $((i_1,i_2), \{(k,i_2,j_2)\})$ and column $((j_1,j_2), \{(k,i_1,j_1)\})$ and all these entries are 1. So we can write $\mathbb{E}\mathbf{F}_{1,1} = \mathbf{A} + \mathbf{B}$ corresponding to the 2 sets of entries. Arguing just as in the previous case, we can obtain $\|\mathbf{A}\|_{2t}^{2t} = n^{c+4d} n^{t(2d+c)}$ where we group the rows on k, i_2, j_1 and $\|\mathbf{B}\|_{2t}^{2t} = n^{c+4d} n^{t(2d+c)}$ where we group the rows on k, i_1, j_2. Therefore, $\|\mathbb{E}\mathbf{F}_{1,1}\|_{2t}^{2t} \leq 2^{2t}(\|\mathbf{A}\|_{2t}^{2t} + \|\mathbf{B}\|_{2t}^{2t}) = 2^{2t+1} n^{c+4d} n^{t(2d+c)}$.

3. The case $\mathbb{E}\mathbf{F}_{0,2}$ is identical to $\mathbb{E}\mathbf{F}_{2,0}$.

Putting them together, $\mathbb{E} \|\mathbf{F}\|_{2t}^{2t} \leq (C't)^{2t} n^{c+4d} n^{t(2d+c)}$ for an absolute constant $C' > 0$. Now, we apply Markov's inequality to get

$$Pr[\|\mathbf{F} - \mathbb{E}\mathbf{F}\| \geq \theta] \leq Pr[\|\mathbf{F} - \mathbb{E}\mathbf{F}\|_{2t}^{2t} \geq \theta^{2t}] \leq \theta^{-2t} \mathbb{E} \|\mathbf{F} - \mathbb{E}\mathbf{F}\|_{2t}^{2t} \leq \theta^{-2t} (C't)^{2t} n^{c+4d} n^{t(2d+c)}$$

We now set $\theta = \varepsilon^{-1/(2t)} (C't) n^{(c+4d)/t} n^{(2d+c)/2}$ to make this expression at most ε. Plug in $\varepsilon = n^{-100}$ and set $t = \log n$ to obtain that $\|\mathbf{F} - \mathbb{E}\mathbf{F}\| \leq C n^{(2d+c)/2} \log n$ holds with probability $1 - n^{-100}$, where $C > 0$ is an absolute constant. ∎

4.2 Graph matrices In this section, we first define graph matrices and then show how to obtain norm bounds for *dense graph matrices*, i.e. the case when $G \sim \mathcal{G}_{n,1/2}$, using our framework. Handling *sparse graph matrices*, i.e. the case when $G \sim \mathcal{G}_{n,p}$ for $p = o(1)$, may not work well with our basic framework as we will explain in Section 5. Instead, our general framework in Section 6 will handle this case well and we obtain sparse graph matrix norm bounds in Section 8.

4.2.1 Definitions Define by $\mathcal{G}_{n,p}$ the Erdős-Rényi random graph on the vertex set $[n]$ with n vertices, where each edge is present independently with probability p. Let the graph be encoded by variables $G_{i,j} \in \Omega = \{-\sqrt{\frac{1-p}{p}}, \sqrt{\frac{p}{1-p}}\}$ where $-\sqrt{\frac{1-p}{p}}$ indicates the presence of the edge $\{i,j\}$ and $\sqrt{\frac{p}{1-p}}$ indicates absence, for all $1 \le i,j \le n$.

So, each $G_{i,j}$ for $i < j$ is sampled from Ω where $G_{i,j}$ takes the value $-\sqrt{\frac{1-p}{p}}$ with probability p and takes the value $\sqrt{\frac{p}{1-p}}$ otherwise. Here, Ω has been normalized so that $\mathbb{E}_{x\sim\Omega}[x] = 0, \mathbb{E}_{x\sim\Omega}[x^2] = 1$. as is standard in p-biased Fourier analysis.

When $p = 1/2$, we are in the setting of *dense graph matrices*. Then, $\mathcal{G}_{n,1/2}$ can be thought of as a sampling of the $G_{i,j}, i < j$ independently and uniformly from $\Omega = \{-1, 1\}$. For a set of edges $E \subseteq \binom{[n]}{2}$, define $G_E := \prod_{e\in E} G_e$. When $p = 1/2$, the G_E correspond to the Fourier basis for functions of the graph.

Define $\mathcal{I}$ to be the set of sub-tuples of $[n]$, including the empty tuple. Graph matrices will have rows and columns indexed by $\mathcal{I}$. Each graph matrix has a succinct representation as a graph with some extra information, that is called a *shape*.

DEFINITION 4.1. (SHAPE) *A shape is a tuple $\tau = (V(\tau), E(\tau), U_\tau, V_\tau)$ where $(V(\tau), E(\tau))$ is a graph and U_τ, V_τ are ordered subsets of the vertices.*

DEFINITION 4.2. (REALIZATION) *Given a shape τ, a realization of τ is an injective map $\varphi : V(\tau) \to [n]$.*

DEFINITION 4.3. (GRAPH MATRICES) *Let τ be a shape. The graph matrix $\mathbf{M}_\tau : \{\pm 1\}^{\binom{n}{2}} \to \mathbb{R}^{\mathcal{I}\times\mathcal{I}}$ is defined to be the matrix-valued function with I, J-th entry defined as follows.*

$$\mathbf{M}_\tau[I,J] := \sum_{\substack{\text{Realization } \varphi \\ \varphi(U_\tau)=I, \varphi(V_\tau)=J}} G_{\varphi(E(\tau))} = \sum_{\substack{\text{Realization } \varphi \\ \varphi(U_\tau)=I, \varphi(V_\tau)=J}} \prod_{(u,v)\in E(\tau)} G_{\varphi(u),\varphi(v)}$$

In other words, we sum over all realizations of τ that map U_τ, V_τ to I, J respectively and for each such realization, we have a term corresponding to the Fourier character that the realization gives.

The following examples illustrate some simple graph matrices.

EXAMPLE 4.1. (ADJACENCY MATRIX) *Let τ be the shape on the left in Fig. 3, with two vertices $V(\tau) = \{u, v\}$ and a single edge $E(\tau) = \{\{u, v\}\}$. U_τ, V_τ are $(u), (v)$ respectively where we use tuples to indicate ordering. Then $\mathbf{M}_\tau$ has nonzero entries $\mathbf{M}_\tau[(i),(j)](G) = G_{i,j}$ for all $i \ne j$. If $G \in \{\pm 1\}^{\binom{n}{2}}$ is thought of as a graph, then $\mathbf{M}_\tau$ has as principal submatrix the ± 1 adjacency matrix of G with zeros on the diagonal, and the other entries are 0.*

EXAMPLE 4.2. *In Fig. 3, consider the shape τ on the right. We have $U_\tau = (u_1, u_2), V_\tau = (v_1), V(\tau) = \{u_1, u_2, v_1, w_1\}$ and $E(\tau) = \{\{u_1, w_1\}, \{u_2, w_1\}, \{w_1, v_1\}\}$. $\mathbf{M}_\tau$ is a matrix with rows and columns indexed by sub-tuples of $[n]$. Its nonzero entries are in rows I and columns J with $|I| = |U_\tau| = 2$ and $|J| = |V_\tau| = 1$ respectively. Specifically, for all distinct a_1, a_2, b_1, the entry corresponding to row (a_1, a_2) and column (b_1) is $\sum_{c_1\in[n]\setminus\{a_1,a_2,b_1\}} G_{a_1,c_1} G_{a_2,c_1} G_{c_1,b_1}$. Here,*

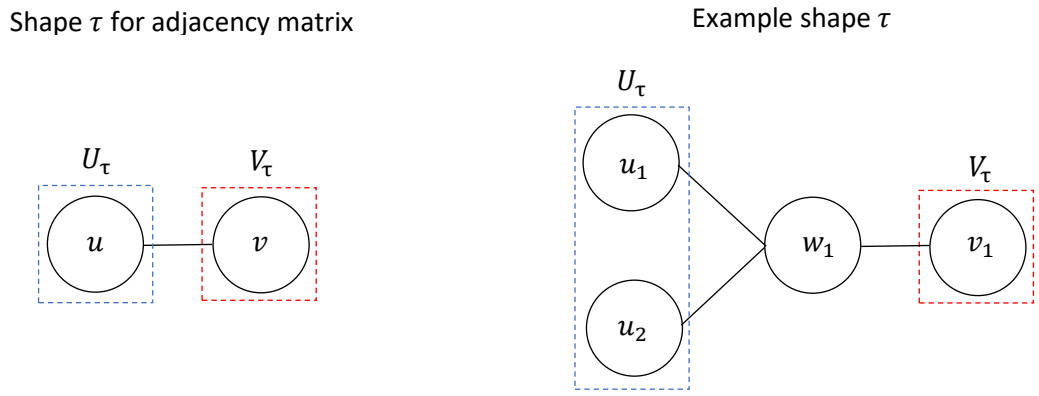

Figure 3: Left: Shape corresponding to adjacency matrix, Right: Example of a more complicated shape

each term is obtained via the realization φ that maps u_1, u_2, w_1, v_1 to a_1, a_2, c_1, b_1 respectively. Succinctly,

$$\mathbf{M}_\tau = \begin{array}{c} \\ \textit{row } (a_1,a_2) \rightarrow \end{array} \begin{array}{c} \textit{column } (b_1) \\ \downarrow \\ \left(\begin{array}{c} \cdots\cdots \sum_{c_1 \in [n]\setminus\{a_1,a_2,b_1\}} G_{a_1,c_1} G_{a_2,c_1} G_{c_1,b_1} \cdots\cdots \\ \\ \end{array} \right) \end{array}$$

Intuitively, graph matrices are symmetrizations of the Fourier basis, where the symmetry is incorporated by summing over all realizations of "free" vertices $V(\tau) \setminus U_\tau \setminus V_\tau$ of the shape τ. For more examples of graph matrices and why they can be a useful tool to work with, see [AMP16].

4.2.2 Norm bounds for dense graph matrices In this section, we study the concentration of the so-called "dense graph matrices" which is a term that refers to graph matrices M_τ in the setting $p = 1/2$. Since the edges of a random graph sampled from $\mathcal{G}_{n,1/2}$ can be viewed as independent Rademacher random variables, we can apply our framework in this setting.

In particular, we will obtain bounds on $\mathbb{E}\,\|\mathbf{M}_\tau - \mathbb{E}\mathbf{M}_\tau\|_{2t}^{2t}$. The $G_{i,j} \in \{-1,1\}$ correspond to the Z_is in Section 3 and for a fixed shape τ, $\mathbf{M}_\tau$ will be the matrix $\mathbf{F}$ we are interested in analyzing. For $I, J \in \mathcal{I}$, $\mathbf{M}_\tau[I,J]$ is a nonzero polynomial only when there exists at least one realization of τ that maps U_τ, V_τ to I, J respectively. In particular, we must have $|I| = |U_\tau|$ and $|J| = |V_\tau|$. In this case, $\mathbf{M}_\tau[I,J]$ is a homogenous polynomial of degree $|E(\tau)|$. ByTheorem 1.2, we have

$$\mathbb{E}\,\|\mathbf{M}_\tau - \mathbb{E}\mathbf{M}_\tau\|_{2t}^{2t} \;\le\; \sum_{\substack{a+b\ge 1 \\ a,b\ge 0}} (16t|E(\tau)|)^{(a+b)t}\, \|\mathbb{E}\mathbf{M}_{\tau,a,b}\|_{2t}^{2t}$$

where for integers $a, b \ge 0$, $\mathbf{M}_{\tau,a,b}$ is defined to be the matrix with rows and columns each indexed by $\mathcal{I} \times \{0,1\}^{\binom{n}{2}}$ such that for all $I, J \in \mathcal{I}$, we have

$$\mathbf{M}_{\tau,a,b}[(I,\alpha),(J,\beta)] \;=\; \begin{cases} \nabla_{\alpha+\beta}\mathbf{M}_\tau[I,J] & \text{if } |\alpha|_0 = a, |\beta|_0 = b, \alpha\cdot\beta = 0 \\ 0 & \text{o.w.} \end{cases}$$

For any multilinear homogenous polynomial f of degree d, since $\mathbb{E}[G_{i,j}] = 0$ for all i, j, we have $\nabla_\alpha f = 0$ whenever $|\alpha|_0 < d$. Therefore, $\mathbb{E}\mathbf{M}_{\tau,a,b} = 0$ for all $a + b < |E(\tau)|$. Moreover, $\mathbb{E}\mathbf{M}_{\tau,a,b} = 0$ whenever

$a+b \neq |E(G)|$ otherwise $\mathbb{E}\mathbf{M}_{\tau,a,b} = \mathbf{M}_{\tau,a,b}$. So, we can further simplify the above expression to

$$\mathbb{E}\left\|\mathbf{M}_\tau - \mathbb{E}\mathbf{M}_\tau\right\|_{2t}^{2t} \leq \sum_{\substack{a+b=|E(\tau)|\\ a,b\geq 0}} (16t|E(\tau)|)^{|E(\tau)|t} \left\|\mathbf{M}_{\tau,a,b}\right\|_{2t}^{2t}$$

It remains to analyze $\left\|\mathbf{M}_{\tau,a,b}\right\|_{2t}^{2t}$ for $a+b=|E(\tau)|$. We will see that analyzing these matrices is much simpler since they are deterministic matrices and simple computations using the Frobenius norm bound will work well. To state our final bounds, we need to define the notion of vertex separators of shapes.

REMARK 4.1. *As we will see, when analyzing the Frobenius norms for these deterministic matrices, the notion of the minimum vertex separator arises naturally. In prior trace method calculations (e.g. [MP16], [AMP16]), this required ingenious combinatorial observations.*

DEFINITION 4.4. (VERTEX SEPARATOR) *For a shape τ, define a vertex separator to be a subset of vertices $S \subseteq V(\tau)$ such that there is no path from U_τ to V_τ in $\tau \setminus S$, which is the shape obtained by deleting all the vertices of S (including all edges they're incident on).*

For a shape τ, denote by S_τ a vertex separator of the smallest size. Also, let I_τ be the set of isolated vertices (vertices with degree 0) in $V(\tau) \setminus U_\tau \setminus V_\tau$, so the presence of these vertices essentially scale the matrix by a scalar factor.

THEOREM 4.1. *For a shape τ and any integer $t \geq 1$,*

$$\mathbb{E}\left\|\mathbf{M}_\tau - \mathbb{E}\mathbf{M}_\tau\right\|_{2t}^{2t} \leq \left(C^{t|E(\tau)|} n^{|V(\tau)|} t^{t|E(\tau)|} |E(\tau)|^{2t|E(\tau)|}\right) n^{t(|V(\tau)|-|S_\tau|+|I_\tau|)}$$

for an absolute constant $C > 0$.

Up to lower order terms, the same result has been shown before in [MP16, AMP16]. To interpret this bound, assume that τ has a constant number of vertices. By setting $t \approx \text{polylog}(n)$, we get

$$\|\mathbf{M}_\tau\| = \widetilde{O}\left(\sqrt{n}^{|V(\tau)|-|S_\tau|+|I_\tau|}\right)$$

with high probability, where $\widetilde{O}$ hides logarithmic factors. This is obtained by applying Markov's inequality on the bound on $\mathbb{E}\left\|\mathbf{M}_\tau\right\|_{2t}^{2t}$. If τ has at least one edge, then $\mathbb{E}\mathbf{M}_\tau = 0$ andTheorem 4.1 yields such bounds. If τ has no edges, then it's quite simple to obtain such a bound and we include it in Lemma 4.2 for the sake of completeness.

Corollary 4.1 makes precise the high probability bound above. Therefore, this power of n is essentially what controls the norm bound and this is utilized heavily in applications (e.g. [BHK+19, GJJ+20, PR20]).

Proof. [Proof of Theorem 4.1] We first argue that we can assume $I_\tau = \emptyset$. This is because of the following reason. Each distinct vertex in τ of degree 0 essentially scales the matrix by a factor of at most n. And in the right hand side of the inequality, each vertex in I_τ contributes a factor of n^{2t} accordingly, from $n^{t|V(\tau)|}$ and from $n^{t|I_\tau|}$, and the other changes only weaken the inequality.

Now, fix $a, b \geq 0$ such that $a+b = |E(\tau)|$ and consider $\mathbf{M}_{\tau,a,b}$. For $I, J \in \mathcal{I}, \alpha, \beta \in \{0,1\}^{\binom{n}{2}}$ such that $|\alpha|_0 = a, |\beta|_0 = b, \alpha \cdot \beta = 0$, by definition,

$$\begin{aligned}\mathbf{M}_{\tau,a,b}[(I,\alpha),(J,\beta)] &= \nabla_{\alpha+\beta}\left(\sum_{\varphi:\varphi(U_\tau)=I,\varphi(V_\tau)=J} \prod_{u,v\in E(\tau)} G_{\varphi(u),\varphi(v)}\right)\\ &= |\{\varphi \mid \varphi(U_\tau) = I, \varphi(V_\tau) = J, \varphi(E(\tau)) = \text{Supp}(\alpha+\beta)\}|\end{aligned}$$

where Supp(.) denotes the support. We will now obtain norm bounds on these deterministic matrices by reinterpreting them as graph matrices for different shapes. Let $P = (E_1, E_2)$ denote the partition of $E(\tau) = E_1 \sqcup E_2$ into two ordered sets E_1, E_2, where $\sqcup$ denotes disjoint union. Let the set of ordered partitions P be $\mathcal{P}$. Then, we can write $\mathbf{M}_{\tau,a,b} = \sum_{P\in\mathcal{P}} \mathbf{M}_{\tau,a,b,P}$ where

$$\mathbf{M}_{\tau,a,b,P}[(I,\alpha),(J,\beta)] \;=\; |\{\varphi \mid \varphi(U_\tau) = I, \varphi(V_\tau) = J, \varphi(E_1) = \mathrm{Supp}(\alpha), \varphi(E_2) = \mathrm{Supp}(\beta)\}|$$

Also, $|\mathcal{P}| \le (4|E(\tau)|)^{|E(\tau)|}$ and so, by Fact 2.2,

$$\left\|\mathbf{M}_{\tau,a,b}\right\|_{2t}^{2t} \le (4|E(\tau)|)^{t|E(\tau)|} \sum_{P\in\mathcal{P}} \left\|\mathbf{M}_{\tau,a,b,P}\right\|_{2t}^{2t}$$

Each $\mathbf{M}_{\tau,a,b,P}$ can be interpreted as a graph matrix for a different shape τ_P, with the same vertex set and no edges. Let $V(\tau_P) = V(\tau), E(\tau_P) = \emptyset$ and set $U_{\tau_P} = U_\tau \cup V(E_1), V(\tau_P) = V_\tau \cup V(E_2)$ using a canonical ordering. Then, $\mathbf{M}_{\tau,a,b}$ is equal to $\mathbf{M}_{\tau_P}$ up to renaming of the rows and columns. For an illustration, see Fig. 4.

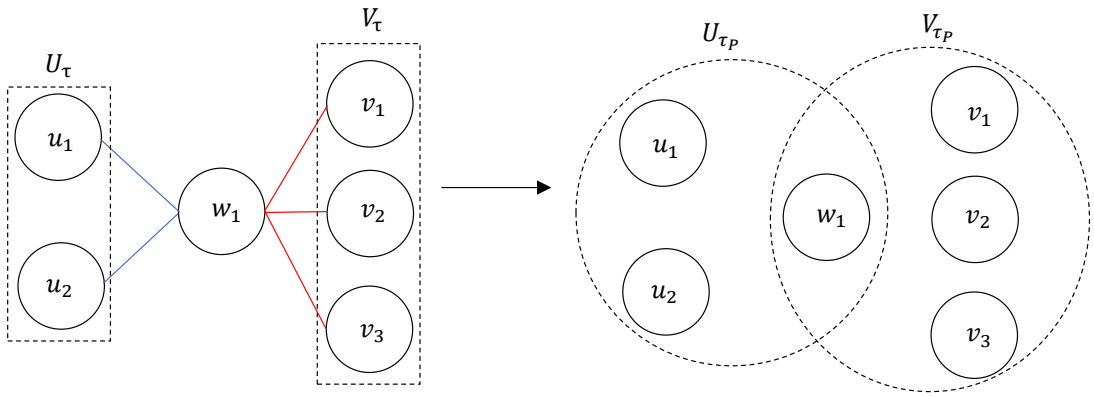

Figure 4: An example illustrating how τ_P is defined. In this example, P constraints the blue and red edges to go to α and β respectively. U_{τ_P}, V_{τ_P} have an ordering on the vertices (not shown here).

This graph matrix has a block diagonal structure indexed by the realizations of the set of common vertices $S = U_{\tau_P} \cap V_{\tau_P}$. Indeed, for $K \in [n]^S$, let $\mathbf{M}_{\tau_P,K}$ be the block of $\mathbf{M}_{\tau_P}$ with $\varphi(S) = K$. Then, $\mathbf{M}_{\tau_P,K}\mathbf{M}_{\tau_P,K'}^\intercal = \mathbf{M}_{\tau_P,K}^\intercal\mathbf{M}_{\tau_P,K'} = 0$ for $K \neq K'$ and so,

$$\begin{aligned}\mathbb{E}\left\|\mathbf{M}_{\tau,a,b}\right\|_{2t}^{2t} \le (4|E(\tau)|)^{t|E(\tau)|} \sum_{P\in\mathcal{P}} \left\|\mathbf{M}_{\tau_P}\right\|_{2t}^{2t} &= (4|E(\tau)|)^{t|E(\tau)|} \sum_{P\in\mathcal{P}} \sum_{T\in[n]^S} \left\|\mathbf{M}_{\tau_P,T}\right\|_{2t}^{2t} \\ &\le (4|E(\tau)|)^{t|E(\tau)|} \sum_{P\in\mathcal{P}} \sum_{T\in[n]^S} \left(\left\|\mathbf{M}_{\tau_P,T}\right\|_2^2\right)^t\end{aligned}$$

where we bounded the Schatten norm by the appropriate power of the Frobenius norm. For any fixed $K \in [n]^S$, the entries of $\mathbf{M}_{\tau_P,K}$ take values in $\{0,1\}$ and the number of nonzero entries is at most $n^{|V(\tau)|-|S|}$ because the realizations of vertices in S are fixed and the other vertices have at most n choices each. Therefore, $\left\|\mathbf{M}_{\tau_P,K}\right\|_2^2 \le n^{|V(\tau)|-|S|}$.

Finally, we bound $|S|$ to estimate how large this term can be over all possibilities of P. We argue that S blocks all paths from U_τ to V_τ. To see this, consider any path from U_τ to V_τ, it must contain an edge $(u,v) \in E(\tau)$ such that $u \in U_{\tau_P}, v \in V_{\tau_P}$. We must either have $(u,v) \in E_1$, in which case $u, v \in U_{\tau_P}$ and $v \in S$, or $(u,v) \in E_2$, in which case $u, v \in V_{\tau_P}$ and $u \in S$. In either case, S must contain either u or v. This argument implies S must be a vertex separator of τ, giving $|S| \ge |S_\tau|$. For a proof by picture, see Fig. 5.

We also have the trivial upper bound $|S| \le |V(\tau)|$. Ultimately, this gives

$$\left\|\mathbf{M}_{\tau,a,b}\right\|_{2t}^{2t} \le (4|E(\tau)|)^{t|E(\tau)|} \sum_{P\in\mathcal{P}} \sum_{T\in[n]^S} n^{t(|V(\tau)|-|S_\tau|)} \quad \le (4|E(\tau)|)^{t|E(\tau)|}(4|E(\tau)|)^{|E(\tau)|} n^{|V(\tau)|} n^{t(|V(\tau)|-|S_\tau|)}$$

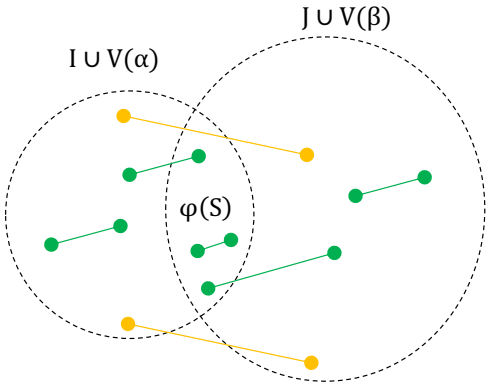

Figure 5: Proof by picture that $|S| \geq |S_\tau|$. Green edges can occur in τ, orange edges cannot, so S blocks all paths from U_τ to V_τ.

Along with our prior discussion, we get

$$\begin{aligned}
\mathbb{E}\,\|\mathbf{M}_\tau - \mathbb{E}\mathbf{M}_\tau\|_{2t}^{2t} &\leq \sum_{a+b=|E(\tau)|} (16t|E(\tau)|)^{|E(\tau)|t}\,\|\mathbf{M}_{\tau,a,b}\|_{2t}^{2t} \\
&\leq \sum_{a+b=|E(\tau)|} (16t|E(\tau)|)^{|E(\tau)|t}(4|E(\tau)|)^{t|E(\tau)|}(4|E(\tau)|)^{|E(\tau)|}n^{|V(\tau)|}n^{t(|V(\tau)|-|S_\tau|)} \\
&\leq \left(C^{t|E(\tau)|}n^{|V(\tau)|}t^{t|E(\tau)|}|E(\tau)|^{2t|E(\tau)|}\right)n^{t(|V(\tau)|-|S_\tau|)}
\end{aligned}$$

for an absolute constant $C > 0$. ∎

REMARK 4.2. *Note that while the proof of the norm bound above still requires some combinatorial analysis, this arises mostly from a mechanical application of the general result Theorem 1.2. Also, one only needs the simpler combinatorics of the fixed-size shapes obtained from the given shape τ, rather than increasingly large shapes formed by combining copies of τ, as in the application of trace method [AMP16].*

In the proof above, our analysis of the shape τ_P which has no edges, applies in general to any shape τ with no edges. For the sake of completeness, we state it explicity in the following lemma.

LEMMA 4.2. *For a shape τ with no edges and any integer $t \geq 1$, $\mathbb{E}\,\|\mathbf{M}_\tau\|_{2t}^{2t} \leq n^{|U_\tau \cap V_\tau|}n^{t(V(\tau)-|U_\tau\cap V_\tau|+|I_\tau|)}$.*

Note that this has the same form as Theorem 4.1 because for a shape τ with no edges, the minimum vertex separator S_τ is just $U_\tau \cap V_\tau$. The following corollary obtains high probability norm bounds for norms of graph matrices via Markov's inequality.

COROLLARY 4.1. *For a shape τ, for any constant $\varepsilon > 0$, with probability $1 - \varepsilon$,*

$$\|\mathbf{M}_\tau\| \leq (C|E(\tau)|\log(n^{|V(\tau)|}/\varepsilon))^{|E(\tau)|} \cdot \sqrt{n}^{|V(\tau)|-|S_\tau|+|I_\tau|}$$

for an absolute constant $C > 0$.

Proof. If $E(\tau) = \emptyset$, we invoke Lemma 4.2. Otherwise, $\mathbb{E}\mathbf{M}_\tau = 0$ and we invoke Theorem 4.1. By an application of Markov's inequality,

$$\begin{aligned}
Pr[\|\mathbf{M}_\tau\| \geq \theta] \leq Pr[\|\mathbf{M}_\tau\|_{2t}^{2t} \geq \theta^{2t}] &\leq \theta^{-2t}\mathbb{E}\,\|\mathbf{M}_\tau\|_{2t}^{2t} \\
&\leq \theta^{-2t}\left((C')^{t|E(\tau)|}n^{|V(\tau)|}t^{t|E(\tau)|}|E(\tau)|^{2t|E(\tau)|}\right)n^{t(|V(\tau)|-|S_\tau|+|I_\tau|)}
\end{aligned}$$

for an absolute constant $C' > 0$. To make this expression at most ε, we simply set $\theta = \left(\varepsilon^{-1/(2t)}(C'')^{|E(\tau)|}n^{|V(\tau)|/(2t)}t^{|E(\tau)|/2}|E(\tau)|^{|E(\tau)|}\right)\sqrt{n}^{|V(\tau)|-|S_\tau|+|I_\tau|}$ for an absolute constant $C'' > 0$. Finally, set $t = \frac{1}{2}\log(n^{|V(\tau)|}/\varepsilon)$ to complete the proof. ∎

5 Why a naïve application of [PMT16] may fail for general product distributions

In this section, we elaborate on the difficulties that arise when working with random variables that are not necessarily Rademacher. In this case, note that we cannot assume that the polynomial entries are multilinear as well.

To recall the setting, we are given a random matrix $\mathbf{F}$ whose entries are low degree polynomials in random variables $Z_1, \ldots, Z_n$ which are independently sampled from arbitrary distributions. And we wish to obtain concentration bounds on how much $\mathbf{F}$ can deviate from its mean, by way of controlling $\mathbb{E}\,\|\mathbf{F} - \mathbb{E}\mathbf{F}\|_{2t}^{2t}$.

Building on the ideas from Section 3, we could attempt to use matrix Efron-Stein,Theorem 1.1 and hope to obtain a similar recursion framework. We now discuss what happens if we do this. Assume $\mathbb{E}[Z_i] = 0, \mathbb{E}[Z_i^2] = 1$. We can proceed similar to the proof of Theorem 1.2. So, we consider $\mathbf{X}$ as a principal submatrix of $\mathbf{X}_{0,0}$ and follow through Lemma 3.1. The main change will happen in Claim 3.1. In particular, the equation $\mathbb{E}[(Z_i - \widetilde{Z}_i)^2|Z] = 2$ is no longer true. Instead, we will have $\mathbb{E}[(Z_i - \widetilde{Z}_i)^2|Z] = 1 + Z_i^2$. So, we get the expression

$$\sum_{i=1}^{n}(1+Z_i^2)\mathbf{F}_{a,b,i}\mathbf{F}_{a,b,i}^\intercal = \sum_{i=1}^{n}\mathbf{F}_{a,b,i}\mathbf{F}_{a,b,i}^\intercal + \sum_{i=1}^{n}Z_i^2\mathbf{F}_{a,b,i}\mathbf{F}_{a,b,i}^\intercal$$

The first term can been handled just as in the basic framework. Unfortunately, the second term will be a source of difficulty. To get around this difficulty, we could attempt to apply the matrix Efron-Stein inequality again on an appropriately constructed matrix. To do this, we can interpret the second term as having been obtained after differentiating with respect to the variable Z_i and then *putting the variable back*. In contrast, we didn't need to put it back when working with Rademacher random variables. But after we do this, when we recurse on these extra matrices, the new second term will contain the left hand side as a sub-term, thereby giving a trivial inequality and stalling the recursion.

To see this more clearly, consider the simplest case $a = b = 0$. Then, the first term $\sum_{i=1}^{n}\mathbf{F}_{a,b,i}\mathbf{F}_{a,b,i}^\intercal$ will be equal to $\mathbf{F}_{0,1}\mathbf{F}_{0,1}^\intercal$ as we saw earlier. To evaluate the second term $\sum_{i=1}^{n}Z_i^2\mathbf{F}_{a,b,i}\mathbf{F}_{a,b,i}^\intercal$ in a similar manner, we define the matrix $\mathbf{H}$ to be the same as $\mathbf{F}_{0,1}$ except that each entry is now multiplied by Z_i where i is the differentiated variable in the column. That is, $\mathbf{H}[I,(J,\mathbf{e}_i)] = Z_i\mathbf{F}_{0,1}[I,(J,\mathbf{e}_i)]$. Observe that in the definition of $\mathbf{H}$, Z_i has been put back after differentiating with respect to it. Then, the second term will be $\mathbf{H}\mathbf{H}^\intercal$ and we can hope to use Efron-Stein again on this matrix $\mathbf{H}$ recursively.

We could do that and proceed similarly to the proof of Lemma 3.1 with appropriate modifications as above. But since $\beta_i = 1$ already, differentiating with respect to Z_i and putting it back, will return the same matrix $\mathbf{H}$! So, we end up with an inequality of the form

$$\mathbb{E}\,\|\mathbf{H}\|_{2t}^{2t} \le \mathrm{O}(t)^t(\mathbb{E}\,\|\mathbf{H}\|_{2t}^{2t} + \text{ other nonnegative terms})$$

Indeed, this is a tautology and will not be useful to us.

For a quick and dirty bound, suppose we had a parameter L such that $1 + Z_i^2 \le L$ for our distributions, then we will be able to obtain a similar framework while incurring a loss of $\sqrt{L}$ at each step of the recursion. But unfortunately, this bound will be lossy. For example, if we do this computation for the centered normalized adjacency matrix of $G \sim \mathcal{G}_{n,p}$, we will obtain a norm bound of $\widetilde{\mathrm{O}}(\frac{\sqrt{n(1-p)}}{\sqrt{p}})$ where $\widetilde{\mathrm{O}}$ hides logarithmic factors.. This bound is tight for constant or even inverse polylogarithmic p. But for $p = n^{-\theta}$ for some constant $0 < \theta < 1$, this is not tight because in this regime, the true norm bound is known to be $\widetilde{\mathrm{O}}(\sqrt{n})$ (see the early works of [FK81, Vu05] and for tighter bounds, see [BGBK20] and references therein).

If we dig into the details of what happened, this example illustrates that the matrix Efron-Stein inequalityTheorem 1.1 becomes a tautology for certain kinds of matrices, that yield $\mathbf{V} = \mathrm{O}(1)\mathbf{X}\mathbf{X}^\intercal +$

other positive semidefinite matrices.

But in our framework in general, the aforementioned bad matrices occur when we differentiate with respect to variables that have already been differentiated on. In other words, the current definition of the variance proxy $\mathbf{V}$ doesn't take into account whether we have already differentiated with respect to some variable Z_i. So, for the general recursion, we dive into the proof due to [PMT16] and modify it using structural properties of the intermediate matrices we obtain in our framework.

6 The general recursion framework

We now assume $Z_1, \ldots, Z_n$ are i.i.d. random variables sampled from a distribution Ω with finite moments. We assume that they are identically distributed for simplicity but our technique easily extends even when they are not identically distributed, as long as they are independent. For each $i \leq n$, define $\widetilde{Z}_i$ to be an independent copy of Z_i and define the vector $Z^{(i)} := (Z_1, \ldots, Z_{i-1}, \widetilde{Z}_i, Z_{i+1}, \ldots, Z_n)$. Define Z' to be the random vector defined by sampling i from $[n]$ uniformly at random and then setting $Z' = Z^{(i)}$.

Let $\mathbf{F} \in \mathbb{R}[Z]^{\mathcal{I}\times\mathcal{J}}$ be a matrix with rows and columns indexed by arbitrary sets $\mathcal{I}, \mathcal{J}$ respectively such that for all $I \in \mathcal{I}, J \in \mathcal{J}$, $\mathbf{F}[I,J]$ are polynomials of $Z_1, \ldots, Z_n$. Let d_p be the maximum degree of $\mathbf{F}[I,J]$ over all entries I, J and let d be the maximum degree of Z_i over all entries $\mathbf{F}[I,J]$ and $i \leq n$.

Similar to the Rademacher case, let $\mathbf{X} := \mathbf{F} - \mathbb{E}\mathbf{F}$. When the input is Z, we denote the matrices as $\mathbf{F}, \mathbf{X}$, etc and when the input is $Z^{(i)}$, denote the corresponding matrices as $\mathbf{F}^{(i)}, \mathbf{X}^{(i)}$, etc. In this section, we will give a general framework using which we can obtain bounds on $\mathbb{E}\,\|\mathbf{F} - \mathbb{E}\mathbf{F}\|_{2t}^{2t}$ for any integer $t \geq 1$. We set up a few preliminaries in order to state the main theorem.

DEFINITION 6.1. (SPACE $\mathcal{S}$) *Let $\mathcal{S}$ be the space of mean-zero polynomials in $Z_1, \ldots, Z_n$ of degree at most d_p.*

For $\alpha \neq 0$, we also define the centered monomials $\chi_\alpha(Z) = \prod_{\alpha_i>0}(Z_i^{\alpha_i} - \mathbb{E}[Z_i^{\alpha_i}])$. By definition, $\chi_\alpha \in \mathcal{S}$ for all $\alpha \neq 0, |\alpha|_1 \leq d_p$. The following proposition is straightforward.

PROPOSITION 6.1. *The set $\{\chi_\alpha(Z) | 1 \leq |\alpha|_1 \leq d_p\}$ forms a basis for $\mathcal{S}$.*

For the general framework, we work over this basis because as we will see in Section 7, the "inner kernel matrix" is convenient to state in this basis. The ∇ operator also works nicely with our polynomials χ_β. Indeed, observe that $\nabla_\alpha(\chi_\beta) = \begin{cases} \chi_{\beta-\alpha} & \text{if } \alpha \trianglelefteq \beta \\ 0 & \text{o.w.} \end{cases}$.

For a polynomial $f(Z)$ in $\mathcal{S}$, denote by $\widehat{f}(\alpha)$ the coefficient of $\chi_\alpha(Z)$ in the expansion of f, that is, $f(Z) = \sum_{0\neq\alpha\in\mathbb{N}^n} \widehat{f}(\alpha)\chi_\alpha(Z)$. We can naturally extend this notation to matrices that have mean 0. So, we can write $\mathbf{X} = \sum_{\alpha\neq 0} \widehat{\mathbf{X}}(\alpha)\chi_\alpha(Z)$ where $\widehat{\mathbf{X}}(\alpha)$ are deterministic matrices. In order to apply our recursion framework, we group this sum into terms based on $|\alpha|_0$. For $k \geq 1$, define $\mathbf{X}_k = \sum_{|\alpha|_0=k} \widehat{\mathbf{X}}(\alpha)\chi_\alpha(Z)$. Then, $\mathbf{X} = \sum_{k\geq 1} \mathbf{X}_k$. Note that when $k > d_p$, $\mathbf{X}_k = 0$.

DEFINITION 6.2. (INDEXING SET $\mathcal{K}$) *We define $\mathcal{K} \subseteq \mathbb{N}^n \times \{0,1\}^n$ to be the set of pairs (α, γ) such that $|\alpha|_1 \leq d_p, \alpha \in \mathbb{N}^n$ and $\gamma \leq \alpha$ with $\gamma \in \{0,1\}^n$.*

REMARK 6.1. *If we assume that the maximum degree of our polynomials d_p is constant, then the size of $\mathcal{K}$ is polynomially large, not exponentially large. Hence, the matrices we will consider below will also be of polynomial size when d_p is constant.*

Define the diagonal matrices $\mathbf{D}_1 \in \mathbb{R}[Z]^{\mathcal{I}\times\mathcal{K}} \times \mathbb{R}[Z]^{\mathcal{I}\times\mathcal{K}}$ and $\mathbf{D}_2 \in \mathbb{R}[Z]^{\mathcal{J}\times\mathcal{K}} \times \mathbb{R}[Z]^{\mathcal{J}\times\mathcal{K}}$ with nonzero entries

$$\mathbf{D}_1[(I,\alpha,\gamma),(I,\alpha,\gamma)] = \sqrt{\mathbb{E}[Z^{2\alpha\cdot(1-\gamma)}]}Z^{\alpha\cdot\gamma}, \qquad \mathbf{D}_2[(J,\alpha,\gamma),(J,\alpha,\gamma)] = \sqrt{\mathbb{E}[Z^{2\alpha\cdot(1-\gamma)}]}Z^{\alpha\cdot\gamma}$$

DEFINITION 6.3. (MATRICES $\mathbf{G}_{k,a,b}, \mathbf{F}_{k,a,b}$) *For integers k, a, b such that $k \geq 1, a, b \geq 0$, define the matrix $\mathbf{G}_{k,a,b}$ to have rows and columns indexed by $\mathcal{I} \times \mathcal{K}$ and $\mathcal{J} \times \mathcal{K}$ respectively such that for all $(I, \alpha_1, \gamma_1) \in \mathcal{I} \times \mathcal{K}$, $(J, \alpha_2, \gamma_2) \in \mathcal{J} \times \mathcal{K}$,*

$$\mathbf{G}_{k,a,b}[(I, \alpha_1, \gamma_1), (J, \alpha_2, \gamma_2)] = \begin{cases} \nabla_{\alpha_1+\alpha_2} \mathbf{X}_k[I, J] & \text{if } |\alpha_1|_0 = a, |\alpha_2|_0 = b, \alpha_1 \cdot \alpha_2 = 0 \\ 0 & \text{o.w.} \end{cases}$$

Also, define $\mathbf{F}_{k,a,b} := \mathbf{D}_1 \mathbf{G}_{k,a,b} \mathbf{D}_2$.

Note that when $k > d_p$, $\mathbf{F}_{k,a,b} = 0$.

PROPOSITION 6.2. *For integers k, a, b such that $k \geq 1, a, b \geq 0$, suppose $a + b < k$. Then each nonzero entry f of $\mathbf{G}_{k,a,b}$ has the property that $\widehat{f}(\alpha)$ is nonzero only when $|\alpha|_0 = k - a - b$*

Proof. The nonzero entries of $\mathbf{X}_k$ only has terms containing exactly k variables and $\nabla_{\alpha_1+\alpha_2}$ either zeroes out the term, or it truncates exactly $|\alpha_1 + \alpha_2|_0 = |\alpha_1|_0 + |\alpha_2|_0 = a + b$ variables. ∎

This also immediately implies that $\mathbb{E}[\mathbf{G}_{k,a,b}] = 0$ whenever $a + b < k$. Finally, when $k = a + b$, we have that $\mathbf{G}_{k,a,b}$ is a deterministic matrix independent of the Z_i. These give rise to the matrices $\mathbf{F}_{a+b,a,b}$ that appears in our main theorem. We are now ready to state the main theorem.

THEOREM 6.1. (GENERAL RECURSION) *Let the tuple of random variables Z and the function $\mathbf{F}$ be as above. Then, for all integers $t \geq 1$,*

$$\mathbb{E} \left\| \mathbf{F} - \mathbb{E}\mathbf{F} \right\|_{2t}^{2t} \leq \sum_{a,b \geq 0, a+b \geq 1} (Ct^2 d d_p^4)^{(a+b)t} \mathbb{E} \left\| \mathbf{F}_{a+b,a,b} \right\|_{2t}^{2t}$$

for an absolute constant $C > 0$.

Note that $\mathbf{F}_{a+b,a,b} = \mathbf{D}_1 \mathbf{G}_{a+b,a,b} \mathbf{D}_2$ where $\mathbf{D}_1, \mathbf{D}_2$ are diagonal matrices and $\mathbf{G}_{a+b,a,b}$ is a deterministic matrix that's independent of Z. To analyze the expected Schatten norm of such matrices, we can resort to far simpler techniques. For instance, we can obtain a simple bound using an appropriate power of the Frobenius norm, and apply standard scalar concentration tools. We will see an example of this in Section 8.

REMARK 6.1. *We have made no attempts to optimize the factors in front of the expectation inTheorem 6.1, which we suspect can be improved.*

We prove the main theorem by repeatedly applying the following technical lemma, the proof of which we defer to the next section.

LEMMA 6.1. *For all integers $t \geq 1$, integers $k \geq 1, a, b \geq 0$ such that $a + b < k$,*

$$\mathbb{E} \left\| \overline{\mathbf{F}}_{k,a,b} \right\|_{2t}^{2t} \leq (Ct^2 d d_p^2)^t (\mathbb{E} \left\| \overline{\mathbf{F}}_{k,a,b+1} \right\|_{2t}^{2t} + \mathbb{E} \left\| \overline{\mathbf{F}}_{k,a+1,b} \right\|_{2t}^{2t})$$

for an absolute constant $C > 0$.

Using this lemma, we can complete the proof of the main theorem.

Proof. [Proof of Theorem 6.1] Using Fact 2.2, we have $\mathbb{E} \left\| \mathbf{X} \right\|_{2t}^{2t} \leq d_p^{2t} \sum_{k=1}^{d_p} \mathbb{E} \left\| \mathbf{X}_k \right\|_{2t}^{2t}$. Note that for any $k \geq 1$, the matrix $\mathbf{X}_k$ is a principal submatrix of $\mathbf{F}_{k,0,0}$ with all other entries being 0, so $\mathbb{E} \left\| \mathbf{X}_k \right\|_{2t}^{2t} = \mathbb{E} \left\| \mathbf{F}_{k,0,0} \right\|_{2t}^{2t} = \frac{1}{2} \mathbb{E} \left\| \overline{\mathbf{F}}_{k,0,0} \right\|_{2t}^{2t}$. Therefore, $\mathbb{E} \left\| \mathbf{X} \right\|_{2t}^{2t} \leq \frac{1}{2} d_p^{2t} \sum_{k=1}^{d_p} \mathbb{E} \left\| \overline{\mathbf{F}}_{k,0,0} \right\|_{2t}^{2t}$. We now apply Lemma 6.1 repeatedly to all our terms until $k = a + b$, ultimately giving

$$\mathbb{E} \left\| \mathbf{X} \right\|_{2t}^{2t} \leq \frac{1}{2} d_p^{2t} (Ct^2 d d_p^2)^{(a+b)t} \sum_{a,b \geq 0, a+b \geq 1} \mathbb{E} \left\| \overline{\mathbf{F}}_{a+b,a,b} \right\|_{2t}^{2t}$$

Observing that $\mathbb{E} \left\| \overline{\mathbf{F}}_{a+b,a,b} \right\|_{2t}^{2t} = 2\mathbb{E} \left\| \mathbf{F}_{a+b,a,b} \right\|_{2t}^{2t}$ completes the proof. ∎

7 A generalization of [PMT16] and proof of Lemma 6.1

In this section, we will prove Lemma 6.1 using the high level strategy described in Section 1. This requires generalizing the results in [PMT16], and the proof techniques may be of independent interest.

7.1 Generalizing [PMT16] via explicit inner kernels In our setting, observe that (Z, Z') has the same distribution as (Z', Z). This is what is known as an *exchangeable pair* of variables, that will be extremely useful for our analysis. In particular, Z, Z' have the same distribution and $\mathbb{E}f(Z, Z') = \mathbb{E}f(Z', Z)$ for every integrable function f.

DEFINITION 7.1. (LAPLACIAN OPERATOR $\mathcal{L}$) *Define the operator $\mathcal{L}$ on the space $\mathcal{S}$ as $\mathcal{L}(f)(Z) = \mathbb{E}[f(Z) - f(Z')|Z]$ for all polynomials $f \in \mathcal{S}$.*

Note that this operator is well-defined since for any $f \in \mathcal{S}$, $\mathbb{E}[L(f)] = \mathbb{E}[\mathbb{E}[f(Z) - f(Z')|Z]] = \mathbb{E}[f(Z) - f(Z')] = 0$ and hence, $L(f) \in \mathcal{S}$.

LEMMA 7.1. *For all $\alpha \in \mathbb{N}^n$, χ_α is an eigenvector of $\mathcal{L}$ with eigenvalue $\frac{|\alpha|_0}{n}$.*

Proof. Recall that Z' is obtained by choosing $i \in [n]$ uniformly at random and then setting $Z' = Z^{(i)}$. Therefore, $\mathcal{L}(\chi_\alpha)(Z) = \mathbb{E}[\chi_\alpha(Z) - \chi_\alpha(Z')|Z] = \frac{1}{n}\sum_{i \le n} \mathbb{E}[\chi_\alpha(Z) - \chi_\alpha(Z^{(i)})|Z]$ When $\alpha_i = 0, \chi_\alpha(Z) - \chi_\alpha(Z^{(i)}) = 0$. Otherwise, $\mathbb{E}[\chi_\alpha(Z) - \chi_\alpha(Z^{(i)})|Z] = \chi_\alpha(Z)$. Therefore, the above expression simplifies to $\frac{|\alpha|_0}{n}\chi_\alpha(Z)$. ∎

THEOREM 7.1. (EXPLICIT KERNEL) *For any mean-centered polynomial $f \in \mathcal{S}$, there exists a polynomial K_f on $2n$ variables $z_1, \dots, z_n, z'_1, \dots, z'_n$, denoted collectively as (z, z'), with the following properties*

1. *$K_f(z', z) = -K_f(z, z')$*
2. *$\mathbb{E}[K_f(Z, Z')|Z] = f(Z)$ where (Z, Z') is the exchangeable pair we consider above.*

Proof. Using Proposition 6.1 and Lemma 7.1, under the basis of polynomials χ_α, the operator $\mathcal{L}$ is a diagonal matrix with nonzero diagonal entries and therefore, $\mathcal{L}^{-1}$ exists and is explicitly given by $\mathcal{L}^{-1}(f)(Z) = \sum_\alpha \frac{n}{|\alpha|_0}\widehat{f}(\alpha)\chi_\alpha(Z)$. We then take $K_f(z, z') = \mathcal{L}^{-1}(f)(z) - \mathcal{L}^{-1}(f)(z')$. The first condition is obvious and for the second condition, we have

$$\mathbb{E}[K_f(Z, Z')|Z] = \mathbb{E}[\mathcal{L}^{-1}(f)(Z) - \mathcal{L}^{-1}(f)(Z')|Z] = \mathcal{L}(\mathcal{L}^{-1}(f)) = f$$

∎

As seen in the proof of Theorem 7.1, $\mathcal{L}$ has a well-defined inverse $\mathcal{L}^{-1}$. We now define the matrix $\mathbf{K}_{k,a,b}$ that we call the *inner kernel*.

DEFINITION 7.2. (THE INNER KERNEL MATRIX $\mathbf{K}_{k,a,b}$) *For integers $k \ge 1, a, b \ge 0$ such that $a + b < k$, define the matrix $\mathbf{K}_{k,a,b} \in \mathbb{R}[Z]^{\mathcal{I}\times\mathcal{K}} \times \mathbb{R}[Z]^{\mathcal{J}\times\mathcal{K}}$ taking $2n$ variables $(z, z') = (z_1, \dots, z_n, z'_1, \dots, z'_n)$ as input as $\mathbf{K}_{k,a,b}(z, z') = \mathcal{L}^{-1}(\mathbf{G}_{k,a,b})(z) - \mathcal{L}^{-1}(\mathbf{G}_{k,a,b})(z')$.*

In the rest of this section except where explicitly stated, fix integers $k \ge 1, a, b \ge 0$ such that $a + b < k$. Then, the inner kernel $\mathbf{K}_{k,a,b}$ is well-defined.

LEMMA 7.2. $\mathbf{K}_{k,a,b}(Z, Z') = \frac{n}{k-a-b}(\mathbf{G}_{k,a,b}(Z) - \mathbf{G}_{k,a,b}(Z'))$

Proof.

$$\begin{aligned}\mathbf{K}_{k,a,b}(Z,Z') &= \mathcal{L}^{-1}(\mathbf{G}_{k,a,b})(Z) - \mathcal{L}^{-1}(\mathbf{G}_{k,a,b})(Z') \\ &= \sum_{|\alpha|_0=k-a-b} \widehat{\mathbf{G}_{k,a,b}}(\alpha)(\mathcal{L}^{-1}(\chi_\alpha)(Z) - \mathcal{L}^{-1}(\chi_\alpha)(Z')) \\ &= \frac{n}{k-a-b} \sum_{|\alpha|_0=k-a-b} \widehat{\mathbf{G}_{k,a,b}}(\alpha)(\chi_\alpha(Z) - \chi_\alpha(Z')) \\ &= \frac{n}{k-a-b}(\mathbf{G}_{k,a,b}(Z) - \mathbf{G}_{k,a,b}(Z'))\end{aligned}$$

□

The following lemma postulates important properties of the the inner kernel, including how it interacts with $\mathbf{D}_1$ and $\mathbf{D}_2$.

LEMMA 7.3. *$\mathbf{K}_{k,a,b}$ satisfies the following properties*

1. $\mathbf{K}_{k,a,b}(z',z) = -\mathbf{K}_{k,a,b}(z,z')$
2. $\mathbb{E}[\mathbf{K}_{k,a,b}(Z,Z')|Z] = \mathbf{G}_{k,a,b}(Z)$
3. $(\mathbf{D}_1(Z) - \mathbf{D}_1(Z'))\mathbf{K}_{k,a,b}(Z,Z') = \mathbf{K}_{k,a,b}(Z,Z')(\mathbf{D}_2(Z) - \mathbf{D}_2(Z')) = 0.$

Proof. The first equality is obvious from the definition. For the second equality, note that $\mathbb{E}[\mathbf{G}_{k,a,b}] = 0$ and $\mathbf{K}_{k,a,b}$ is defined by replacing each entry f of $\mathbf{G}_{k,a,b}$ by the kernel polynomial K_f as exhibited in Theorem 7.1. Now, we prove the third equality.

Consider the matrix $(\mathbf{D}_1(Z) - \mathbf{D}_1(Z'))\mathbf{K}_{k,a,b}(Z,Z')$ whose $[(I,\alpha_1,\gamma_1),(J,\alpha_2,\gamma_2)]$ entry is given by

$$\frac{n}{k-a-b}\sqrt{\mathbb{E}[Z^{2\alpha_1\cdot(1-\gamma_1)}]}(Z^{\alpha_1\cdot\gamma_1} - (Z')^{\alpha_1\cdot\gamma_1})(\nabla_{\alpha_1+\alpha_2}\mathbf{X}_k[I,J](Z) - \nabla_{\alpha_1+\alpha_2}\mathbf{X}_k[I,J](Z'))$$

where we have used Lemma 7.2. We will argue that this term is identically 0. We must have $Z' = Z^{(i)}$ for some $i \le n$. If $(\alpha_1\cdot\gamma_1)_i = 0$, then $Z^{\alpha_1\cdot\gamma_1} = (Z')^{\alpha_1\cdot\gamma_1}$ and the above term is 0. Otherwise, $(\alpha_1+\alpha_2)_i \neq 0$ and so $\nabla_{\alpha_1+\alpha_2}$ on any polynomial f will only contain the terms independent of Z_i, in which case $\nabla_{\alpha_1+\alpha_2}\mathbf{X}_k[I,J](Z) = \nabla_{\alpha_1+\alpha_2}\mathbf{X}_k[I,J](Z')$. In this case was well, the above term is 0. The proof of the other equality is analogous. □

The reason we call $\mathbf{K}_{k,a,b}$ the inner kernel is because, as seen above, it serves as a kernel for the inner matrix $\mathbf{G}$ in the decomposition $\mathbf{F} = \mathbf{DGD}$. Since we will need to work with Hermitian dilations, we define $\mathbf{D} = \begin{bmatrix}\mathbf{D}_1 & 0\\ 0 & \mathbf{D}_2\end{bmatrix}$. We will use the following basic fact extensively in our manipulations.

FACT 7.1. *For any matrix $\mathbf{A} \in \mathbb{R}[Z]^{\mathcal{I}\times\mathcal{K}} \times \mathbb{R}[Z]^{\mathcal{J}\times\mathcal{K}}$, $\mathbf{D}\overline{\mathbf{A}}\mathbf{D} = \overline{\mathbf{D}_1\mathbf{A}\mathbf{D}_2}$.*

Proof. We have

$$\begin{aligned}\mathbf{D}\overline{\mathbf{A}}\mathbf{D} = \begin{bmatrix}\mathbf{D}_1 & 0\\ 0 & \mathbf{D}_2\end{bmatrix}\begin{bmatrix}0 & \mathbf{A}\\ \mathbf{A}^\intercal & 0\end{bmatrix}\begin{bmatrix}\mathbf{D}_1 & 0\\ 0 & \mathbf{D}_2\end{bmatrix} &= \begin{bmatrix}0 & \mathbf{D}_1\mathbf{A}\\ \mathbf{D}_2\mathbf{A}^\intercal & 0\end{bmatrix}\begin{bmatrix}\mathbf{D}_1 & 0\\ 0 & \mathbf{D}_2\end{bmatrix} \\ &= \begin{bmatrix}0 & \mathbf{D}_1\mathbf{A}\mathbf{D}_2\\ \mathbf{D}_2\mathbf{A}^\intercal\mathbf{D}_1 & 0\end{bmatrix} \\ &= \overline{\mathbf{D}_1\mathbf{A}\mathbf{D}_2}\end{aligned}$$

□

We start with a generalized version of a result from [PMT16].

LEMMA 7.4. *Let* $\mathbf{K} = \overline{\mathbf{K}}_{k,a,b}$. *For any symmetric matrix valued function* $\mathbf{R}$ *on the variables* Z *of the same dimensions as* $\mathbf{K}$, *such that* $\mathbb{E}\,\|\mathbf{K}(Z,Z')\mathbf{R}(Z)\| < \infty$, *we have*

$$\mathbb{E}[\overline{\mathbf{F}}_{k,a,b}(Z)\mathbf{R}(Z)] = \frac{1}{2}\mathbb{E}[\mathbf{D}(Z)\mathbf{K}(Z,Z')\mathbf{D}(Z)(\mathbf{R}(Z) - \mathbf{R}(Z'))]$$

Proof. By Lemma 7.3, we have

$$\begin{aligned}\mathbb{E}[\overline{\mathbf{F}}_{k,a,b}(Z)\mathbf{R}(Z)] &= \mathbb{E}[\mathbf{D}(Z)\overline{\mathbf{G}}_{k,a,b}(Z)\mathbf{D}(Z)\mathbf{R}(Z)] \\ &= \mathbb{E}[\mathbf{D}(Z)\mathbb{E}[\mathbf{K}(Z,Z')|Z]\mathbf{D}(Z)\mathbf{R}(Z)] \\ &= \mathbb{E}[\mathbf{D}(Z)\mathbf{K}(Z,Z')\mathbf{D}(Z)\mathbf{R}(Z)]\end{aligned}$$

where the first equality follow from condition 2 of Lemma 7.3 and the second follows from the pull-through property of expectations. Continuing,

$$\begin{aligned}\mathbb{E}[\overline{\mathbf{F}}_{k,a,b}(Z)\mathbf{R}(Z)] &= \mathbb{E}[\mathbf{D}(Z)\mathbf{K}(Z,Z')\mathbf{D}(Z)\mathbf{R}(Z)] \\ &= \mathbb{E}[\mathbf{D}(Z')\mathbf{K}(Z',Z)\mathbf{D}(Z')\mathbf{R}(Z')] \\ &= -\mathbb{E}[\mathbf{D}(Z')\mathbf{K}(Z,Z')\mathbf{D}(Z')\mathbf{R}(Z')] \\ &= -\mathbb{E}[\mathbf{D}(Z)\mathbf{K}(Z,Z')\mathbf{D}(Z')\mathbf{R}(Z')] \\ &= -\mathbb{E}[\mathbf{D}(Z)\mathbf{K}(Z,Z')\mathbf{D}(Z)\mathbf{R}(Z')]\end{aligned}$$

Here, the second equality follows from the fact that (Z,Z') has the same distribution as (Z',Z), so we can exchange them. The third, fourth and fifth equalities follow from conditions $1,3,3$ of Lemma 7.3 respectively. Adding the two displays, we get the result. ∎

DEFINITION 7.3. (MATRICES $\mathbf{U}_{k,a,b}, \mathbf{V}_{k,a,b}$) *We define the following matrices*

$$\mathbf{U}_{k,a,b} = \mathbb{E}[(\overline{\mathbf{F}}_{k,a,b}(Z) - \overline{\mathbf{F}}_{k,a,b}(Z'))^2|Z]$$

$$\mathbf{V}_{k,a,b} = \mathbb{E}[(\mathbf{D}(Z)\overline{\mathbf{K}}_{k,a,b}(Z,Z')\mathbf{D}(Z))^2|Z]$$

The definition of $\mathbf{U}_{k,a,b}$ is essentially unchanged from [PMT16], where it is called the *conditional variance*. The definition of $\mathbf{V}_{k,a,b}$ is slightly different in our setting. This lets us exploit the specific product structure exhibited by $\overline{\mathbf{F}}_{k,a,b}$ and the special properties of the inner kernel from Lemma 7.3. We will now prove a lemma which is similar to a lemma shown in [PMT16].

LEMMA 7.5. *For any* $s > 0$ *and for any integer* $t \geq 1$,

$$\mathbb{E}\,\|\overline{\mathbf{F}}_{k,a,b}\|_{2t}^{2t} \leq \left(\frac{2t-1}{4}\right)^t \mathbb{E}\left\|s\mathbf{U}_{k,a,b} + s^{-1}\mathbf{V}_{k,a,b}\right\|_t^t$$

To prove this, we will use the following inequality.

LEMMA 7.6. (POLYNOMIAL MEAN VALUE TRACE INEQUALITY, [PMT16]) *For all matrices* $\mathbf{A},\mathbf{B},\mathbf{C} \in \mathbb{H}^d$, *all integers* $q \geq 1$ *and all* $s > 0$,

$$\mathrm{tr}[\mathbf{C}(\mathbf{A}^q - \mathbf{B}^q)]| \leq \frac{q}{4}\,\mathrm{tr}[(s(\mathbf{A}-\mathbf{B})^2 + s^{-1}\mathbf{C}^2)(\mathbf{A}^{q-1} + \mathbf{B}^{q-1})]$$

Proof. [Proof of Lemma 7.5] We start by invoking Lemma 7.4 by setting $\mathbf{R}(Z) = \overline{\mathbf{F}}_{k,a,b}^{2t-1}(Z)$ to get

$$\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t} = \mathbb{E}\,\mathrm{tr}[\overline{\mathbf{F}}_{k,a,b}\cdot\overline{\mathbf{F}}_{k,a,b}^{2t-1}] = \frac{1}{2}\mathbb{E}[\mathbf{D}(Z)\overline{\mathbf{K}}_{k,a,b}(Z,Z')\mathbf{D}(Z)(\overline{\mathbf{F}}_{k,a,b}^{2t-1}(Z)-\overline{\mathbf{F}}_{k,a,b}^{2t-1}(Z'))]$$

Applying Lemma 7.6,

$$\begin{aligned}
&\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t}\\
&\le (\frac{2t-1}{8})\mathbb{E}\,\mathrm{tr}[(s(\overline{\mathbf{F}}_{k,a,b}(Z)-\overline{\mathbf{F}}_{k,a,b}(Z'))^2+s^{-1}(\mathbf{D}(Z)\overline{\mathbf{K}}_{k,a,b}(Z,Z')\mathbf{D}(Z))^2)(\overline{\mathbf{F}}_{k,a,b}^{2t-2}(Z)+\overline{\mathbf{F}}_{k,a,b}^{2t-2}(Z'))]\\
&= (\frac{2t-1}{4})\mathbb{E}\,\mathrm{tr}[(s(\overline{\mathbf{F}}_{k,a,b}(Z)-\overline{\mathbf{F}}_{k,a,b}(Z'))^2+s^{-1}(\mathbf{D}(Z)\overline{\mathbf{K}}_{k,a,b}(Z,Z')\mathbf{D}(Z))^2)\overline{\mathbf{F}}_{k,a,b}^{2t-2}(Z)]
\end{aligned}$$

where the last line used the fact that (Z,Z') has the same distribution as (Z',Z) and applied condition 3 of Lemma 7.3. Using the definitions of $\mathbf{U}_{k,a,b}$ and $\mathbf{V}_{k,a,b}$, we get

$$\begin{aligned}
\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t} &\le \frac{2t-1}{4}\mathbb{E}\,\mathrm{tr}[(s\mathbf{U}_{k,a,b}+s^{-1}\mathbf{V}_{k,a,b})\overline{\mathbf{F}}_{k,a,b}^{2t-2}]\\
&\le \frac{2t-1}{4}\left(\mathbb{E}\left\|s\mathbf{U}_{k,a,b}+s^{-1}\mathbf{V}_{k,a,b}\right\|_t^t\right)^{1/t}(\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t})^{(t-1)/t}
\end{aligned}$$

where we used Hölder's inequality for the trace and Hölder's inequality for the expectation. Rearranging gives the result. ∎

7.2 Proof of Lemma 6.1 Lemma 7.5 suggests that in order to bound $\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t}$, it suffices to bound $\mathbb{E}\left\|\mathbf{U}_{k,a,b}\right\|_t^t$ and $\mathbb{E}\left\|\mathbf{V}_{k,a,b}\right\|_t^t$. Indeed, this will be our strategy. To bound $\mathbb{E}\left\|\mathbf{U}_{k,a,b}\right\|_t^t$, we will bound it via the matrices that we define below.

DEFINITION 7.4. (MATRICES $\mathbf{\Delta}_1^{k,a,b}, \mathbf{\Delta}_2^{k,a,b}, \mathbf{\Delta}_3^{k,a,b}$) *Define the matrices*

$$\mathbf{\Delta}_1^{k,a,b} = \mathbb{E}[((\mathbf{D}(Z)-\mathbf{D}(Z'))\overline{\mathbf{G}}_{k,a,b}(Z)\mathbf{D}(Z))^2|Z]$$

$$\mathbf{\Delta}_2^{k,a,b} = \mathbb{E}[(\mathbf{D}(Z)(\overline{\mathbf{G}}_{k,a,b}(Z)-\overline{\mathbf{G}}_{k,a,b}(Z'))\mathbf{D}(Z))^2|Z]$$

$$\mathbf{\Delta}_3^{k,a,b} = \mathbb{E}[(\mathbf{D}(Z)\overline{\mathbf{G}}_{k,a,b}(Z)(\mathbf{D}(Z)-\mathbf{D}(Z')))^2|Z]$$

LEMMA 7.7. $\mathbf{U}_{k,a,b} \preceq 3(\mathbf{\Delta}_1^{k,a,b}+\mathbf{\Delta}_2^{k,a,b}+\mathbf{\Delta}_3^{k,a,b})$.

To prove this lemma, we will use the following lemma.

LEMMA 7.8. *We have the relations*

$$(\mathbf{D}(Z)-\mathbf{D}(Z'))(\overline{\mathbf{G}}_{k,a,b}(Z)\mathbf{D}(Z)-\overline{\mathbf{G}}_{k,a,b}(Z')\mathbf{D}(Z'))=0$$

$$(\overline{\mathbf{G}}_{k,a,b}(Z)-\overline{\mathbf{G}}_{k,a,b}(Z'))(\mathbf{D}(Z)-\mathbf{D}(Z'))=0$$

Proof. [Proof sketch] The proof is similar to the proof of third equality in Lemma 7.3. When Z' is set to $Z^{(i)}$ for some $i \le n$, when a diagonal entry of $\mathbf{D}(Z)-\mathbf{D}(Z')$ is nonzero, then the corresponding row of $\overline{\mathbf{G}}_{k,a,b}(Z)\mathbf{D}(Z)-\overline{\mathbf{G}}_{k,a,b}(Z')\mathbf{D}(Z')$ will be 0. The second equality is analogous. ∎

Proof. [Proof of Lemma 7.7] We have

$$\begin{aligned}
&(\overline{\mathbf{F}}_{k,a,b}(Z) - \overline{\mathbf{F}}_{k,a,b}(Z'))^2 \\
&= (\mathbf{D}(Z)\overline{\mathbf{G}}_{k,a,b}(Z)\mathbf{D}(Z) - \mathbf{D}(Z')\overline{\mathbf{G}}_{k,a,b}(Z')\mathbf{D}(Z'))^2 \\
&= \left(\mathbf{D}(Z)\overline{\mathbf{G}}_{k,a,b}(Z)(\mathbf{D}(Z) - \mathbf{D}(Z')) + \mathbf{D}(Z)(\overline{\mathbf{G}}_{k,a,b}(Z) - \overline{\mathbf{G}}_{k,a,b}(Z'))\mathbf{D}(Z') + (\mathbf{D}(Z) - \mathbf{D}(Z'))\overline{\mathbf{G}}_{k,a,b}(Z')\mathbf{D}(Z')\right)^2 \\
&= \left(\mathbf{D}(Z)\overline{\mathbf{G}}_{k,a,b}(Z)(\mathbf{D}(Z) - \mathbf{D}(Z')) + \mathbf{D}(Z)(\overline{\mathbf{G}}_{k,a,b}(Z) - \overline{\mathbf{G}}_{k,a,b}(Z'))\mathbf{D}(Z) + (\mathbf{D}(Z) - \mathbf{D}(Z'))\overline{\mathbf{G}}_{k,a,b}(Z)\mathbf{D}(Z)\right)^2
\end{aligned}$$

where the last equality follows from Lemma 7.8. Taking expectations conditioned on Z and applying Fact 2.1, we immediately get $\mathbf{U}_{k,a,b} \preceq 3(\mathbf{\Delta}_1^{k,a,b} + \mathbf{\Delta}_2^{k,a,b} + \mathbf{\Delta}_3^{k,a,b})$. ∎

In subsequent sections, we will prove the following technical bounds on the matrices we have considered so far.

LEMMA 7.9. *For all integers $t \geq 1$, $\mathbb{E}\left\|\mathbf{\Delta}_2^{k,a,b}\right\|_t^t \leq \frac{(2d_p)^t}{n^t}(\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b+1}\right\|_{2t}^{2t} + \mathbb{E}\left\|\overline{\mathbf{F}}_{k,a+1,b}\right\|_{2t}^{2t})$.*

LEMMA 7.10. *For all integers $t \geq 1$, $\mathbb{E}\left\|\mathbf{V}_{k,a,b}\right\|_t^t \leq (2d_p)^t n^t(\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b+1}\right\|_{2t}^{2t} + \mathbb{E}\left\|\overline{\mathbf{F}}_{k,a+1,b}\right\|_{2t}^{2t})$.*

LEMMA 7.11. *For all integers $t \geq 1$, $\mathbb{E}\left\|\mathbf{\Delta}_1^{k,a,b}\right\|_t^t \leq \frac{(8dd_p)^t}{n^t}\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t}$.*

LEMMA 7.12. *For all integers $t \geq 1$, $\mathbb{E}\left\|\mathbf{\Delta}_3^{k,a,b}\right\|_t^t \leq \frac{(4d_p)^t}{n^t}\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t}$.*

Assuming the above lemmas, we can complete the proof of Lemma 6.1, which we restate for convenience.

LEMMA 6.1. *For all integers $t \geq 1$, integers $k \geq 1, a, b \geq 0$ such that $a + b < k$,*

$$\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t} \leq (Ct^2dd_p^2)^t(\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b+1}\right\|_{2t}^{2t} + \mathbb{E}\left\|\overline{\mathbf{F}}_{k,a+1,b}\right\|_{2t}^{2t})$$

for an absolute constant $C > 0$.

Proof. [Proof of Lemma 6.1] Using Lemma 7.5, Lemma 7.7, we get that for any $s > 0$,

$$\begin{aligned}
\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t} &\leq (\frac{2t-1}{4})^t\mathbb{E}\left\|s\mathbf{U}_{k,a,b} + s^{-1}\mathbf{V}_{k,a,b}\right\|_t^t \\
&\leq t^t(s^t\mathbb{E}\left\|\mathbf{U}_{k,a,b}\right\|_t^t + s^{-t}\mathbb{E}\left\|\mathbf{V}_{k,a,b}\right\|_t^t) \\
&\leq (9st)^t(\mathbb{E}\left\|\mathbf{\Delta}_1^{k,a,b}\right\|_t^t + \mathbb{E}\left\|\mathbf{\Delta}_2^{k,a,b}\right\|_t^t + \mathbb{E}\left\|\mathbf{\Delta}_3^{k,a,b}\right\|_t^t) + t^ts^{-t}\mathbb{E}\left\|\mathbf{V}_{k,a,b}\right\|_t^t
\end{aligned}$$

Let $\rho = s/n$. Since the inequality is true for any choice of $s > 0$, it is true for any choice of $\rho > 0$. Now, using Lemma 7.11, Lemma 7.12,

$$\begin{aligned}
(9st)^t(\mathbb{E}\left\|\mathbf{\Delta}_1^{k,a,b}\right\|_t^t + \mathbb{E}\left\|\mathbf{\Delta}_3^{k,a,b}\right\|_t^t) &\leq (9st)^t\left(\frac{(8dd_p)^t}{n^t} + \frac{(4d_p)^t}{n^t}\right)\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t} \\
&= \rho^t(C_1tdd_p)^t\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t}
\end{aligned}$$

for an absolute constant $C_1 > 0$. Using Lemma 7.9, Lemma 7.10,

$$\begin{aligned}
(9st)^t\mathbb{E}\left\|\mathbf{\Delta}_2^{k,a,b}\right\|_t^t + t^ts^{-t}\mathbb{E}\left\|\mathbf{V}_{k,a,b}\right\|_t^t &\leq \left((9st)^t\frac{(2d_p)^t}{n^t} + t^ts^{-t}(2d_p)^tn^t\right)(\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b+1}\right\|_{2t}^{2t} + \mathbb{E}\left\|\overline{\mathbf{F}}_{k,a+1,b}\right\|_{2t}^{2t}) \\
&\leq (\rho^tC_2^t + \rho^{-t}C_3^t)(td_p)^t(\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b+1}\right\|_{2t}^{2t} + \mathbb{E}\left\|\overline{\mathbf{F}}_{k,a+1,b}\right\|_{2t}^{2t})
\end{aligned}$$

for absolute constants $C_2, C_3 > 0$. Therefore,

$$\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t} \leq \rho^t (C_1 t d d_p)^t \mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t} + (\rho^t C_2^t + \rho^{-t} C_3^t)(t d_p)^t (\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b+1}\right\|_{2t}^{2t} + \mathbb{E}\left\|\overline{\mathbf{F}}_{k,a+1,b}\right\|_{2t}^{2t})$$

We choose $\rho > 0$ so that $\rho^t (C_1 t d d_p)^t = \frac{1}{2}$ to get

$$\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t} \leq \frac{1}{2}\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t} + \frac{1}{2}(C t^2 d d_p^2)^t (\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b+1}\right\|_{2t}^{2t} + \mathbb{E}\left\|\overline{\mathbf{F}}_{k,a+1,b}\right\|_{2t}^{2t})$$

for an absolute constant $C > 0$. Rearranging yields the result. □

7.3 Bounding $\mathbf{\Delta}_2^{k,a,b}$ and $\mathbf{V}_{k,a,b}$ The next lemma relates $\mathbf{V}_{k,a,b}$ to $\mathbf{\Delta}_2^{k,a,b}$ upto a factor of n^2 which will be enough for us. We can then focus on bounding $\mathbf{\Delta}_2^{k,a,b}$.

LEMMA 7.13. $\mathbf{V}_{k,a,b} \preceq n^2 \mathbf{\Delta}_2^{k,a,b}$

Proof. Using Lemma 7.2,

$$\begin{aligned}
\mathbf{V}_{k,a,b} &= \mathbb{E}[(\mathbf{D}(Z)\overline{\mathbf{K}}_{k,a,b}(Z,Z')\mathbf{D}(Z))^2|Z] \\
&= \mathbb{E}[(\mathbf{D}(Z)\left(\frac{n}{k-a-b}(\overline{\mathbf{G}}_{k,a,b}(Z) - \overline{\mathbf{G}}_{k,a,b}(Z'))\right)\mathbf{D}(Z))^2|Z] \\
&\preceq n^2 \mathbb{E}[(\mathbf{D}(Z)(\overline{\mathbf{G}}_{k,a,b}(Z) - \overline{\mathbf{G}}_{k,a,b}(Z'))\mathbf{D}(Z))^2|Z] \\
&= n^2 \mathbf{\Delta}_2^{k,a,b}
\end{aligned}$$

□

For $1 \leq i \leq n$ and $1 \leq l \leq d$, let $\mathbf{e}_{i,l} \in \mathbb{N}^n$ denote the vector α with $\alpha_i = l$ and $\alpha_j = 0$ for $j \neq i$. We note the following simple proposition.

PROPOSITION 7.1. *For any polynomial f such that the degree of Z_i is at most d, $f(Z) - f(Z^{(i)}) = \sum_{1 \leq l \leq d} (Z_i^l - \widetilde{Z}_i^{\,l}) \nabla_{\mathbf{e}_{i,l}}(f)$*

We now restate and prove Lemma 7.9.

LEMMA 7.9. *For all integers $t \geq 1$, $\mathbb{E}\left\|\mathbf{\Delta}_2^{k,a,b}\right\|_t^t \leq \frac{(2d_p)^t}{n^t}(\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b+1}\right\|_{2t}^{2t} + \mathbb{E}\left\|\overline{\mathbf{F}}_{k,a+1,b}\right\|_{2t}^{2t})$.*

Proof. Consider

$$\begin{aligned}
\mathbf{\Delta}_2^{k,a,b} &= \mathbb{E}[(\mathbf{D}(Z)(\overline{\mathbf{G}}_{k,a,b}(Z) - \overline{\mathbf{G}}_{k,a,b}(Z'))\mathbf{D}(Z))^2|Z] \\
&= \mathbb{E}\left[\begin{bmatrix} \mathbf{M}\mathbf{M}^\intercal & 0 \\ 0 & \mathbf{M}^\intercal\mathbf{M} \end{bmatrix}|Z\right] \\
&= \begin{bmatrix} \mathbb{E}[\mathbf{M}\mathbf{M}^\intercal|Z] & 0 \\ 0 & \mathbb{E}[\mathbf{M}^\intercal\mathbf{M}|Z] \end{bmatrix}
\end{aligned}$$

where $\mathbf{M} = \mathbf{D}_1(Z)(\mathbf{G}_{k,a,b}(Z) - \mathbf{G}_{k,a,b}(Z'))\mathbf{D}_2(Z)$. Using Proposition 7.1,

$$\begin{aligned}
\mathbb{E}[\mathbf{M}\mathbf{M}^T|Z] &= \mathbb{E}[\mathbf{D}_1(Z)(\mathbf{G}_{k,a,b}(Z) - \mathbf{G}_{k,a,b}(Z'))\mathbf{D}_2(Z) \cdot \mathbf{D}_2(Z)(\mathbf{G}_{k,a,b}(Z) - \mathbf{G}_{k,a,b}(Z'))^\intercal \mathbf{D}_1(Z)|Z] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbf{D}_1(Z)(\mathbf{G}_{k,a,b}(Z) - \mathbf{G}_{k,a,b}(Z^{(i)}))\mathbf{D}_2(Z) \cdot \mathbf{D}_2(Z)(\mathbf{G}_{k,a,b}(Z) - \mathbf{G}_{k,a,b}(Z^{(i)}))^\intercal \mathbf{D}_1(Z)|Z] \\
&= \frac{1}{n}\sum_{i=1}^n \sum_{l=1}^d \mathbb{E}[(Z_i^l - \widetilde{Z}_i^{\,l})^2|Z] \cdot \mathbf{D}_1(Z)(\nabla_{\mathbf{e}_{i,l}}\mathbf{G}_{k,a,b})(Z)\mathbf{D}_2(Z) \cdot \mathbf{D}_2(Z)(\nabla_{\mathbf{e}_{i,l}}\mathbf{G}_{k,a,b})(Z)^\intercal \mathbf{D}_1(Z)
\end{aligned}$$

Define $\mathbf{N}_{i,l}(Z) := \mathbf{D}_1(Z)(\nabla_{\mathbf{e}_{i,l}}\mathbf{G}_{k,a,b})(Z)\mathbf{D}_2(Z)$. Then,

$$\mathbb{E}[\mathbf{M}\mathbf{M}^T|Z] = \frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{d}\mathbb{E}[(Z_i^l - \widetilde{Z}_i^{\,l})^2|Z]\cdot\mathbf{N}_{i,l}(Z)\mathbf{N}_{i,l}(Z)^\intercal \preceq \frac{2}{n}\sum_{i=1}^{n}\sum_{l=1}^{d}(Z_i^{2l}+\mathbb{E}[Z_i^{2l}])\cdot\mathbf{N}_{i,l}(Z)\mathbf{N}_{i,l}(Z)^\intercal$$

Similarly, $\mathbb{E}[\mathbf{M}^\intercal\mathbf{M}|Z] \preceq \frac{2}{n}\sum_{i=1}^{n}\sum_{l=1}^{d}(Z_i^{2l}+\mathbb{E}[Z_i^{2l}])\cdot\mathbf{N}_{i,l}(Z)^\intercal\mathbf{N}_{i,l}(Z)$

CLAIM 7.1. *We have the relations*

$$\sum_{i=1}^{n}\sum_{l=1}^{d}(Z_i^{2l}+\mathbb{E}[Z_i^{2l}])\cdot\mathbf{N}_{i,l}(Z)\mathbf{N}_{i,l}(Z)^\intercal = (b+1)\mathbf{F}_{k,a,b+1}\mathbf{F}_{k,a,b+1}^\intercal$$

$$\sum_{i=1}^{n}\sum_{l=1}^{d}(Z_i^{2l}+\mathbb{E}[Z_i^{2l}])\cdot\mathbf{N}_{i,l}(Z)^\intercal\mathbf{N}_{i,l}(Z) = (a+1)\mathbf{F}_{k,a+1,b}^\intercal\mathbf{F}_{k,a+1,b}$$

Using this claim, we have

$$\mathbb{E}[\mathbf{M}\mathbf{M}^T|Z] \preceq \frac{2(b+1)}{n}\mathbf{F}_{k,a,b+1}\mathbf{F}_{k,a,b+1}^\intercal \preceq \frac{2d_p}{n}\mathbf{F}_{k,a,b+1}\mathbf{F}_{k,a,b+1}^\intercal$$

$$\mathbb{E}[\mathbf{M}^\intercal\mathbf{M}|Z] \preceq \frac{2(a+1)}{n}\mathbf{F}_{k,a+1,b}^\intercal\mathbf{F}_{k,a+1,b} \preceq \frac{2d_p}{n}\mathbf{F}_{k,a+1,b}^\intercal\mathbf{F}_{k,a+1,b}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left\|\Delta_2^{k,a,b}\right\|_t^t = \mathbb{E}\left\|\mathbb{E}[\mathbf{M}\mathbf{M}^\intercal|Z]\right\|_t^t + \mathbb{E}\left\|\mathbb{E}[\mathbf{M}^\intercal\mathbf{M}|Z]\right\|_t^t &\le \frac{(2d_p)^t}{n^t}(\mathbb{E}\left\|\mathbf{F}_{k,a,b+1}\right\|_{2t}^{2t} + \mathbb{E}\left\|\mathbf{F}_{k,a+1,b}\right\|_{2t}^{2t}) \\ &\le \frac{(2d_p)^t}{n^t}(\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b+1}\right\|_{2t}^{2t} + \mathbb{E}\left\|\overline{\mathbf{F}}_{k,a+1,b}\right\|_{2t}^{2t})\end{aligned}$$

□

It remains to prove the claim.

Proof. [Proof of Claim 7.1] We will prove the first relation, the second is analogous. For a fixed $i \le n, l \le d$, consider any nonzero entry $[(I_1,\alpha_1,\gamma_1),(I_2,\alpha_2,\gamma_2)]$ of $\sum_{i=1}^{n}\sum_{l=1}^{d}(Z_i^{2l}+\mathbb{E}[Z_i^{2l}])\mathbf{N}_{i,l}(Z)\mathbf{N}_{i,l}(Z)^\intercal$, where $I_1, I_2 \in \mathcal{I}, (\alpha_1,\gamma_1),(\alpha_2,\gamma_2) \in \mathcal{K}$. We must have $|\alpha_1|_0 = |\alpha_2|_0 = a$, in which case the entry is equal to

$$\begin{aligned}\sum_{\substack{(J,\alpha_3,\gamma_3)\in\mathcal{J}\times\mathcal{K}\\ |\alpha_3|=b\\ \alpha_1\alpha_3=\alpha_2\alpha_3=0}} &(Z_i^{2l}+\mathbb{E}[Z_i^{2l}])\cdot(\sqrt{\mathbb{E}[Z^{2\alpha_1\cdot(1-\gamma_1)+2\alpha_3\cdot(1-\gamma_3)}]}Z^{\alpha_1\cdot\gamma_1+\alpha_3\cdot\gamma_3}\nabla_{\mathbf{e}_{i,l}}\nabla_{\alpha_1+\alpha_3}\mathbf{X}_k[I_1,J]) \\ &\cdot(\sqrt{\mathbb{E}[Z^{2\alpha_2\cdot(1-\gamma_2)+2\alpha_3\cdot(1-\gamma_3)}]}Z^{\alpha_2\cdot\gamma_2+\alpha_3\cdot\gamma_3}\nabla_{\mathbf{e}_{i,l}}\nabla_{\alpha_2+\alpha_3}\mathbf{X}_k[I_2,J])\end{aligned}$$

Note that the term inside the summation is nonzero only when $\mathbf{e}_{i,l}\cdot(\alpha_1+\alpha_3) = \mathbf{e}_{i,l}\cdot(\alpha_2+\alpha_3) = 0$. Hence, this sum can be written as

$$\begin{aligned}\sum_{\substack{(J,\alpha_3,\gamma_3)\in\mathcal{J}\times\mathcal{K}\\ |\alpha_3|=b+1\\ \mathbf{e}_{i,l}\le\alpha_3,\alpha_1\alpha_3=\alpha_2\alpha_3=0}} &(\sqrt{\mathbb{E}[Z^{2\alpha_1\cdot(1-\gamma_1)+2\alpha_3\cdot(1-\gamma_3)}]}Z^{\alpha_1\cdot\gamma_1+\alpha_3\cdot\gamma_3}\nabla_{\alpha_1+\alpha_3}\mathbf{X}_k[I_1,J]) \\ &\cdot(\sqrt{\mathbb{E}[Z^{2\alpha_2\cdot(1-\gamma_2)+2\alpha_3\cdot(1-\gamma_3)}]}Z^{\alpha_2\cdot\gamma_2+\alpha_3\cdot\gamma_3}\nabla_{\alpha_2+\alpha_3}\mathbf{X}_k[I_2,J])\end{aligned}$$

When we add this entry over all $i \le n, l \le d$, this simplifies to

$$(b+1)\cdot \sum_{\substack{(J,\alpha_3,\gamma_3)\in\mathcal{J}\times\mathcal{K}\\ |\alpha_3|=b+1\\ \alpha_1\alpha_3=\alpha_2\alpha_3=0}} (\sqrt{\mathbb{E}[Z^{2\alpha_1\cdot(1-\gamma_1)+2\alpha_3\cdot(1-\gamma_3)}]}Z^{\alpha_1\cdot\gamma_1+\alpha_3\cdot\gamma_3}\nabla_{\alpha_1+\alpha_3}\mathbf{X}_k[I_1,J]) \cdot (\sqrt{\mathbb{E}[Z^{2\alpha_2\cdot(1-\gamma_2)+2\alpha_3\cdot(1-\gamma_3)}]}Z^{\alpha_2\cdot\gamma_2+\alpha_3\cdot\gamma_3}\nabla_{\alpha_2+\alpha_3}\mathbf{X}_k[I_2,J])$$

The factor of $(b+1)$ came because the index i could have been chosen from among all the active indices in α_3. But this is precisely the $[(I_1,\alpha_1,\gamma_1),(I_2,\alpha_2,\gamma_2)]$ entry of $(b+1)\mathbf{F}_{k,a,b+1}\mathbf{F}_{k,a,b+1}^\intercal$, proving the claim. ∎

We restate and prove Lemma 7.10.

LEMMA 7.10. *For all integers $t \ge 1$, $\mathbb{E}\left\|\mathbf{V}_{k,a,b}\right\|_t^t \le (2d_p)^t n^t(\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b+1}\right\|_{2t}^{2t} + \mathbb{E}\left\|\overline{\mathbf{F}}_{k,a+1,b}\right\|_{2t}^{2t})$.*

Proof. Using Lemma 7.13 and Lemma 7.9, we get

$$\mathbb{E}\left\|\mathbf{V}_{k,a,b}\right\|_t^t \le n^{2t}\mathbb{E}\left\|\mathbf{\Delta}_2^{k,a,b}\right\|_t^t \le (2d_p)^t n^t(\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b+1}\right\|_{2t}^{2t} + \mathbb{E}\left\|\overline{\mathbf{F}}_{k,a+1,b}\right\|_{2t}^{2t})$$

∎

7.4 Bounding $\mathbf{\Delta}_1^{k,a,b}$ and $\mathbf{\Delta}_3^{k,a,b}$ Define $\sqcup$ to be the disjoint union of sets. For $1 \le i \le n$ and $1 \le l \le d$, define the diagonal matrices $\mathbf{\Pi}_{i,l}, \mathbf{\Pi}'_{i,l}, \mathbf{\Pi}_i, \mathbf{\Pi}'_i \in \mathbb{R}^{(\mathcal{I}\times\mathcal{K})\sqcup(\mathcal{J}\times\mathcal{K})} \times \mathbb{R}^{(\mathcal{I}\times\mathcal{K})\sqcup(\mathcal{J}\times\mathcal{K})}$ (the same dimensions as $\mathbf{D}$) as

$$\mathbf{\Pi}_{i,l}[(I,\alpha,\beta),(I,\alpha,\beta)] = \begin{cases} 1 & \text{if } (\alpha\cdot\gamma)_i \ne 0 \text{ and } \alpha_i = l \\ 0 & \text{o.w.} \end{cases} \qquad \mathbf{\Pi}_i[(I,\alpha,\beta),(I,\alpha,\beta)] = \begin{cases} 1 & \text{if } (\alpha\cdot\gamma)_i \ne 0 \\ 0 & \text{o.w.} \end{cases}$$

$$\mathbf{\Pi}'_{i,l}[(I,\alpha,\beta),(I,\alpha,\beta)] = \begin{cases} 1 & \text{if } \alpha_i \ne 0 \text{ and } \alpha_i = l \\ 0 & \text{o.w.} \end{cases} \qquad \mathbf{\Pi}'_i[(I,\alpha,\beta),(I,\alpha,\beta)] = \begin{cases} 1 & \text{if } \alpha_i \ne 0 \\ 0 & \text{o.w.} \end{cases}$$

for all $I \in \mathcal{I}\sqcup\mathcal{J}$. Note that for all $i \le n$, $\mathbf{\Pi}_i = \sum_{l=1}^d \mathbf{\Pi}_{i,l}$.

Also, for all $1 \le i \le n$, we define the permutation matrices $\mathbf{\Sigma}_i \in \mathbb{R}^{(\mathcal{I}\times\mathcal{K})\sqcup(\mathcal{J}\times\mathcal{K})} \times \mathbb{R}^{(\mathcal{I}\times\mathcal{K})\sqcup(\mathcal{J}\times\mathcal{K})}$ as follows. Consider the permutation σ_1 on $\mathcal{I}\times\mathcal{K}$ that transposes (I,α,γ) and $(I,\alpha,\gamma+\mathbf{e}_i)$ for all $(I,\alpha,\gamma) \in \mathcal{I}\times\mathcal{K}$ such that $\alpha_i \ne 0$. Here, $\mathbf{e}_i \in \{0,1\}^n$ has exactly one nonzero entry, which is in the ith position, and $\gamma+\mathbf{e}_i$ is the usual addition over $\mathbb{F}_2$. σ_1 leaves other positions fixed. Let $\mathbf{\Sigma}_i^{(1)}$ be the permutation matrix for σ. Similarly, let $\mathbf{\Sigma}_i^{(2)}$ be the permutation matrix of the permutation σ_2 on $\mathcal{J}\times\mathcal{K}$ that transposes (J,α,γ) and $(J,\alpha,\gamma+\mathbf{e}_i)$ for all $(J,\alpha,\gamma) \in \mathcal{J}\times\mathcal{K}$ such that $\alpha_i \ne 0$, and leaves all other positions fixed. Then, we define $\mathbf{\Sigma}_i = \begin{bmatrix} \mathbf{\Sigma}_i^{(1)} & 0 \\ 0 & \mathbf{\Sigma}_i^{(2)} \end{bmatrix}$. The following fact is easy to verify.

FACT 7.2. *$\mathbf{\Pi}'_{i,l}\mathbf{\Sigma}_i = \mathbf{\Sigma}_i\mathbf{\Pi}'_{i,l}$ and $\mathbf{\Pi}'_i\mathbf{\Sigma}_i = \mathbf{\Sigma}_i\mathbf{\Pi}'_i$.*

We are now ready to prove Lemma 7.11 which we restate for convenience.

LEMMA 7.11. *For all integers $t \ge 1$, $\mathbb{E}\left\|\mathbf{\Delta}_1^{k,a,b}\right\|_t^t \le \frac{(8dd_p)^t}{n^t}\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t}$.*

Proof. Firstly,

$$\begin{aligned}\mathbf{\Delta}_1^{k,a,b} &= \mathbb{E}[((\mathbf{D}(Z)-\mathbf{D}(Z'))\overline{\mathbf{G}}_{k,a,b}(Z)\mathbf{D}(Z))^2|Z] \\ &= \mathbb{E}[(\mathbf{D}(Z)-\mathbf{D}(Z'))\overline{\mathbf{G}}_{k,a,b}(Z)\mathbf{D}(Z)\cdot\mathbf{D}(Z)\overline{\mathbf{G}}_{k,a,b}(Z)(\mathbf{D}(Z)-\mathbf{D}(Z'))|Z] \\ &= \mathbb{E}[(\mathbf{D}(Z)-\mathbf{D}(Z'))\mathbf{M}(Z)(\mathbf{D}(Z)-\mathbf{D}(Z'))|Z]\end{aligned}$$

where we define $\mathbf{M}(Z) = \overline{\mathbf{G}}_{k,a,b}(Z)\mathbf{D}(Z) \cdot \mathbf{D}(Z)\overline{\mathbf{G}}_{k,a,b}(Z)$. Recall that $Z' = Z^{(i)}$ for some i randomly chosen from $[n]$ uniformly. Observing that $\mathbf{D}(Z) - \mathbf{D}(Z^{(i)}) = \mathbf{\Pi}_i(\mathbf{D}(Z) - \mathbf{D}(Z^{(i)}))$ for all i, we get

$$\begin{aligned}
\mathbf{\Delta}_1^{k,a,b} &= \mathbb{E}[\mathbb{E}_{i\in[n]}[(\mathbf{D}(Z) - \mathbf{D}(Z^{(i)}))\mathbf{M}(Z)(\mathbf{D}(Z) - \mathbf{D}(Z^{(i)}))]|Z] \\
&= \mathbb{E}[\mathbb{E}_{i\in[n]}[\mathbf{\Pi}_i(\mathbf{D}(Z) - \mathbf{D}(Z^{(i)}))\mathbf{M}(Z)(\mathbf{D}(Z) - \mathbf{D}(Z^{(i)}))\mathbf{\Pi}_i]|Z] \\
&\preceq 2\Big(\mathbb{E}[\mathbb{E}_{i\in[n]}[\mathbf{\Pi}_i\mathbf{D}(Z)\mathbf{M}(Z)\mathbf{D}(Z)\mathbf{\Pi}_i]|Z] + \mathbb{E}[\mathbb{E}_{i\in[n]}[\mathbf{\Pi}_i\mathbf{D}(Z^{(i)})\mathbf{M}(Z)\mathbf{D}(Z^{(i)})\mathbf{\Pi}_i]|Z]\Big) \\
&\preceq 2\Big(\mathbb{E}_{i\in[n]}[\mathbf{\Pi}_i\overline{\mathbf{F}}_{k,a,b}^2\mathbf{\Pi}_i] + \mathbb{E}[\mathbb{E}_{i\in[n]}[\mathbf{\Pi}_i\mathbf{D}(Z^{(i)})\mathbf{M}(Z)\mathbf{D}(Z^{(i)})\mathbf{\Pi}_i]|Z]\Big) \\
&\preceq 2(\mathbf{\Delta}_{10} + \mathbf{\Delta}_{11})
\end{aligned}$$

where we define

$$\mathbf{\Delta}_{10} = \mathbb{E}_{i\in[n]}[\mathbf{\Pi}_i\overline{\mathbf{F}}_{k,a,b}^2\mathbf{\Pi}_i], \qquad \mathbf{\Delta}_{11} = \mathbb{E}[\mathbb{E}_{i\in[n]}[\mathbf{\Pi}_i\mathbf{D}(Z^{(i)})\mathbf{M}(Z)\mathbf{D}(Z^{(i)})\mathbf{\Pi}_i]|Z]$$

Invoking Lemma 2.1 over the interval $[0,\infty)$ with the convex continuous function $f(x) = x^t$, $\mathbf{B}_i = \overline{\mathbf{F}}_{k,a,b}^2$, $\mathbf{A}_i = \frac{1}{\sqrt{d_p}}\mathbf{\Pi}_i$ where we observe that $\sum_{i=1}^n \mathbf{A}_i\mathbf{A}_i^T = \frac{1}{d_p}\sum_{i=1}^n \mathbf{\Pi}_i^2 \preceq \mathbf{I}$, we get

$$\begin{aligned}
\mathbb{E}\,\|\mathbf{\Delta}_{10}\|_t^t = \mathbb{E}\,\mathrm{tr}[\mathbf{\Delta}_{10}^t] = \mathbb{E}\,\mathrm{tr}[\Big(\mathbb{E}_{i\in[n]}[\mathbf{\Pi}_i\overline{\mathbf{F}}_{k,a,b}^2\mathbf{\Pi}_i]\Big)^t] &= \frac{1}{n^t}\mathbb{E}\,\mathrm{tr}[\Big(\sum_{i=1}^n \mathbf{\Pi}_i\overline{\mathbf{F}}_{k,a,b}^2\mathbf{\Pi}_i\Big)^t] \\
&\le \frac{d_p^{t-1}}{n^t}\mathbb{E}\,\mathrm{tr}[\Big(\sum_{i=1}^n \mathbf{\Pi}_i\overline{\mathbf{F}}_{k,a,b}^{2t}\mathbf{\Pi}_i\Big)] \\
&\le \frac{d_p^{t-1}}{n^t}\mathbb{E}\,\mathrm{tr}[\Big(\sum_{i=1}^n \mathbf{\Pi}_i^2\Big)\overline{\mathbf{F}}_{k,a,b}^{2t}] \\
&\le \frac{d_p^t}{n^t}\mathbb{E}\,\mathrm{tr}[\overline{\mathbf{F}}_{k,a,b}^{2t}] \\
&= \frac{d_p^t}{n^t}\mathbb{E}\,\|\overline{\mathbf{F}}_{k,a,b}\|_{2t}^{2t}
\end{aligned}$$

Now, consider

$$\begin{aligned}
\mathbf{\Delta}_{11} &= \mathbb{E}[\mathbb{E}_{i\in[n]}[\mathbf{\Pi}_i\mathbf{D}(Z^{(i)})\mathbf{M}(Z)\mathbf{D}(Z^{(i)})\mathbf{\Pi}_i]|Z] \\
&= \mathbb{E}[\mathbb{E}_{i\in[n]}[(\sum_{l=1}^d \mathbf{\Pi}_{i,l})\mathbf{D}(Z^{(i)})\mathbf{M}(Z)\mathbf{D}(Z^{(i)})(\sum_{l=1}^d \mathbf{\Pi}_{i,l})]|Z] \\
&\preceq d\cdot\mathbb{E}[\mathbb{E}_{i\in[n]}[\sum_{l=1}^d \mathbf{\Pi}_{i,l}\mathbf{D}(Z^{(i)})\mathbf{M}(Z)\mathbf{D}(Z^{(i)})\mathbf{\Pi}_{i,l}]|Z] \\
&= d\cdot\mathbb{E}_{i\in[n]}[\sum_{l=1}^d \frac{\mathbb{E}[Z_i^{2l}]}{Z_i^{2l}}\mathbf{\Pi}_{i,l}\mathbf{D}(Z)\mathbf{M}(Z)\mathbf{D}(Z)\mathbf{\Pi}_{i,l}] \\
&= \frac{d}{n}\sum_{i=1}^n\sum_{l=1}^d \frac{\mathbb{E}[Z_i^{2l}]}{Z_i^{2l}}\mathbf{\Pi}_{i,l}\mathbf{D}(Z)\mathbf{M}(Z)\mathbf{D}(Z)\mathbf{\Pi}_{i,l} \\
&= \frac{d}{n}\sum_{i=1}^n\sum_{l=1}^d \mathbf{\Pi}_{i,l}\mathbf{\Sigma}_i\mathbf{D}(Z)\mathbf{M}(Z)\mathbf{D}(Z)\mathbf{\Sigma}_i^\intercal\mathbf{\Pi}_{i,l} \\
&= \frac{d}{n}\sum_{i=1}^n\sum_{l=1}^d \mathbf{\Pi}_{i,l}\mathbf{\Sigma}_i\overline{\mathbf{F}}_{k,a,b}^2\mathbf{\Sigma}_i^\intercal\mathbf{\Pi}_{i,l}
\end{aligned}$$

We now invoke Lemma 2.1 on dd_p terms with $\mathbf{B}_{i,l} = \overline{\mathbf{F}}^2_{k,a,b}$ and $\mathbf{A}_{i,l} = \frac{1}{\sqrt{d_p}}\mathbf{\Pi}_{i,l}\mathbf{\Sigma}_i$ where we observe that

$$\sum_{i=1}^{n}\sum_{l=1}^{d}\mathbf{A}_{i,l}\mathbf{A}_{i,l}^T = \frac{1}{d_p}\sum_{i=1}^{n}\sum_{l=1}^{d}\mathbf{\Pi}_{i,l}\mathbf{\Sigma}_i\mathbf{\Sigma}_i^\intercal\mathbf{\Pi}_{i,l}^\intercal = \frac{1}{d_p}\sum_{i=1}^{n}\sum_{l=1}^{d}\mathbf{\Pi}_{i,l}^2 \preceq \mathbf{I}$$

to get

$$\begin{aligned}\mathbb{E}\,\|\mathbf{\Delta}_{11}\|_t^t = \mathbb{E}\,\mathrm{tr}[\mathbf{\Delta}_{11}^t] &\le \frac{d^t}{n^t}\mathbb{E}\,\mathrm{tr}[(\sum_{i=1}^{n}\sum_{l=1}^{d}\mathbf{\Pi}_{i,l}\mathbf{\Sigma}_i\overline{\mathbf{F}}^2_{k,a,b}\mathbf{\Sigma}_i^\intercal\mathbf{\Pi}_{i,l})^t]\\ &\le \frac{(dd_p)^t}{n^t}\mathbb{E}\,\mathrm{tr}[\left(\frac{1}{d_p}\sum_{i=1}^{n}\sum_{l=1}^{d}\mathbf{\Pi}_{i,l}\mathbf{\Sigma}_i\overline{\mathbf{F}}^{2t}_{k,a,b}\mathbf{\Sigma}_i^\intercal\mathbf{\Pi}_{i,l}\right)]\\ &= \frac{(dd_p)^t}{n^t}\mathbb{E}\,\mathrm{tr}[\left(\frac{1}{d_p}\sum_{i=1}^{n}\sum_{l=1}^{d}\mathbf{\Sigma}_i^\intercal\mathbf{\Pi}_{i,l}\mathbf{\Pi}_{i,l}\mathbf{\Sigma}_i\overline{\mathbf{F}}^{2t}_{k,a,b}\right)]\end{aligned}$$

To simplify this, we use Fact 7.2 to get

$$\sum_{i=1}^{n}\sum_{l=1}^{d}\mathbf{\Sigma}_i^\intercal(\mathbf{\Pi}_{i,l})^2\mathbf{\Sigma}_i \preceq \sum_{i=1}^{n}\sum_{l=1}^{d}\mathbf{\Sigma}_i^\intercal(\mathbf{\Pi}'_{i,l})^2\mathbf{\Sigma}_i = \sum_{i=1}^{n}\sum_{l=1}^{d}\mathbf{\Pi}'_{i,l}\mathbf{\Sigma}_i^\intercal\mathbf{\Sigma}_i\mathbf{\Pi}'_{i,l} = \sum_{i=1}^{n}\sum_{l=1}^{d}\mathbf{\Pi}'_{i,l}\mathbf{\Pi}'_{i,l} \preceq d_p\mathbf{I}$$

Therefore, $\mathbb{E}\,\|\mathbf{\Delta}_{11}\|_t^t \le \frac{(dd_p)^t}{n^t}\mathbb{E}\,\mathrm{tr}[\overline{\mathbf{F}}^{2t}_{k,a,b}] = \frac{(dd_p)^t}{n^t}\mathbb{E}\,\|\overline{\mathbf{F}}_{k,a,b}\|_{2t}^{2t}$. Putting them together and using Fact 2.2,

$$\mathbb{E}\left\|\mathbf{\Delta}_1^{k,a,b}\right\|_t^t \le 4^t(\mathbb{E}\,\|\mathbf{\Delta}_{10}\|_t^t + \mathbb{E}\,\|\mathbf{\Delta}_{11}\|_t^t) \le \frac{(8dd_p)^t}{n^t}\mathbb{E}\,\|\overline{\mathbf{F}}_{k,a,b}\|_{2t}^{2t}$$

□

We now restate and prove Lemma 7.12.

LEMMA 7.12. *For all integers $t \ge 1$, $\mathbb{E}\left\|\mathbf{\Delta}_3^{k,a,b}\right\|_t^t \le \frac{(4d_p)^t}{n^t}\mathbb{E}\,\|\overline{\mathbf{F}}_{k,a,b}\|_{2t}^{2t}$.*

Proof. Recall that $Z' = Z^{(i)}$ for i sampled uniformly from $[n]$. Then,

$$\begin{aligned}\mathbf{\Delta}_3^{k,a,b} &= \mathbb{E}[(\mathbf{D}(Z)\overline{\mathbf{G}}_{k,a,b}(Z)(\mathbf{D}(Z)-\mathbf{D}(Z')))^2|Z]\\ &= \mathbb{E}[\mathbb{E}_{i\in[n]}[(\mathbf{D}(Z)\overline{\mathbf{G}}_{k,a,b}(Z)(\mathbf{D}(Z)-\mathbf{D}(Z^{(i)})))^2]|Z]\\ &= \mathbb{E}[\mathbb{E}_{i\in[n]}[(\mathbf{D}(Z)\overline{\mathbf{G}}_{k,a,b}(Z)\mathbf{\Pi}_i(\mathbf{D}(Z)-\mathbf{D}(Z^{(i)})))^2]|Z]\end{aligned}$$

where we use the fact that $\mathbf{D}(Z)-\mathbf{D}(Z^{(i)}) = \mathbf{\Pi}_i(\mathbf{D}(Z)-\mathbf{D}(Z^{(i)}))$ for all i. Define $\mathbf{M}(Z) = \mathbf{D}(Z)\overline{\mathbf{G}}_{k,a,b}$ to get

$$\begin{aligned}\mathbf{\Delta}_3^{k,a,b} &= \mathbb{E}[\mathbb{E}_{i\in[n]}[\mathbf{M}(Z)\mathbf{\Pi}_i(\mathbf{D}(Z)-\mathbf{D}(Z^{(i)}))^2\mathbf{\Pi}_i\mathbf{M}(Z)^\intercal]|Z]\\ &\preceq 2(\mathbb{E}[\mathbb{E}_{i\in[n]}[\mathbf{M}(Z)\mathbf{\Pi}_i\mathbf{D}(Z)^2\mathbf{\Pi}_i\mathbf{M}(Z)^\intercal]|Z] + \mathbb{E}[\mathbb{E}_{i\in[n]}[\mathbf{M}(Z)\mathbf{\Pi}_i\mathbf{D}(Z^{(i)})^2\mathbf{\Pi}_i\mathbf{M}(Z)^\intercal]|Z])\\ &= 2(\mathbb{E}_{i\in[n]}[\mathbf{M}(Z)\mathbf{\Pi}_i\mathbf{D}(Z)^2\mathbf{\Pi}_i\mathbf{M}(Z)^\intercal] + \mathbb{E}[\mathbb{E}_{i\in[n]}[\mathbf{M}(Z)\mathbf{\Pi}_i\mathbf{D}(Z^{(i)})^2\mathbf{\Pi}_i\mathbf{M}(Z)^\intercal]|Z])\\ &= 2(\mathbf{\Delta}_{30}+\mathbf{\Delta}_{31})\end{aligned}$$

where we define

$$\mathbf{\Delta}_{30} = \mathbb{E}_{i\in[n]}[\mathbf{M}(Z)\mathbf{\Pi}_i\mathbf{D}(Z)^2\mathbf{\Pi}_i\mathbf{M}(Z)^\intercal], \qquad \mathbf{\Delta}_{31} = \mathbb{E}[\mathbb{E}_{i\in[n]}[\mathbf{M}(Z)\mathbf{\Pi}_i\mathbf{D}(Z^{(i)})^2\mathbf{\Pi}_i\mathbf{M}(Z)^\intercal]|Z]$$

We have

$$\begin{aligned}\boldsymbol{\Delta}_{30} = \mathbb{E}_{i\in[n]}[\mathbf{M}(Z)\boldsymbol{\Pi}_i\mathbf{D}(Z)^2\boldsymbol{\Pi}_i\mathbf{M}(Z)^\intercal] &= \mathbb{E}_{i\in[n]}[\mathbf{M}(Z)\mathbf{D}(Z)\boldsymbol{\Pi}_i\boldsymbol{\Pi}_i\mathbf{D}(Z)\mathbf{M}(Z)^\intercal] \\ &= \mathbf{M}(Z)\mathbf{D}(Z)(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\Pi}_i^2)\mathbf{D}(Z)\mathbf{M}(Z)^\intercal \\ &\preceq \frac{d_p}{n}\mathbf{M}(Z)\mathbf{D}(Z)\mathbf{D}(Z)\mathbf{M}(Z)^\intercal \\ &= \frac{d_p}{n}\overline{\mathbf{F}}_{k,a,b}^2\end{aligned}$$

For the other term, using Fact 7.2,

$$\begin{aligned}\boldsymbol{\Delta}_{31} = \mathbb{E}[\mathbb{E}_{i\in[n]}[\mathbf{M}(Z)\boldsymbol{\Pi}_i\mathbf{D}(Z^{(i)})^2\boldsymbol{\Pi}_i\mathbf{M}(Z)^\intercal]|Z] &= \mathbb{E}_{i\in[n]}[\mathbf{M}(Z)\boldsymbol{\Pi}_i\boldsymbol{\Sigma}_i\mathbf{D}(Z)^2\boldsymbol{\Sigma}_i\boldsymbol{\Pi}_i\mathbf{M}(Z)^\intercal] \\ &\preceq \mathbb{E}_{i\in[n]}[\mathbf{M}(Z)\boldsymbol{\Pi}_i'\boldsymbol{\Sigma}_i\mathbf{D}(Z)^2\boldsymbol{\Sigma}_i\boldsymbol{\Pi}_i'\mathbf{M}(Z)^\intercal] \\ &= \mathbb{E}_{i\in[n]}[\mathbf{M}(Z)\boldsymbol{\Sigma}_i\boldsymbol{\Pi}_i'\mathbf{D}(Z)^2\boldsymbol{\Pi}_i'\boldsymbol{\Sigma}_i\mathbf{M}(Z)^\intercal] \\ &= \mathbb{E}_{i\in[n]}[\mathbf{D}(Z)\overline{\mathbf{G}}_{k,a,b}\boldsymbol{\Sigma}_i\boldsymbol{\Pi}_i'\mathbf{D}(Z)^2\boldsymbol{\Pi}_i'\boldsymbol{\Sigma}_i\overline{\mathbf{G}}_{k,a,b}\mathbf{D}(Z)]\end{aligned}$$

Observe that $\overline{\mathbf{G}}_{k,a,b}\boldsymbol{\Sigma}_i = \overline{\mathbf{G}}_{k,a,b}$ because the entries of $\overline{\mathbf{G}}$ only depend on α and not on γ, so permuting the γs will not have any effect on the matrix. Therefore,

$$\begin{aligned}\boldsymbol{\Delta}_{31} &\preceq \mathbb{E}_{i\in[n]}[\mathbf{D}(Z)\overline{\mathbf{G}}_{k,a,b}\boldsymbol{\Pi}_i'\mathbf{D}(Z)^2\boldsymbol{\Pi}_i'\overline{\mathbf{G}}_{k,a,b}\mathbf{D}(Z)] \\ &\preceq \mathbb{E}_{i\in[n]}[\mathbf{D}(Z)\overline{\mathbf{G}}_{k,a,b}\mathbf{D}(Z)\boldsymbol{\Pi}_i'\boldsymbol{\Pi}_i'\mathbf{D}(Z)\overline{\mathbf{G}}_{k,a,b}\mathbf{D}(Z)] \\ &= \mathbb{E}_{i\in[n]}\overline{\mathbf{F}}_{k,a,b}\boldsymbol{\Pi}_i'\boldsymbol{\Pi}_i'\overline{\mathbf{F}}_{k,a,b} \\ &= \frac{1}{n}\sum_{i=1}^{n}\overline{\mathbf{F}}_{k,a,b}\boldsymbol{\Pi}_i'\boldsymbol{\Pi}_i'\overline{\mathbf{F}}_{k,a,b} \\ &\preceq \frac{d_p}{n}\overline{\mathbf{F}}_{k,a,b}^2\end{aligned}$$

where we used the fact that $\sum_{i=1}^{n}\boldsymbol{\Pi}_i'\boldsymbol{\Pi}_i' \preceq d_p\mathbf{I}$. Putting them together,

$$\mathbb{E}\left\|\boldsymbol{\Delta}_3^{k,a,b}\right\|_t^t \le 2^t(\mathbb{E}\left\|\boldsymbol{\Delta}_{30}\right\|_t^t + \mathbb{E}\left\|\boldsymbol{\Delta}_{31}\right\|_t^t) \le 2^t\cdot 2\frac{d_p^t}{n^t}\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t} \le \frac{(4d_p)^t}{n^t}\mathbb{E}\left\|\overline{\mathbf{F}}_{k,a,b}\right\|_{2t}^{2t}$$

□

8 Application: Sparse graph matrices

We now consider sparse graph matrices, i.e., the setting $G \sim \mathcal{G}_{n,p}$ for $p \le \frac{1}{2}$. The main difference from dense graph matrices is the contribution of the edge factors. Naïvely bounding the contribution of each edge by it's absolute value, as explained in Section 5, each edge in the shape contributes a factor of $\sqrt{\frac{1-p}{p}}$. But in many cases, these bounds are not tight. In fact, they are not tight even in the basic case of the adjacency matrix. In this section, we obtain tighter bounds using our general recursion. As we will see, the improved bound will contain the edge factors only for edges within the vertex separator.

Let $\mathbf{M}_\tau$ be the graph matrix corresponding to shape τ where we use p-biased Fourier characters $G_{i,j}$. In this section, we obtain bounds on $\mathbb{E}\left\|\mathbf{M}_\tau - \mathbb{E}\mathbf{M}_\tau\right\|_{2t}^{2t}$ and use it to obtain high probability bounds on $\left\|\mathbf{M}_\tau\right\|$. Since many of the details are similar to Section 4.2.2 and the proof of Theorem 4.1, we will pass lightly over some details. We recommend the reader to read that section first.

The $G_{i,j}$ correspond to the Z_is in Section 6 and $\mathbf{F}$ corresponds to $\mathbf{M}_\tau$. Let $\mathcal{I}$ denote the set of sub-tuples of $[n]$. Each nonzero entry of $\mathbf{M}_\tau$ is a homogenous polynomial of degree $|E(\tau)|$. If $E(\tau) = \emptyset$, then, $\mathbf{M}_\tau - \mathbb{E}\mathbf{M}_\tau = 0$

so we can focus on the case when τ has at least one edge. Moreover, since degree-0 vertices in $V(\tau) \setminus U_\tau \setminus V_\tau$ simply scale the matrix by a factor of at most n, we can handle them separately and for our main analysis, we assume there are no such vertices in τ.

We will use Theorem 6.1 but the matrices and the statement can be drastically simplified in our application. Instate the notation of Section 6. Since we are dealing with multilinear polynomials, in the definition of $\mathcal{K}$, we can restrict our attention to $\alpha \in \{0,1\}^{\binom{n}{2}}$ because for any other $\alpha \in \mathbb{N}^n$, the corresponding row or column of $\mathbf{G}_{a+b,a,b}$ and hence $\mathbf{F}_{a+b,a,b}$, will be 0. So, we can accordingly redefine $\mathcal{K}$ to only contain these (α, γ), hence $\mathcal{K} \subseteq \{0,1\}^n \times \{0,1\}^n$.

Next, the diagonal matrices $\mathbf{D}_1, \mathbf{D}_2$ will both be equal to the diagonal matrix $\mathbf{D} \in \mathbb{R}[Z]^{\mathcal{I}\times\mathcal{K}} \times \mathbb{R}[Z]^{\mathcal{I}\times\mathcal{K}}$ with nonzero entries

$$\mathbf{D}[(I,\alpha,\gamma),(I,\alpha,\gamma)] = \sqrt{\mathbb{E}[\prod_{i,j} G_{ij}^{2\alpha_{ij}(1-\gamma)_{ij}}]} \prod_{i,j} G_i^{\alpha_{ij}\gamma_{ij}} = \prod_{i,j} G_i^{\alpha_{ij}\gamma_{ij}}$$

where we used the fact that for any i, j, $\mathbb{E}[G_{ij}^2] = 1$.

For integers $a, b \geq 0$ such that $a + b = |E(\tau)|$, define the matrix $\mathbf{M}_{\tau,a,b}$ to be the matrix $\mathbf{G}_{a+b,a,b}$. We use this notation in order to be streamlined with Section 4.2.2. That is, $\mathbf{M}_{\tau,a,b}$ has rows and columns indexed by $\mathcal{I} \times \mathcal{K}$ such that for all $(I,\alpha_1,\gamma_1),(J,\alpha_2,\gamma_2) \in \mathcal{I} \times \mathcal{K}$,

$$\mathbf{M}_{\tau,a,b}[(I,\alpha_1,\gamma_1),(J,\alpha_2,\gamma_2)] = \begin{cases} \nabla_{\alpha_1+\alpha_2}\mathbf{M}_\tau[I,J] & \text{if } |\alpha_1|_0 = a, |\alpha_2|_0 = b, \alpha_1 \cdot \alpha_2 = 0 \\ 0 & \text{o.w.} \end{cases}$$

This is almost identical to the $\mathbf{M}_{\tau,a,b}$ matrix defined in Section 4.2.2, with the difference being that the row and column indices now have γ in them. Therefore, for $I, J \in \mathcal{I}, (\alpha_1,\gamma_1),(\alpha_2,\gamma_2) \in \mathcal{K}$ such that $|\alpha_1|_0 = a, |\alpha_2|_0 = b, \alpha_1 \cdot \alpha_2 = 0$, the entry in row (I,α_1,γ_1) and column (J,α_2,γ_2) is the number of realizations φ of τ such that

- U_τ, V_τ map to I, J respectively under φ, and
- Under φ, the edges of τ map to the edges in α_1 and α_2 viewed as a set.

By Theorem 6.1, for integers $t \geq 1$,

$$\begin{aligned} \mathbb{E}\,\|\mathbf{M}_\tau - \mathbb{E}\mathbf{M}_\tau\|_{2t}^{2t} &\leq \sum_{a,b\geq 0, a+b\geq 1} (Ct^2 dd_p^4)^{(a+b)t} \mathbb{E}\,\|\mathbf{F}_{a+b,a,b}\|_{2t}^{2t} \\ &= \sum_{a,b\geq 0, a+b=|E(\tau)|} (Ct^2|E(\tau)|^4)^{t|E(\tau)|} \mathbb{E}\,\|\mathbf{D}\mathbf{M}_{\tau,a,b}\mathbf{D}\|_{2t}^{2t} \end{aligned}$$

for an absolute constant $C > 0$.

Now, we would like to analyze $\mathbb{E}\,\|\mathbf{D}\mathbf{M}_{\tau,a,b}\mathbf{D}\|_{2t}^{2t}$. Just as in the proof of Theorem 4.1, let P specify which edges of $E(\tau)$ go to α_1, α_2 respectively and in what order. Moreover, we now store extra information in P that indicates which entries of γ_1, γ_2 (relative to α_1, α_2) are set to 1. Let the set of such information P be denoted $\mathcal{P}$, then $|\mathcal{P}| \leq (4|E(\tau)|)^{t|E(\tau)|} 2^{|E(\tau)|}$. Thus,

$$\mathbb{E}\,\|\mathbf{D}\mathbf{M}_{\tau,a,b}\mathbf{D}\|_{2t}^{2t} \leq (8|E(\tau)|)^{t|E(\tau)|} \sum_{P\in\mathcal{P}} \mathbb{E}\,\|\mathbf{D}\mathbf{M}_{\tau,a,b,P}\mathbf{D}\|_{2t}^{2t}$$

where we define $\mathbf{M}_{\tau,a,b,P}$ similar to $\mathbf{M}_{\tau,a,b}$ with the extra condition that $\varphi, \alpha_1, \alpha_2, \gamma_1, \gamma_2$ must respect P.

At this point, in contrast to the proof of Theorem 4.1, note that the matrices $\mathbf{M}_{\tau,a,b,P}$ here have rows and columns indexed by $\mathcal{I} \times \mathcal{K}$. We will again define the shape τ_P that is equal to the nonzero block of the matrix $\mathbf{D}\mathbf{M}_{\tau,a,b,P}\mathbf{D}$, up to renaming of the rows and columns. $V(\tau_P), U_{\tau_P}, V_{\tau_P}$ are defined the same way as in Section 4.2.2 but to incorporate the action of $\mathbf{D}$ on these entries, we simply keep the edges that are active in γ_1 or γ_2, as prescribed by P. For an illustration, see Fig. 6.

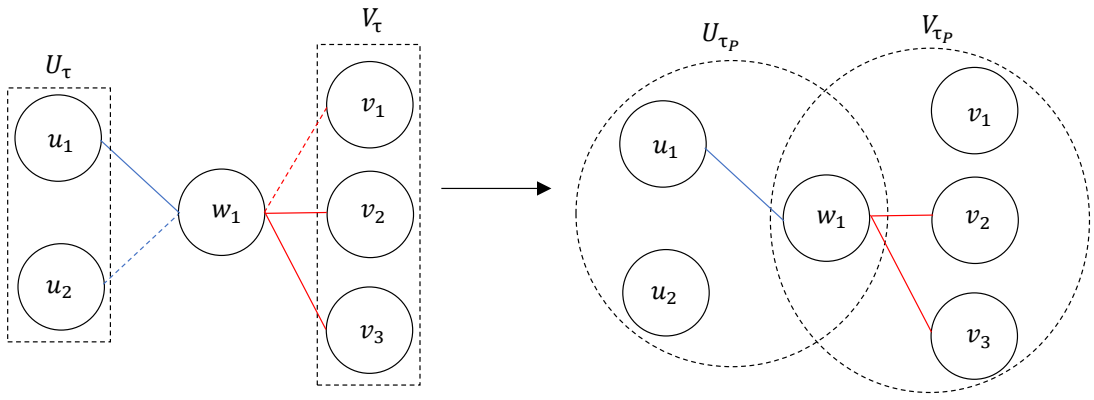

Figure 6: An example illustrating how τ_P is defined. In this example, P constraints the blue and red edges to go to α_1 and α_2 respectively. Moreover, P indicates that some edges are active in γ_1, γ_2 (indicated by a solid edge) and some are not active (indicated by a dashed edge) in γ_1, γ_2. We keep the solid edges in τ_P. U_{τ_P}, V_{τ_P} also have an ordering on the vertices (not shown here).

Then, by similar renaming of the rows and columns of $\mathbf{DM}_{\tau,a,b,P}\mathbf{D}$ and dropping the γs, we obtain $\mathbf{M}_{\tau_P}$. We therefore obtain the bound

$$\mathbb{E}\left\|\mathbf{DM}_{\tau,a,b}\mathbf{D}\right\|_{2t}^{2t} \leq (8|E(\tau)|)^{t|E(\tau)|}\sum_{P\in\mathcal{P}}\mathbb{E}\left\|\mathbf{M}_{\tau_P}\right\|_{2t}^{2t}$$

We would like to analyze norm bounds on the matrices $\mathbf{M}_{\tau_P}$. Observe that τ_P are shapes with the properties

- there are no vertices in $V(\tau_P) \setminus U_{\tau_P} \setminus V_{\tau_P}$
- each edge is either entirely contained in U_{τ_P} or entirely contained in V_{τ_P}

Call such shapes *simple*.

In the following lemma, whose proof is deferred to the next section, we prove norm bounds on simple shapes. Recall that in Lemma 4.2, we analyzed the norm bounds of simple shapes with no edges (because in this case, the graph distribution doesn't matter). The analysis for simple shapes is very similar but this time, we use scalar concentration tools to bound the Frobenius norm.

For a set S of vertices, denote by $E(S)$ the set of edges with both endpoints in S.

LEMMA 8.1. *For all even integers $t \geq 2$, if τ is a simple shape,*

$$\mathbb{E}\left\|\mathbf{M}_\tau\right\|_{2t}^{2t} \leq \left(n^{|V(\tau)|}(Ct)^{t|E(\tau)|}|V(\tau)|^{t|V(\tau)|}\right)\max_{U_\tau\cap V_\tau\subseteq S\subseteq V(\tau)}\left(\frac{1-p}{p}\right)^{t|E(S)|}n^{t(|V(\tau)|-|S|)}$$

for an absolute constant $C > 0$.

For simple shapes, the main difference from norm bounds on corresponding dense graph matrices is that each edge within S contributes a factor of $\sqrt{\frac{1-p}{p}}$. Edge contributions are unavoidable when handling sparse graph matrices, but we have identified that we need not consider all edges in the shape but only a subset of it. Using this lemma, we can obtain norm bounds on general graph matrices. We first recall the definition of a vertex separator.

DEFINITION 4.4. (VERTEX SEPARATOR) *For a shape τ, define a vertex separator to be a subset of vertices $S \subseteq V(\tau)$ such that there is no path from U_τ to V_τ in $\tau \setminus S$, which is the shape obtained by deleting all the vertices of S (including all edges they're incident on).*

Let I_τ be the set of isolated vertices (vertices of degree 0) in $V(\tau) \setminus U_\tau \setminus V_\tau$, so they essentially scale the matrix by a scalar factor. We now state the main theorem of this section.

THEOREM 8.1. *For all even integers $t \geq 2$, for any shape τ,*

$$\mathbb{E}\,\|\mathbf{M}_\tau - \mathbb{E}\mathbf{M}_\tau\|_{2t}^{2t} \leq \left(n^{|V(\tau)|}|V(\tau)|^{t|V(\tau)|}(Ct^3|E(\tau)|^5)^{t|E(\tau)|}\right) \max_{\text{vertex separator } S} \left(\frac{1-p}{p}\right)^{t|E(S)|} n^{t(|V(\tau)|-|S|+|I_\tau|)}$$

where the maximum is over all vertex separators S.

To interpret this bound, if we assume that there are a constant number of vertices in τ, then by choosing $t \approx \text{polylog}(n)$, we get

$$\|\mathbf{M}_\tau\| = \widetilde{O}\left(\max_{\text{vertex separator } S} \left(\sqrt{\frac{1-p}{p}}\right)^{|E(S)|} \sqrt{n}^{|V(\tau)-|S|+|I_\tau|}\right)$$

with high probability, where $\widetilde{O}$ hides logarithmic factors. This result follows from Theorem 8.1 if τ has at least one edge, but also applies if τ has no edges, in which case we can directly use the far simpler Lemma 4.2. A precise form of the above characterization is given in Corollary 8.1.

Theorem 8.1 gives us the right dependence on p, n for norm bounds in the case of sparse graph matrices. The same bound, up to lower order terms, was also obtained in [JPR+21] via the trace power method, where they use these bounds to prove semidefinite-programming lower bounds for the maximum independent set problem on sparse graphs.

Proof. [Proof of Theorem 8.1] If $E(\tau) = \emptyset$, then $\mathbf{M}_\tau = \mathbb{E}\mathbf{M}_\tau$ and we are done. So, assume $E(\tau) \neq \emptyset$. Since vertices in I_τ only scale the matrix by a factor of at most n, we can handle them separately and our bound has the appropriate power of n coming from these. Therefore, we can assume $I_\tau = \emptyset$. Continuing our prior discussions, for an absolute constant $C_1 > 0$,

$$\begin{aligned}\mathbb{E}\,\|\mathbf{M}_\tau - \mathbb{E}\mathbf{M}_\tau\|_{2t}^{2t} &\leq \sum_{a,b\geq 0, a+b=|E(\tau)|} (C_1 t^2|E(\tau)|^4)^{t|E(\tau)|}\mathbb{E}\left\|\mathbf{D}\mathbf{M}_{\tau,a,b}\mathbf{D}\right\|_{2t}^{2t} \\ &\leq \sum_{a,b\geq 0, a+b=|E(\tau)|} (C_1 t^2|E(\tau)|^4)^{t|E(\tau)|}(8|E(\tau)|)^{t|E(\tau)|} \sum_{\psi\in\Gamma_{a,b}} \mathbb{E}\left\|\mathbf{M}_\psi\right\|_{2t}^{2t}\end{aligned}$$

where $\Gamma_{a,b}$ are the set of simple shapes we obtain for $\mathbf{D}\mathbf{M}_{\tau,a,b}\mathbf{D}$, as per our discussion above. Using Lemma 8.1, for an absolute constant $C_2 > 0$, we have

$$\begin{aligned}&\mathbb{E}\,\|\mathbf{M}_\tau - \mathbb{E}\mathbf{M}_\tau\|_{2t}^{2t} \\ &\leq \left(n^{|V(\tau)|}|V(\tau)|^{t|V(\tau)|}(C_2 t^3|E(\tau)|^5)^{t|E(\tau)|}\right) \sum_{a,b\geq 0, a+b=|E(\tau)|} \sum_{\psi\in\Gamma_{a,b}} \max_{U_\psi\cap V_\psi \subseteq S\subseteq V(\psi)} \left(\frac{1-p}{p}\right)^{t|E(S)|} n^{t(|V(\psi)|-|S|)}\end{aligned}$$

For any a, b, consider any simple shape $\psi \in \Gamma_{a,b}$ that can be obtained. As observed in the proof of Theorem 4.1 (see in particular Fig. 5), $U_\psi \cap V_\psi$ must be a vertex separator of τ. Therefore, any $S \supseteq U_\psi \cap V_\psi$ must be a vertex separator of τ. It's easy to see that as S ranges over all sets such that $U_\psi \cap V_\psi \subseteq S \subseteq V(\psi)$, it ranges over all vertex separators of τ.

Also, the number of different ψ is at most $4^{|E(\tau)|}$ since each edge can go either to U_ψ or V_ψ and for each such choice, it can either be active in γ or not. Therefore,

$$\begin{aligned}&\mathbb{E}\,\|\mathbf{M}_\tau - \mathbb{E}\mathbf{M}_\tau\|_{2t}^{2t} \\ &\leq \left(n^{|V(\tau)|}|V(\tau)|^{t|V(\tau)|}(C_2 t^3|E(\tau)|^5)^{t|E(\tau)|}\right) 4^{|E(\tau)|} \max_{\text{vertex separator } S} \left(\frac{1-p}{p}\right)^{t|E(S)|} n^{t(|V(\tau)|-|S|)} \\ &\leq \left(n^{|V(\tau)|}|V(\tau)|^{t|V(\tau)|}(Ct^3|E(\tau)|^5)^{t|E(\tau)|}\right) \max_{\text{vertex separator } S} \left(\frac{1-p}{p}\right)^{t|E(S)|} n^{t(|V(\tau)|-|S|)}\end{aligned}$$

for an absolute constant $C > 0$. ∎

The following corollary obtains high probability norm bounds for norms of graph matrices via Markov's inequality. We assume the graph has at least one edge, otherwise it is deterministic and its norm bound was already analyzed in Lemma 4.2, Corollary 4.1, where we observe that the distinction between sparse and dense graph matrices does not matter if the random matrix is deterministic.

COROLLARY 8.1. *For a shape τ with at least one edge, for any constant $\varepsilon > 0$, with probability $1 - \varepsilon$,*

$$\|\mathbf{M}_\tau\| \leq \left(|V(\tau)|^{|V(\tau)|/2}(C|E(\tau)|^5 \log^3(n^{|V(\tau)|}/\varepsilon))^{|E(\tau)|/2}\right) \cdot \max_{\text{vertex separator } S} \left(\sqrt{\frac{1-p}{p}}\right)^{|E(S)|} \sqrt{n}^{|V(\tau)-|S|+|I_\tau|}$$

for an absolute constant $C > 0$.

Proof. Since $|E(\tau)| \geq 1$, $\mathbb{E}\mathbf{M}_\tau = 0$. By an application of Markov's inequality,

$$\begin{aligned} Pr[\|\mathbf{M}_\tau\| \geq \theta] &\leq Pr[\|\mathbf{M}_\tau\|_{2t}^{2t} \geq \theta^{2t}] \\ &\leq \theta^{-2t}\mathbb{E}\,\|\mathbf{M}_\tau\|_{2t}^{2t} \\ &\leq \theta^{-2t}\left(n^{|V(\tau)|}|V(\tau)|^{t|V(\tau)|}(C't^3|E(\tau)|^5)^{t|E(\tau)|}\right) \max_{\text{vertex separator } S} \left(\frac{1-p}{p}\right)^{t|E(S)|} n^{t(|V(\tau)|-|S|+|I_\tau|)} \end{aligned}$$

for an absolute constant $C' > 0$. We now set

$$\begin{aligned} \theta = &\left(\varepsilon^{-1/(2t)}(C'')^{|E(\tau)|}n^{|V(\tau)|/(2t)}|V(\tau)|^{|V(\tau)|/2}t^{3|E(\tau)|/2}|E(\tau)|^{5|E(\tau)|/2}\right) \\ &\cdot \max_{\text{vertex separator } S} \left(\sqrt{\frac{1-p}{p}}\right)^{|E(S)|} \sqrt{n}^{|V(\tau)-|S|+|I_\tau|} \end{aligned}$$

for an absolute constant $C'' > 0$, to make this expression at most ε. Set $t = \frac{1}{2}\log(n^{|V(\tau)|}/\varepsilon)$ to complete the proof. ∎

8.1 Norm bounds on simple graph matrices In this section, we will prove Lemma 8.1. First, we recall the following scalar concentration result from [SS11].

8.1.1 Schudy-Sviridenko moment bound The definitions and main bound in this section are from [SS11].

DEFINITION 8.1. *A random variable Z is central moment bounded with real parameter $L > 0$ if for any integer $i \geq 1$,*

$$\mathbb{E}[|Z - \mathbb{E}[Z]|^i] \leq i \cdot L \cdot \mathbb{E}[|Z - \mathbb{E}[Z]|^{i-1}]$$

PROPOSITION 8.1. *The p-biased Bernoulli random variable Z is central moment bounded with real parameter $L = \sqrt{\frac{1-p}{p}}$.*

Proof. We have $\mathbb{E}[Z] = 0$ and for $p \leq \frac{1}{2}$, $|Z| \leq \sqrt{\frac{1-p}{p}}$, therefore,

$$\begin{aligned} \mathbb{E}[|Z - \mathbb{E}[Z]|^i] &= p\sqrt{\frac{p}{1-p}}^i + (1-p)\sqrt{\frac{1-p}{p}}^i \\ &\leq \sqrt{\frac{1-p}{p}}\left(p\sqrt{\frac{p}{1-p}}^{i-1} + (1-p)\sqrt{\frac{1-p}{p}}^{i-1}\right) \\ &= \sqrt{\frac{1-p}{p}}\mathbb{E}[|Z - \mathbb{E}[Z]|^{i-1}] \end{aligned}$$

therefore, we can take $L = \sqrt{\frac{1-p}{p}}$. ∎

For a given multilinear polynomial $f(x)$ on variables $x_1, \ldots, x_n$, we can naturally associate with it a hypergraph H on vertices $[n]$ and weighted hyperedges $E(H)$ where each $h \in E(H)$ corresponds to a distinct term of $f(x)$. Each hyperedge h is a subset $V(h)$ of vertices and has a real valued weight w_h which is the coefficient of that monomial in f. Therefore,

$$f(x) = \sum_{h \in E(H)} w_h \prod_{v \in V(h)} x_v$$

Assume f has degree d_p, then each hyperedge of H has at most d_p vertices. Now, for a given collection of independent random variables $Y_1, \ldots, Y_n$, a multilinear poynomial f with associated hypergraph H and weights w, and an integer $r \geq 0$, define

$$\mu_r(f, Y) = \max_{S \subseteq [n], |S| = r} \left(\sum_{h \in E(H), S \subseteq V(h)} |w_h| \prod_{v \in V(h) \setminus S} \mathbb{E}[|Y_v|] \right)$$

LEMMA 8.2. ([SS11], LEMMA 5.1) *Given n independent central moment bounded random variables $Y_1, \ldots, Y_n$ with the same parameter $L > 0$ and a degree d_p multilinear polynomial $f(x)$. Let $t \geq 2$ be an even integer, then*

$$\mathbb{E}[|f(Y) - \mathbb{E}[f(Y)]|^t] \leq \max \left\{ \left(\sqrt{t R_4^{d_p} \mathrm{Var}[f(Y)]} \right)^t, \max_{r \in [d_p]} (t^r R_4^{d_p} L^r \mu_r(f, Y))^t \right\}$$

where $R_4 \geq 1$ is some absolute constant.

In our setting, we can also bound the variance in terms of the μ_r as was shown in [SS11], which will simplify our calculations.

LEMMA 8.3. ([SS11], LEMMA 1.5) *For the same setting as in Lemma 8.2,*

$$\mathrm{Var}[f(Y)] \leq 2 d_p 4^{d_p} \max_{r \in [d_p]} (\mu_0(f, Y) \mu_r(f, Y) 4^r L^r)$$

8.1.2 Proof of Lemma 8.1

We are ready to prove Lemma 8.1 which we restate for convenience.

LEMMA 8.1. *For all even integers $t \geq 2$, if τ is a simple shape,*

$$\mathbb{E} \|\mathbf{M}_\tau\|_{2t}^{2t} \leq \left(n^{|V(\tau)|} (Ct)^{t|E(\tau)|} |V(\tau)|^{t|V(\tau)|} \right) \max_{U_\tau \cap V_\tau \subseteq S \subseteq V(\tau)} \left(\frac{1-p}{p} \right)^{t|E(S)|} n^{t(|V(\tau)| - |S|)}$$

for an absolute constant $C > 0$.

We will prove it the same way as Lemma 4.2, by bounding the schatten norm of each diagonal block by an appropriate power of its Frobenius norm. In this case, to bound the expected power of the Frobenius norm, we use the scalar concentration inequality from the previous section.

Proof. [Proof of Lemma 8.1] First, we note that $\mathbf{M}_\tau$ has a block diagonal structure indexed by the realizations of the set of common vertices $S_0 = U_{\tau_p} \cap V_{\tau_p}$. For $T \in [n]^{S_0}$, let $\mathbf{M}_{\tau,T}$ be the block of $\mathbf{M}_\tau$ with $\varphi(S_0) = T$. Then, $\mathbf{M}_{\tau,T} \mathbf{M}_{\tau,T'}^\intercal = \mathbf{M}_{\tau,T}^\intercal \mathbf{M}_{\tau,T'} = 0$ for $T \neq T'$ and so,

$$\mathbb{E} \|\mathbf{M}_\tau\|_{2t}^{2t} = \sum_{T \in [n]^{S_0}} \mathbb{E} \|\mathbf{M}_{\tau,T}\|_{2t}^{2t} \leq \sum_{T \in [n]^{S_0}} \mathbb{E} (\|\mathbf{M}_{\tau,T}\|_2^2)^t$$

where we bounded the Schatten norm by a power of the Frobenius norm.

Fix $T \in [n]^{S_0}$ and consider $\mathbb{E}\,\|\mathbf{M}_{\tau,T}\|_2^2$. Let $\mathcal{R}$ be the set of realizations φ of τ such that $\varphi(S_0) = T$. Then, for $\varphi \in \mathcal{R}$ and $e \in E(S_0)$, the value of $\varphi(e)$ is fixed. Using this,

$$\begin{aligned}\|\mathbf{M}_{\tau,T}\|_2^2 &= \sum_{\varphi\in\mathcal{R}} \prod_{e\in E(\tau)} G_{\varphi(e)}^2 = \prod_{e\in E(S_0)} G_{\varphi(e)}^2 \sum_{\varphi\in\mathcal{R}} \prod_{e\in E(\tau)\setminus E(S_0)} G_{\varphi(e)}^2 \\ &\le L^{|E(S_0)|} \sum_{\varphi\in\mathcal{R}} \prod_{e\in E(\tau)\setminus E(S_0)} G_{\varphi(e)}^2\end{aligned}$$

where $L = \frac{1-p}{p}$ is an upper bound on G_{ij}^2 for $p \le \frac{1}{2}$. For convenience, we define the quantity $A = \max_{S_0\subseteq S\subseteq V(\tau)} L^{|E(S)|} n^{|V(\tau)|-|S|}$.

CLAIM 8.1. $\mathbb{E}(\|\mathbf{M}_{\tau,T}\|_2)^t \le (Ct)^{t|E(\tau)|}|V(\tau)|^{t|V(\tau)|}A^t$ *for an absolute constant* $C > 0$.

Using this claim, we have

$$\begin{aligned}\mathbb{E}\,\|\mathbf{M}_\tau\|_{2t}^{2t} &\le \sum_{T\in[n]^{S_0}} \mathbb{E}(\|\mathbf{M}_{\tau,T}\|_2)^t \\ &\le n^{|S_0|}(Ct)^{t|E(\tau)|}|V(\tau)|^{t|V(\tau)|}A^t \\ &= n^{|V(\tau)|}(Ct)^{t|E(\tau)|}|V(\tau)|^{t|V(\tau)|} \max_{U_\tau\cap V_\tau\subseteq S\subseteq V(\tau)} \left(\frac{1-p}{p}\right)^{t|E(S)|} n^{t(|V(\tau)|-|S|)}\end{aligned}$$

as required. ∎

It remains to prove the claim.

Proof. [Proof of Claim 8.1] For $1 \le i, j \le n$, define the variables $Y_{ij} = G_{ij}^2$ with $\mathbb{E}[|Y_{ij}|] = 1$. Let $f(Y)$ be the polynomial $L^{|E(S_0)|}\sum_{\varphi\in\mathcal{R}}\prod_{e\in E(\tau)\setminus E(S_0)} Y_{\varphi(e)}$. It suffices to prove that $\mathbb{E}[f(Y)^t] \le (Ct)^{t|E_1|}A^t$. We will first prove that $\mathbb{E}[(f(Y) - \mathbb{E}[f(Y)])^t] \le (C't)^{t|E(\tau)|}|V(\tau)|^{t|V(\tau)|}A^t$ for a sufficiently large constant $C' > 0$.

f is a homogeneous multilinear polynomial of degree $|E(\tau)\setminus E(S_0)|$. If we had $E(\tau)\setminus E(S_0) = \emptyset$, then f is a constant and so, the inequality is obvious because $f(Y) = \mathbb{E}[f(Y)]$. Now, assume $E(\tau)\setminus E(S_0) \ne \emptyset$. We invoke Lemma 8.2. Let f have associated hypergraph H and weights w. Then,

$$\mathbb{E}[|f(Y) - \mathbb{E}[f(Y)]|^t] \le \max\left\{\left(\sqrt{tR_4^{|E(\tau)\setminus E(S_0)|}\mathrm{Var}[f(Y)]}\right)^t, \max_{r\in[|E(\tau)\setminus E(S_0)|]}(t^r R_4^{|E(\tau)\setminus E(S_0)|}L^r\mu_r(f,Y))^t\right\}$$

For all $r \ge 0$, we will prove that $L^r\mu_r(f,Y) \le |V(\tau)|^{|V(\tau)|}A$. By definition,

$$\mu_r(f,Y) = \max_{F\subseteq\binom{[n]}{2},|F|=r} \sum_{h\in E(H),F\subseteq V(h)} |w_h|$$

Consider any set of edge labels $F \subseteq \binom{[n]}{2}, |F| = r$. Then, $\sum_{h\in E(H),F\subseteq V(h)}|w_h|$ is at most $L^{|E(S_0)|}c$ where c is the number of realizations $\varphi \in \mathcal{R}$ such that $\varphi(E(\tau))$ contains F. Suppose F contains v new labels apart from $\varphi(S_0) = T$. Then $c \le |V(\tau)|^v n^{|V(\tau)|-|S_0|-v}$ because we can first choose and label the set of vertices that get these v labels and then label the remaining vertices freely, each of which has at most n choices.

Observe that $L^{|E(S_0)|}L^r n^{|V(\tau)|-|S_0|-v} \le A$ because in the definition of S, we can set S to be the union of S and any valid choice of these v vertices. Putting this together, we get

$$L^r\mu_r(f,Y) \le L^r \max_{F\subseteq\binom{[n]}{2},|F|=r} \sum_{h\in E(H),F\subseteq V(h)} |w_h| \le |V(\tau)|^{|V(\tau)|}A$$

which implies

$$\max_{r\in[|E(\tau)\setminus E(S_0)|]}(t^r R_4^{|E(\tau)\setminus E(S_0)|}L^r\mu_r(f,Y))^t \le |V(\tau)|^{t|V(\tau)|}(R_4 t)^{t|E(\tau)|}A^t$$

and using Lemma 8.3,

$$\begin{aligned}\mathrm{Var}[f(Y)] &\le 2|E(\tau)|4^{|E(\tau)|}\max_{r\in[|E(\tau)\setminus E(S_0)|]}(\mu_0(f,Y)\mu_r(f,Y)4^r L^r)\\ &\le 2|E(\tau)|16^{|E(\tau)|}|V(\tau)|^{2|V(\tau)|}A^2\end{aligned}$$

Putting them together, we get

$$\begin{aligned}\mathbb{E}[(f(Y)-\mathbb{E}[f(Y)])^t] &\le \max\left\{\left(\sqrt{2tR_4^{|E(\tau)|}|E(\tau)|16^{|E(\tau)|}|V(\tau)|^{2|V(\tau)|}A^2}\right)^t, |V(\tau)|^{t|V(\tau)|}(R_4 t)^{t|E(\tau)|}A^t\right\}\\ &\le (C't)^{t|E(\tau)|}|V(\tau)|^{t|V(\tau)|}A^t\end{aligned}$$

for an absolute constant $C'>0$. Finally, $\mathbb{E}[f(Y)] \le L^{|E(S_0)|}|\mathcal{R}| \le L^{|E(S_0)|}n^{|V(\tau)\setminus S_0|} \le A$ which gives

$$\begin{aligned}\mathbb{E}[f(Y)^t] \le 2^t(\mathbb{E}[(f(Y)-\mathbb{E}[f(Y)])^t]+\mathbb{E}[f(Y)]^t) &\le 2^t((C't)^{t|E(\tau)|}|V(\tau)|^{t|V(\tau)|}A^t + A^t)\\ &\le (Ct)^{t|E(\tau)|}|V(\tau)|^{t|V(\tau)|}A^t\end{aligned}$$

for an absolute constant $C>0$. ∎

Acknowledgements

We thank Aaron Potechin, Chris Jones and Jeff Xu for helpful discussions regarding graph matrices. We thank Pierre Youssef for helpful pointers to references. We thank anonymous reviewers for their suggestions and thank Chris Jones and Aaron Potechin for proofreading an earlier version of this manuscript.

References

[ABY20] Richard Aoun, Marwa Banna, and Pierre Youssef. Matrix poincaré inequalities and concentration. *Advances in Mathematics*, 371:107251, 2020.

[AD21] Srinivasan Arunachalam and João F Doriguello. Matrix hypercontractivity, streaming algorithms and ldcs: the large alphabet case. *arXiv preprint arXiv:2109.02600*, 2021.

[AMP16] Kwangjun Ahn, Dhruv Medarametla, and Aaron Potechin. Graph matrices: Norm bounds and applications. *arXiv*, pages arXiv–1604, 2016.

[AW15] Radosław Adamczak and Paweł Wolff. Concentration inequalities for non-lipschitz functions with bounded derivatives of higher order. *Probability Theory and Related Fields*, 162(3):531–586, 2015.

[BARDW08] Avraham Ben-Aroya, Oded Regev, and Ronald De Wolf. A hypercontractive inequality for matrix-valued functions with applications to quantum computing and ldcs. In *Proceedings of the 49th IEEE Symposium on Foundations of Computer Science*, pages 477–486. IEEE, 2008.

[BBH+12] Boaz Barak, Fernando G. S. L. Brandão, Aram Wettroth Harrow, Jonathan A. Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. *CoRR*, abs/1205.4484, 2012.

[BBLM05] Stéphane Boucheron, Olivier Bousquet, Gábor Lugosi, and Pascal Massart. Moment inequalities for functions of independent random variables. *The Annals of Probability*, 33(2):514 – 560, 2005.

[BGBK20] Florent Benaych-Georges, Charles Bordenave, and Antti Knowles. Spectral radii of sparse random matrices. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 56, pages 2141–2161. Institut Henri Poincaré, 2020.

[BGS19] Sergey G Bobkov, Friedrich Götze, and Holger Sambale. Higher order concentration of measure. *Communications in Contemporary Mathematics*, 21(03):1850043, 2019.

[BHK+19] Boaz Barak, Samuel Hopkins, Jonathan Kelner, Pravesh K Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *SIAM Journal on Computing*, 48(2):687–735, 2019.

[BHKX22] Mitali Bafna, Jun-Ting Hsieh, Pravesh K Kothari, and Jeff Xu. Polynomial-time power-sum decomposition of polynomials. *arXiv preprint arXiv:2208.00122*, 2022.

[CH19] Hao-Chung Cheng and Min-Hsiu Hsieh. Matrix Poincaré, φ-Sobolev inequalities, and quantum ensembles. *Journal of Mathematical Physics*, 60(3):032201, 2019.

[Cha05] Sourav Chatterjee. *Concentration inequalities with exchangeable pairs*. Stanford University, 2005.

[Cha06] Sourav Chatterjee. Stein's method for concentration inequalities. *arXiv preprint math/0604352*, 2006.

[CHT17] Hao-Chung Cheng, Min-Hsiu Hsieh, and Marco Tomamichel. Exponential decay of matrix φ-entropies on markov semigroups with applications to dynamical evolutions of quantum ensembles. *Journal of Mathematical Physics*, 58(9):092202, 2017.

[DM15] Yash Deshpande and Andrea Montanari. Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems. In *Conference on Learning Theory*, pages 523–562. PMLR, 2015.

[FK81] Zoltán Füredi and János Komlós. The eigenvalues of random symmetric matrices. *Combinatorica*, 1(3):233–241, 1981.

[GJJ+20] Mrinalkanti Ghosh, Fernando Granha Jeronimo, Chris Jones, Aaron Potechin, and Goutham Rajendran. Sum-of-squares lower bounds for sherrington-kirkpatrick via planted affine planes. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 954–965. IEEE, 2020.

[GKS21] Ankit Garg, Tarun Kathuria, and Nikhil Srivastava. Scalar poincaré implies matrix poincaré. *Electronic Communications in Probability*, 26:1–4, 2021.

[GM15] Rong Ge and Tengyu Ma. Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms. *arXiv preprint arXiv:1504.05287*, 2015.

[HKP15] Samuel B Hopkins, Pravesh K Kothari, and Aaron Potechin. SoS and planted clique: Tight analysis of mpw moments at all degrees and an optimal lower bound at degree four. *arXiv preprint arXiv:1507.05230*, 2015.

[Hop18] Samuel Hopkins. *Statistical inference and the sum of squares method*. PhD thesis, Cornell University, 2018.

[HP03] Frank Hansen and Gert K Pedersen. Jensen's operator inequality. *Bulletin of the London Mathematical Society*, 35(4):553–564, 2003.

[HSS15] Samuel B Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *Conference on Learning Theory*, pages 956–1006. PMLR, 2015.

[HSS19] Samuel B Hopkins, Tselil Schramm, and Jonathan Shi. A robust spectral algorithm for overcomplete tensor decomposition. In *Conference on Learning Theory*, pages 1683–1722. PMLR, 2019.

[HSSS16] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 178–191, 2016.

[HT21a] De Huang and Joel A. Tropp. From Poincaré inequalities to nonlinear matrix concentration. *Bernoulli*, 27(3):1724 – 1744, 2021.

[HT21b] De Huang and Joel A Tropp. Nonlinear matrix concentration via semigroup methods. *Electronic Journal of Probability*, 26:1–31, 2021.

[Jon22] Chris Jones. *Symmetrized Fourier Analysis of Convex Relaxations for Combinatorial Optimization Problems*. PhD thesis, The University of Chicago, 2022.

[JPR+21] Chris Jones, Aaron Potechin, Goutham Rajendran, Madhur Tulsiani, and Jeff Xu. Sum-of-squares lower bounds for sparse independent set. *IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, 2021.

[KP20] Bohdan Kivva and Aaron Potechin. Exact nuclear norm, completion and decomposition for random overcomplete tensors via degree-4 sos. *arXiv preprint arXiv:2011.09416*, 2020.

[Lat06] Rafał Latała. Estimates of moments and tails of gaussian chaoses. *The Annals of Probability*, 34(6):2315–2331, 2006.

[MP16] Dhruv Medarametla and Aaron Potechin. Bounds on the norms of uniform low degree graph matrices. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2016)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2016.

[MPW15] Raghu Meka, Aaron Potechin, and Avi Wigderson. Sum-of-squares lower bounds for planted clique. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 87–96, 2015.

[MRX20] Sidhanth Mohanty, Prasad Raghavendra, and Jeff Xu. Lifting sum-of-squares lower bounds: degree-2 to degree-4. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 840–853, 2020.

[MW19] Ankur Moitra and Alexander S Wein. Spectral methods from tensor networks. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 926–937, 2019.

[O'D08] R. O'Donnell. Some topics in analysis of Boolean functions. In *STOC*, pages 569–578, 2008.

[O'D14] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[Pet94] Dénes Petz. A survey of certain trace inequalities. *Banach Center Publications*, 30(1):287–298, 1994.

[PMT16] Daniel Paulin, Lester Mackey, and Joel A. Tropp. Efron–stein inequalities for random matrices. *Ann. Probab.*, 44(5):3431–3473, 09 2016.

[PR20] Aaron Potechin and Goutham Rajendran. Machinery for proving sum-of-squares lower bounds on certification problems. *arXiv preprint arXiv:2011.04253*, 2020.

[PR22] Aaron Potechin and Goutham Rajendran. Sub-exponential time sum-of-squares lower bounds for principal components analysis. *Advances in Neural Information Processing Systems*, 2022.

[Raj22a] Goutham Rajendran. Combinatorial Optimization via the Sum of Squares Hierarchy. *arXiv preprint arXiv:2208.04374*, 2022. Master's thesis, University of Chicago.
[Raj22b] Goutham Rajendran. Nonlinear random matrices and applications to the sum of squares hierarchy. 2022.
[RS15] Prasad Raghavendra and Tselil Schramm. Tight lower bounds for planted clique in the degree-4 SOS program. *arXiv preprint arXiv:1507.05136*, 2015.
[SS11] Warren Schudy and Maxim Sviridenko. Bernstein-like concentration and moment inequalities for polynomials of independent random variables: multilinear case. *arXiv preprint arXiv:1109.5193*, 2011.
[SS17] Tselil Schramm and David Steurer. Fast and robust tensor decomposition with applications to dictionary learning. In *Conference on Learning Theory*, pages 1760–1793. PMLR, 2017.
[Ste72] Charles Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*, pages 583–602. University of California Press, 1972.
[Ste86] Charles Stein. Approximate computation of expectations. IMS, 1986.
[Tro15] Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.
[Vu05] Van H Vu. Spectral norm of random matrices. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 423–430, 2005.